UNIVERSIDAD NACIONAL DEL LITORAL

DOCTORADO EN INGENIERÍA

Algoritmos avanzados para la detección del síndrome de apnea-hipopnea obstructiva del sueño

Román Emanuel Rolón

FICH
FACULTAD DE INGENIERÍA Y CIENCIAS HÍDRICAS

INTEC
INSTITUTO DE DESARROLLO TECNOLÓGICO PARA LA INDUSTRIA QUÍMICA

CIMEC
CENTRO DE INVESTIGACIÓN DE MÉTODOS COMPUTACIONALES

sinc(*i*)
INSTITUTO DE INVESTIGACIÓN EN SEÑALES, SISTEMAS E INTELIGENCIA COMPUTACIONAL

Tesis de Doctorado **2018**

UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas
Instituto de Desarrollo Tecnológico para la Industria Química
Centro de Investigación de Métodos Computacionales
Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional

ALGORITMOS AVANZADOS PARA LA DETECCIÓN DEL SÍNDROME DE APNEA-HIPOPNEA OBSTRUCTIVA DEL SUEÑO

Román Emanuel Rolón

Tesis remitida al Comité Académico del Doctorado
como parte de los requisitos para la obtención
del grado de
DOCTOR EN INGENIERÍA
Mención Inteligencia Computacional, Señales y Sistemas
de la
UNIVERSIDAD NACIONAL DEL LITORAL

2018

Secretaría de Posgrado, Facultad de Ingeniería y Ciencias Hídricas, Ciudad Universitaria,
Paraje “El Pozo”, S3000, Santa Fe, Argentina

UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas
Instituto de Desarrollo Tecnológico para la Industria Química
Centro de Investigación de Métodos Computacionales
Instituto de Investigaciones en Señales, Sistemas e Inteligencia Computacional

ALGORITMOS AVANZADOS PARA LA DETECCIÓN DEL SÍNDROME DE APNEA-HIPOPNEA OBSTRUCTIVA DEL SUEÑO

Román Emanuel Rolón

Lugar de Trabajo:

sinc(*i*)
Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional
Facultad de Ingeniería y Ciencias Hídricas
Universidad Nacional del Litoral

Director:
Dr. Hugo Leonardo Rufiner — sinc(*i*)-FICH-UNL-CONICET

Co-director:
Dr. Rubén Daniel Spies — IMAL-FIQ-UNL-CONICET

Jurado Evaluador:
Dr. Hugo Aimar — IMAL-FIQ-UNL-CONICET
Dr. Gastón Schlotthauer — IBB-UNER-CONICET
Dr. Marcelo Risk — IMTIB-CONICET
Dr. Juan Carlos Gómez — FCEIA-UNR-CONICET

2018

UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas

Santa Fe, 27 de Marzo de 2019.

Como miembros del Jurado Evaluador de la Tesis de Doctorado en Ingeniería titulada ***"Algoritmos avanzados para la detección del síndrome de apnea-hipopnea obstructiva del sueño"***, desarrollada por el Ing. Román Emanuel ROLÓN, en el marco de la Mención "Inteligencia Computacional, Señales y Sistemas", certificamos que hemos evaluado la Tesis y recomendamos que sea aceptada como parte de los requisitos para la obtención del título de Doctor en Ingeniería.

La aprobación final de esta disertación estará condicionada a la presentación de dos copias encuadernadas de la versión final de la Tesis ante el Comité Académico del Doctorado en Ingeniería.

Dr. Hugo Aimar

Dr. Gastón Schlotthauer

Dr. Marcelo Risk

Dr. Juan Carlos Gómez

Santa Fe, 27 de Marzo de 2019.

Certifico haber leído la Tesis, preparada bajo mi dirección en el marco de la Mención "Inteligencia Computacional, Señales y Sistemas" y recomiendo que sea aceptada como parte de los requisitos para la obtención del título de Doctor en Ingeniería.

Dr. Rubén Spies
Codirector de Tesis

Dr. Leonardo Rufiner
Director de Tesis

Universidad Nacional del Litoral
Facultad de Ingeniería y Ciencias Hídricas

Secretaría de Posgrado

Ciudad Universitaria
C.C. 217
Ruta Nacional Nº 168 - Km. 472,4
(3000) Santa Fe
Tel: (54) (0342) 4575 229
Fax: (54) (0342) 4575 224
E-mail: posgrado@fich.unl.edu.ar

Formato de tesis:

La presente tesis se encuentra organizada bajo el formato de Tesis por Compilación, aprobado en la resolución Nº 255/17 (Expte. Nº 888317-17) por el Comité Académico de la Carrera Doctorado en Ingeniería, Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral (UNL). De dicha resolución:

"*En el caso de optar por la Tesis por Compilación, ésta consistirá en una descripción técnica de al menos 30 páginas, redactada en español e incluyendo todas las investigaciones abordadas en la tesis. Se deberán incluir las secciones habituales indicadas a continuación en la Sección Contenidos de la Tesis. Los artículos científicos publicados por el autor, en el idioma original de las publicaciones, deberán incluirse en un Anexo con el formato unificado al estilo general de la Tesis indicado en la Sección Formato. El Anexo deberá estar encabezado por una sección donde el tesista detalle para cada una de las publicaciones cuál ha sido su contribución. Esta sección deberá estar avalada por su director de Tesis. El documento central de la Tesis debe incluir referencias explícitas a todas las publicaciones anexadas y presentar una conclusión que muestre la coherencia de dichos trabajos con el hilo conceptual y metodológico de la tesis. Los artículos presentados en los anexos podrán ser artículos publicados, aceptados para publicación (en prensa) o en revisión.*"

A Alexia, por su amor y apoyo incondicional,
a Lorenzo y Nicolás que son la luz que ilumina mi vida.

Agradecimientos:

Este es uno de esos momentos donde desbordan las palabras de felicidad, gratitud y satisfacción concentrándose todas ellas en una sola que es ¡¡gracias!!. Han pasado ya varios años desde aquel momento en que *Alexia*, el gran amor de mi vida, me incentivó y brindó todo su apoyo depositando en mi su confianza para que me postulara a una beca interna de doctorado del Consejo Federal de Investigaciones Científicas y Técnicas (CONICET). En primer lugar, quiero darle las gracias a *Alexia* por ser el pilar de mi vida, por su paciencia, por su amor, por su confianza, por su contención y sobre todo por ser la madre de mis dos hijos *Lorenzo* y *Nicolás*, que son el motor de mi vida y llenan mis días de amor. Ellos también han sido fruto de esta tesis ya que ambos nacieron durante el transcurso de la misma.

Quiero agradecer a mis directores, quienes depositaron toda su confianza en mi desde el primer momento, sin dudarlo, y quienes me han acompañado a lo largo de todos estos años tanto en mi formación profesional como en la humana.

La familia es un regalo de Dios y un gran tesoro que tenemos las personas. Es por ello que mis agradecimientos se extienden hacia toda mi familia y amigos que he cosechado a lo largo de toda mi vida.

Nada de esto podría haber sido posible sin la existencia y aval de las instituciones. Por esta razón, quiero agradecer al CONICET por haber financiado mis estudios de postgrado otorgándome una beca interna doctoral de tiempo completo. Además, quiero dar las gracias al Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, sinc(i), por abrirme las puertas y brindarme todo su apoyo para poder investigar. Asimismo, quiero agradecer al Instituto de Matemática Aplicada del Litoral, IMAL, por su colaboración y compromiso.

Índice general

Índice de figuras

Índice de tablas

Resumen

El *Síndrome de Apnea-Hipopnea Obstructiva del Sueño* (SAHOS) es uno de los trastornos del sueño más comunes en la población general que afecta a múltiples órganos. Se estima que esta patología afecta entre el 3 % y 5 % de la población adulta en todo el mundo y aumenta con la edad. Si bien el SAHOS es más frecuente en adultos, impacta también a niños con una prevalencia cercana al 3 %. Los eventos respiratorios asociados al SAHOS durante el sueño ocurren como consecuencia de una alteración anatómico-funcional de la vía aérea superior que producen su estrechamiento parcial (hipopnea) o su bloqueo total (apnea). Para establecer la severidad del SAHOS, se define el *Índice de Apnea-Hipopnea*. Éste índice representa el número de eventos de apnea-hipopnea por hora de sueño. El estudio patrón de oro para el correcto diagnóstico del SAHOS es la *Polisomnografía* nocturna. Dado que este estudio de referencia no solo requiere del registro simultáneo de varias señales fisiológicas, sino también necesita una infraestructura especial y personal calificado, es de difícil acceso y muy costoso en términos de tiempo y dinero.

Esta tesis aborda el diseño, desarrollo, implementación y evaluación de tres métodos especialmente creados para el reconocimiento automático de los eventos de apnea-hipopnea a partir del análisis y procesamiento de señales de saturación de oxígeno en sangre (SaO_2). En particular, se presentan dos métodos de selección de características llamados MDAS y MDCS ambos basados en representaciones ralas de señales de SaO_2. Además, se introduce una nueva medida de discriminabilidad binaria denotada por DCAF, la cual es capaz de cuantificar eficientemente el grado de discriminabilidad de cada uno de los átomos en un diccionario dado. Los métodos MDAS y MDCS hacen uso de la medida DCAF para detectar los átomos más discriminativos de un diccionario dado y, a partir de ellos, realizar la selección de características. En particular, el método MDCS utiliza la medida DCAF para

identificar y extraer los átomos de un diccionario que tienen mayor información en términos de la clasificación para formar un nuevo sub-diccionario discriminativo. En base a los experimentos desarrollados en esta tesis, se comparó el desempeño de la nueva medida DCAF con el de varias otras medidas de información del estado del arte. Los resultados muestran que DCAF logró un muy buen desempeño. Por otro lado, se comparó el desempeño del nuevo método MDCS con otros tres métodos del estado de arte, superando significativamente el desempeño de todos ellos.

Esta tesis también introduce una extensión del problema de clasificación binaria a uno multi-clase. En este contexto, se propone una generalización de la medida DCAF (la cual tiene en cuenta solo dos clases en los datos) a más de dos clases. En particular, la nueva medida de discriminabilidad combinada no solo tiene en cuenta la probabilidad condicional de activación de los átomos en un diccionario dada la clase y el valor de su correspondiente coeficiente de activación, sino que también incorpora el efecto que éste tiene sobre el error total de representación. Asimismo, se presenta un nuevo método iterativo llamado DAS-KSVD para el aprendizaje de diccionarios estructurados en el contexto de problemas de clasificación multi-clase, que utiliza ésta medida. El nuevo método permite detectar los átomos más discriminativos para cada una de las clases. Utilizando una base de datos de dígitos manuscritos ampliamente utilizada en la literatura, se realizó un análisis del desempeño del método DAS-KSVD obteniéndose tasas de reconocimiento superiores a las obtenidas por técnicas semejantes del estado del arte. También se utilizó el nuevo método DAS-KSVD en un problema de clasificación multi-clase asociado al SAHOS. Los resultados muestran que la construcción de diccionarios estructurados de clase específica constituye una técnica muy prometedora para ser usada como soporte al diagnóstico del SAHOS.

Abstract

Obstructive Sleep Apnea-Hypopnea Syndrome (OSAHS) is one of the most common sleep disorders in the general population which involves multiple organs. It is estimated that such a pathology affects between 3 % and 5 % of the adult population around the world and it increases with age. Although OSAHS is more common in adults, it also impacts children with a prevalence close to 3 %. The respiratory events associated with OSAHS during sleeping occur as a consequence of a functional-anatomic disturbance of the upper airway producing its partial obstruction (Hypopnea) or total blockage (Apnea). To establish the severity of the syndrome, the *Apnea-Hypopnea Index* is defined. This index represents the number of apnea-hypopnea events per hour of sleep. The gold standard study for the correct diagnosis of OSAHS is the nocturnal *Polysomnography*. Since this reference study not only requires the simultaneous recording of several physiological signals, but also it needs special infrastructure and qualified personal staff, is difficult to access and very costly in terms of time and money.

This thesis addresses the design, development, implementation and evaluation of three methods specially created for the automatic recognition of apnea-hypopnea events from the analysis and processing of blood oxygen saturation (SaO_2) signals. In particular, two methods for feature selection called MDAS and MDCS both based on sparse representations of SaO_2 signals, are presented. Also, a new binary discriminative measure denoted by DCAF, which is capable of efficiently quantify the degree of discriminability of each one of the atoms in a given dictionary, is introduced. The MDAS and MDCS methods make use of the DCAF measure to detect the most discriminative atoms in a given dictionary and, taking them into account, perform the selection of features. In particular, the MDCS method uses the DCAF measure to identify and extract the dictionary atoms that have the most useful information

in terms of classification to form a new discriminative sub-dictionary. Based on the experiments performed in this thesis, the performance of the new DCAF measure was compared with several other state-of-the-art information measures. The results show that DCAF achieved a very good performance. On the other hand, the new MDCS method was compared with three other state-of-the-art methods, significantly outperforming all of them.

This thesis also introduces an extension of the binary classification problem to a multi-class one. In this context, a generalization of the DCAF measure (which takes into account only two classes in the data) to more than two classes is proposed. In particular, the new combined discriminative measure not only takes into account the conditional probability of activation of atoms in a given dictionary, the class and the value of its corresponding activation coefficient, but also incorporates the effect it has on the total misrepresentation. Likewise, a new iterative method called DAS-KSVD is presented for the learning of structured dictionaries in the context of multi-class classification problems, which uses this measure. The new method allows to detect the most discriminative atoms for each of the classes. Using a database of handwritten digits widely used in the literature, an analysis of the performance of the DAS-KSVD method was carried out, obtaining recognition rates higher than those yielded by similar state-of-the-art techniques. The new DAS-KSVD method was also used in a multi-class classification problem associated to SAHOS. Results show that building class-specific structured dictionaries constitutes a very promising technique to be used as support for OSAHS diagnosing.

1 Introducción

1.1. Descripción del problema

El *Síndrome de Apnea-Hipopnea Obstructiva del Sueño* (SAHOS) es un trastorno del sueño muy común en la población general. Se estima que esta patología, la cual ha sido distinguida entre 81 tipos de trastornos del sueño [1], afecta alrededor del 3 % y 5 % de las mujeres y los hombres adultos, respectivamente, y esta prevalencia aumenta progresivamente con la edad [2]. Si bien el SAHOS es más frecuente en adultos, afecta también a niños con una prevalencia cercana al 3 % [3]. El SAHOS es un cuadro caracterizado por la presencia de eventos respiratorios durante el sueño normal de las personas. Estos eventos ocurren como consecuencia de una alteración anatómico-funcional de las paredes de la vía aérea superior (incluida la lengua) produciendo su estrechamiento (hipopnea) u obstrucción total (apnea). La Figura 1.1 muestra lo que sucede normalmente en la apnea del sueño. En el sueño normal (parte izquierda) se puede observar como el flujo de aire circula libremente a través de la vía respiratoria. Cuando ocurre un evento de hipopnea (parte central), causado por una obstrucción parcial de la vía aérea superior, se produce una vibración (ronquido) en sus tejidos flexibles al paso del aire y se restringe el paso del oxígeno. Por último, cuando ocurre un evento de apnea (parte derecha), el flujo de aire es totalmente bloqueado restringiendo por completo la circulación del oxígeno.

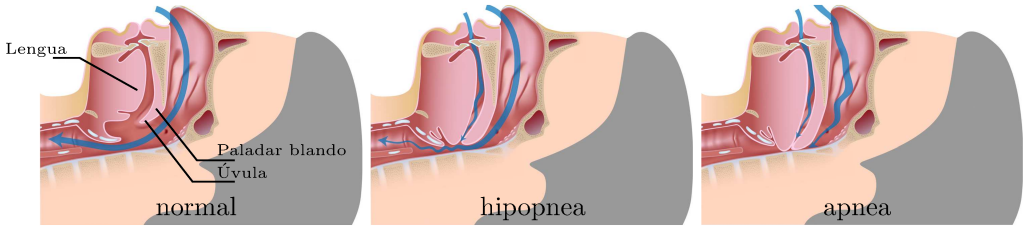


Figura 1.1: Ilustración del SAHOS.

Al finalizar cada evento de apnea-hipopnea (AH), se produce una desaturación de la hemoglobina, la cual origina un patrón característico en los registros de la oxime-

tría de pulso, que genera hipoxemia intermitente. A su vez, los ciclos de hipoxemia-reoxigenación de la hipoxemia intermitente promueven el estrés oxidativo y la angiogénesis, aumentan la activación simpática con elevación de la presión arterial e inflamación sistémica y vascular con disfunción endotelial que contribuye a producir morbilidad crónica multiorgánica, disfunción metabólica, deterioro cognitivo y el aumento tumoral maligno [4]. De hecho, esta relación se ha investigado durante la última década estableciendo una asociación entre la malignidad de diferentes tipos de cáncer y el insomnio o falta de buen sueño [5]. Además, las desaturaciones bruscas y frecuentes durante el sueño asociadas al SAHOS constituyen una de las principales causas de alteraciones cardíacas, cerebrales y metabólicas del SAHOS [6]. El grado de severidad del SAHOS se cuantifica mediante el *Índice de Apnea-Hipopnea* (IAH), el cual representa la tasa de eventos de AH por hora de sueño. Dependiendo del valor de tal índice, el SAHOS se clasifica en *normal* si el IAH es menor (o igual) a 5, *leve* si el IAH es mayor a 5 y menor (o igual) a 15, *moderado* si el IAH es mayor a 15 y menor (o igual) a 30 o *severo* si el IAH es mayor a 30 [3].

El estudio de referencia para el correcto diagnóstico del SAHOS es la *Polisomnografía* (PSG) nocturna [3, 7]. La PSG consiste en la medición simultánea de variables fisiológicas tales como señales electroencefalográficas (EEG), electrocardiográficas (ECG), electrooculares (EOG), electromiográficas (EMG), esfuerzo respiratorio, flujo respiratorio y saturación de oxígeno en sangre (SaO_2), entre muchas otras (ver Figura 1.2). Dada su notable complejidad intrínseca, se ha demostrado que la PSG dificulta en gran medida el sueño normal de las personas [8]. Por otro lado, la PSG requiere de una infraestructura especial (unidades de sueño) y de personal calificado (técnicos y médicos especialistas), lo cual hace que sea muy costosa en términos de su accesibilidad, como así también en tiempo y dinero.

Como se mencionó anteriormente, el SAHOS es una patología muy prevalente en la población general. En una muestra de 602 personas de ambos sexos que trabajan, con edades entre 30 y 60 años, se determinó que el 24 % y 9 % de los hombres y las mujeres, respectivamente, padecen de un SAHOS leve [9]. Asimismo, se ha establecido que el sexo masculino [10], los antecedentes familiares de trastornos del sueño y ronquido [11] y la obesidad [12], son los principales factores de riesgo para el desarrollo del síndrome. Además, el índice de masa corporal, el incremento de la edad, la menopausia y el consumo de alcohol, entre muchos otros, son también

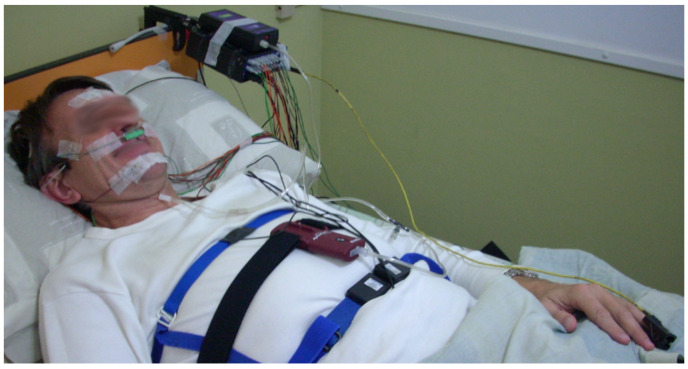


Figura 1.2: PSG nocturna.

factores de riesgo que promueven la patología.

El SAHOS es un problema de salud pública con un impacto negativo en la calidad de vida de las personas y con consecuencias clínicas severas. Este síndrome tiene graves repercusiones socio-sanitarias, laborales y está relacionado con los accidentes de tránsito. Existe evidencia científica que sostiene que el SAHOS no tratado está directamente relacionado con el desarrollo de hipertensión arterial [13] y la progresión de enfermedades cardiovasculares y cerebrovasculares [6]. También se vincula a este síndrome con un aumento de la morbimortalidad asociada no sólo a complicaciones cardiorrespiratorias y neurológicas, sino también con accidentes de tránsito [14]. En definitiva, el SAHOS promueve significativamente el deterioro de la calidad de vida de las personas.

Dada la elevada prevalencia del SAHOS, la medicina de atención primaria es determinante en la identificación de los pacientes que lo sufren y, por lo tanto, necesita de medios diagnósticos confiables, sencillos y de bajo costo. Un adecuado uso de los equipos simplificados correctamente validados permitiría, seleccionando los casos, descentralizar el diagnóstico de las unidades de referencia habitualmente saturadas, hacia unidades de diagnóstico más pequeñas abastecidas con equipos simplificados [3]. Pero esta descentralización del proceso diagnóstico deberá ser acompañada de una apropiada formación del personal y una suficiente coordinación con las unidades de sueño de referencia, para el estudio de los casos difíciles o dudosos [15]. Se deberían organizar redes de complejidad creciente que permitan la rápida consulta con un experto en medicina del sueño y la posibilidad de realizar de ser necesario una PSG para el diagnóstico y tratamiento de este verdadero problema de salud pública

que es el SAHOS [15]. El diseño de equipos de diagnóstico que puedan ser manejados por personal no experto para la búsqueda de pacientes graves que al identificarse más rápidamente puedan tener acceso a un correcto tratamiento es prioritario.

1.2. Eventos respiratorios

Con el propósito de lograr una interpretación precisa y, de esa manera, poder clasificar los eventos respiratorios identificados durante una PSG, la Academia Americana de Medicina del Sueño (AASM, del inglés *American Academy of Sleep Medicine*) ha elaborado un manual de reglas, terminología y especificaciones técnicas destinadas a sistematizar el procedimiento [16, 17]. A continuación se describen los criterios principales que utilizan los médicos especialistas para identificar y etiquetar los eventos de apnea y de hipopnea.

- Evento de apnea:

    1. Reducción en la señal de flujo respiratorio mayor a 90 % respecto del registro basal.
    2. Duración de al menos 10 segundos.
    3. La reducción en la amplitud de la señal de flujo respiratorio debe mantenerse por lo menos durante el 90 % del evento.

- Evento de hipopnea:

    1. Reducción en la señal de flujo respiratorio mayor a 50 % respecto del registro basal.
    2. Duración de al menos 10 segundos.
    3. La reducción en la amplitud de la señal de flujo respiratorio debe mantenerse por lo menos durante el 90 % del evento.
    4. Debe producirse una reducción en la señal de SaO_2 de al menos 3 % respecto del registro basal.

La Figura 1.3 muestra una pequeña porción de la representación gráfica temporal de dos señales biomédicas provenientes de la PSG. La curva representada en

color azul (arriba) corresponde a la señal de flujo respiratorio, mientras que la curva trazada en color rojo (abajo) representa la señal de SaO_2. Asimismo, se han distinguido las etiquetas de los eventos de apnea y de hipopnea en ambas gráficas. Como se puede apreciar en esta figura, existe una relación causal entre la aparición de los eventos de AH, marcados en base a la señal de flujo respiratorio, y la reducción del nivel de oxígeno en sangre. Sin embargo, el intervalo de tiempo para el cual el flujo de aire se encuentra bloqueado y el comienzo de la desaturación en la señal de SaO_2 es muy variable. Además, la duración de los eventos de AH son también variables. Por otro lado, dado que existe la posibilidad de que ocurran eventos de AH que estén acompañados por desaturaciones que no sean fácilmente detectables por el ojo humano, los algoritmos inteligentes pueden ser de gran ayuda para llevar a cavo la detección de los eventos de AH.

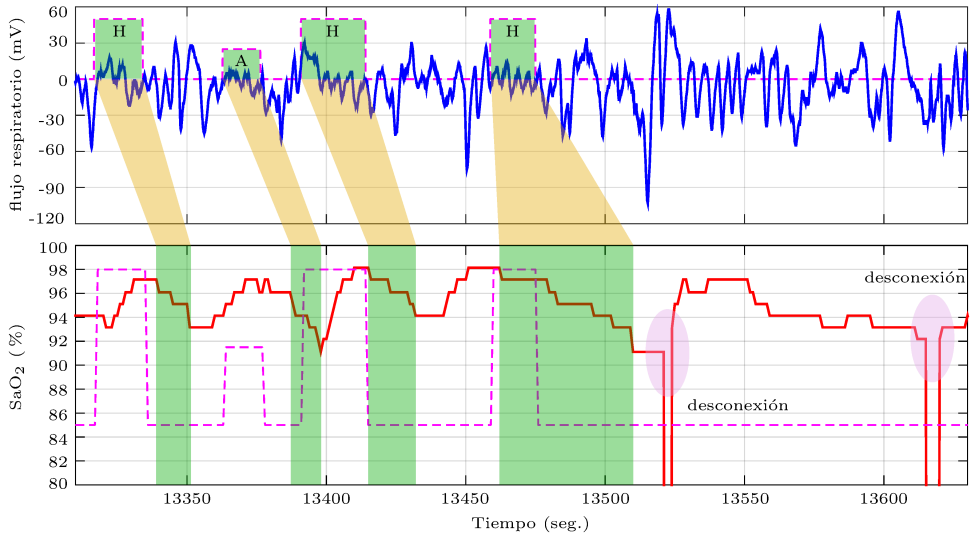


Figura 1.3: Representación gráfica de una porción de señal de flujo respiratorio (arriba) y de la señal de SaO_2 (abajo). Además, se resaltan los eventos de apnea y de hipopnea que fueron marcados por el médico experto en la señal de flujo respiratorio y las correspondientes desaturaciones en la señal de SaO_2.

1.3. Métodos de diagnóstico

Considerando la terminología usada por el consenso argentino de trastornos respiratorios vinculados al sueño [18], los métodos de diagnóstico del SAHOS se clasifican en cuatro niveles. Esta clasificación ha sido diseñada en orden de complejidad decreciente comenzando por el nivel 1 y finalizando en el nivel 4, los cuales corresponden a estudios completos de referencia (PSG nocturna) y a estudios simplificados, respectivamente. A continuación se describen brevemente las características principales de cada uno de los métodos de diagnóstico del SAHOS.

1.3.1. Polisomnografía en una unidad de sueño (nivel 1)

Este estudio es el patrón de oro para el correcto diagnóstico del SAHOS [3, 18]. La PSG es realizada en general por la noche, salvo en casos excepcionales en que los pacientes trabajen durante la noche o en turnos rotativos. Como se mencionó anteriormente, la PSG consiste en obtener registros simultáneos de una gran variedad de señales biomédicas durante el sueño normal del paciente.

El análisis combinado de todo el grupo de señales biomédicas registradas durante la PSG permite que médico especialista pueda realizar un buen diagnóstico del paciente y, por lo tanto, determinar la presencia (o no) del SAHOS. Asimismo, se utilizan distintos sub-grupos de señales que permiten, por ejemplo, determinar la vigilia y reconocer las distintas etapas del sueño [16]. Para esta tarea se utiliza en general el sub-grupo compuesto por las señales EEG, EOG y EMG.

Las señales de flujo respiratorio son registradas por sensores de flujo aéreo que pueden ser termistores, o bien cánulas nasales. Los termistores registran las diferencias de temperatura entre el aire inspirado y el espirado, mientras que las cánulas nasales detectan la limitación al flujo inspiratorio en una curva de presión. Los movimientos respiratorios se detectan mediante el uso de bandas ubicadas en el tórax y el abdomen del paciente. La SaO_2 se evalúa a través de un oxímetro de pulso, el cual es ubicado generalmente en el dedo índice de la mano. Los ronquidos son captados por un micrófono y por el análisis de la señal ECG, y en modo simultáneo, se controla la posición del paciente en la cama mediante un sensor de posición o a través del registro manual del personal técnico.

1.3.2. Polisomnografía domiciliaria (nivel 2)

Este estudio es una PSG completa realizada en el domicilio del paciente. A diferencia de la PSG en una unidad de sueño, la PSG domiciliaria no requiere de la presencia física del personal técnico mientras se realiza adquisición de datos. Además, este estudio registra las mismas señales que el nivel 1.

1.3.3. Poligrafía (nivel 3)

Dada la elevada prevalencia del SAHOS, el elevado costo del estudio de referencia para el correcto diagnóstico y las unidades de sueño colapsadas, surge la necesidad de desarrollar nuevas tecnologías más "simples" que permitan diagnosticar esta patología tanto en personas adultas como en niños. La *Poligrafía* (PLG) respiratoria es precisamente un estudio más simple que la PSG ya que requiere el registro simultáneo de un menor número de señales biomédicas y se puede realizar en el domicilio del paciente sin necesidad de la presencia física del personal técnico [19].

La PLG realizada en casa permite que el paciente tenga un sueño más fisiológico. Este estudio tiene la desventaja de que, al no registrarse señales EEG, no es posible evaluar los estados del sueño, detectar micro-despertares y conocer el tiempo real de sueño, entre otras. En estos casos lo que se realiza habitualmente es considerar el tiempo total de sueño como la duración completa del registro para construir los diversos índices. Notar que esta consideración acerca del tiempo total de sueño podría subestimar el IAH del paciente generando falsos negativos.

Si bien se ha encontrado una alta correlación entre los valores de los AHIs generados por PSGs y PLGs y, la PLG no requiere del registro de señales EEG, éste método aún requiere del registro de otras señales biomédicas cuya adquisición afecta el sueño normal de las personas. Por lo tanto, el futuro de las técnicas de diagnóstico simplificado del SAHOS está fuertemente orientado al desarrollo de nuevas tecnologías de *no contacto* (o contacto mínimo) que utilice el menor número posible de señales fisiológicas. Dado que la oximetría de pulso y, más recientemente, el sonido traqueal aportan al desarrollo de nuevas técnicas confiables, de fácil acceso, económicas y fundamentalmente no invasivas, se han convertido en alternativas muy valiosas para ser utilizadas en el diagnóstico simplificado del SAHOS [20, 21, 22].

1.3.4. Estudios simplificados (nivel 4)

Consiste en el registro continuo de a lo sumo dos señales cardiorrespiratoria durante todo el período de sueño del paciente. Las señales que se utilizan comúnmente para el desarrollo de este tipo de métodos de diagnóstico son señales de flujo respiratorio, de ronquido, de SaO_2 y de sonido traqueal [20, 21, 22, 23]. Es importante señalar que, en estos tipos de estudios, no se tiene información certera acerca del tiempo de sueño del paciente para la obtención del IAH debido a que no se registran señales EEG. Por lo tanto, la estimación de tal índice se realiza en base a la duración total del registro (en horas), en lugar de utilizar el tiempo (real) de sueño. En este contexto, se han propuesto distintas alternativas para determinar cuando un paciente esta despierto o dormido a partir de señales de SaO_2, como por ejemplo en [24].

La oximetría de pulso es una técnica simple, económica, no invasiva y de fácil acceso que permite medir la desaturación de oxígeno de la hemoglobina en sangre. Esta técnica consiste esencialmente en un dispositivo similar a un clip, generalmente llamado sonda, el cual es colocado en una parte del cuerpo de la persona. Este dispositivo se coloca usualmente en el dedo índice de la mano (Figura 1.4) o también en el lóbulo de la oreja. La sonda emite una luz que permite medir los niveles de oxigeno en la sangre.

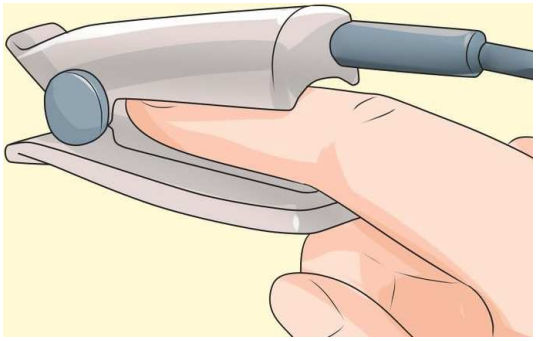


Figura 1.4: Oximetría de pulso.

1.4. Objetivos de la tesis

Teniendo en cuenta lo presentado previamente, en esta tesis se plantea la búsqueda de nuevas técnicas de procesamiento de señales que permitan realizar el diagnóstico simplificado del SAHOS.

La principal contribución realizada consiste en el desarrollo de métodos de selección de características discriminativas basados en técnicas novedosas de procesamiento de señales que permiten la detección automática de los eventos de AH empleando sólo las señales de SaO_2.

En base a esto, los objetivos perseguidos en esta tesis se listan en los siguientes puntos:

- Detectar los eventos de AH empleando sólo las señales de SaO_2.
- Proponer nuevas medidas de discriminabilidad para la selección de características.
- Seleccionar información relevante para la detección de los eventos de AH.
- Desarrollar nuevos métodos discriminativos para el diagnóstico simplificado del SAHOS.
- Usar bases de datos con estudios reales para la aplicación y validación de los métodos desarrollados.
- Analizar y validar los resultados.
- Comparar los métodos propuestos con otros métodos del estado del arte.

1.5. Antecedentes

Como se mencionó en la Sección 1.3, el estudio de referencia para el correcto diagnóstico del SAHOS es la PSG en una unidad de sueño. Sin embargo, dada la elevada prevalencia de esta patología, las unidades de sueño existentes se encuentran desbordadas por la excesiva demanda de estudios. Esto hace que muy pocas personas puedan acceder a este tipo de estudio, el cual es además muy costoso y términos de

tiempo y dinero. Por esta razón, se han desarrollado numerosos métodos alternativos para el diagnóstico simplificado del SAHOS que utilizan un número reducido de señales biomédicas como, por ejemplo, señales de flujo respiratorio [25, 26], ECG [27], de sonido traqueal [21] y de SaO_2 [20, 22], entre otras.

En particular, la oximetría de pulso es una técnica utilizada en numerosas áreas de la medicina ya que permite determinar de forma sencilla y no invasiva los niveles de saturación de oxígeno en la sangre y la frecuencia cardíaca de las personas. Teniendo en cuenta los valores de SaO_2 en la señal de oximetría de pulso, se pueden construir diferentes índices de desaturaciones denotados por ODI4, ODI3 y ODI2 (ODI, del inglés *Oxygen Desaturation Index*), los que representan el número de veces por hora de sueño que la señal cae por debajo del 4 %, 3 % y 2 % del nivel medio (referencia), respectivamente [28]. Es muy importante señalar que, si bien, el concepto de nivel medio es muy intuitivo, este carece de una única definición ya que existen autores que lo definen de forma diferente [29, 30, 31]. Esto se debe en mayor medida a las tendencias típicas que se observan en el registro de las señales de SaO_2 y, en menor medida, a las desconexiones del oxímetro de pulso.

En particular, la búsqueda de personas sospechosas de padecer el SAHOS se puede abordar mediante dos tipos de enfoques. Un enfoque *global* consiste en obtener características generales de la señal de SaO_2, como por ejemplo su valor medio, varianza y entropía, entre otros, con el único objetivo de determinar si una persona está sana o enferma, independientemente del grado de severidad de la patología. Por otro lado, se puede abordar un enfoque más bien *local*, el cual permite realizar un análisis más profundo del grado de severidad del SAHOS [23, 20]. En esta tesis se aborda este enfoque mediante la detección de los eventos de AH a partir del uso de técnicas avanzadas de procesamiento de señales de SaO_2.

En los últimos años se ha observado un creciente interés en la investigación y el desarrollo de nuevos dispositivos portátiles que permitan ayudar al médico especialista en el diagnóstico simplificado del SAHOS. En este sentido, se introdujeron una gran variedad de métodos basados en técnicas de Aprendizaje Maquinal y de redes neuronales artificiales. Asimismo, a partir de la implementación de estas técnicas, se crearon nuevos métodos fundados en el Análisis no Lineal [32, 33], Estadística de Alto Orden, Análisis Espectral [34], Transformada Wavelet Discreta (DWT del inglés *Discrete Wavelet Transform* [35], Análisis de Componentes Independientes y

Descomposición Empírica en Modos [36, 23], entre muchos otros. Además, se desarrollaron diversos métodos para el Reconocimiento de Patrones y se aplicaron con éxito para asistir al médico especialista en la detección automática de eventos de AH [37, 20].

En la actualidad resulta de gran interés la investigación y desarrollo de nuevos métodos de aprendizaje maquinal basados en una técnica novedosa de procesamiento de señales llamada Representaciones Ralas (*Sparse Representations*) [38, 39, 40]. Esta técnica tiene como objetivo producir soluciones que corresponden a la representación más compacta, mediante la combinación lineal de unas pocas formas de onda básicas (átomos) tomadas de un conjunto grande (diccionario). Unas de las principales ventajas que ofrece este tipo de representaciones son su gran robustez al ruido y la reducción de la dimensión del problema, entre otras. En particular, se ha encontrado que la incorporación de estos aspectos en modelos de los sistemas sensoriales biológicos genera representaciones internas con propiedades similares a las de los sistemas reales, en particular a las encontradas en la corteza auditiva o visual primaria de los mamíferos [41, 42]. Sin embargo, no se han encontrado aplicaciones de esta poderosa técnica de procesamiento de señales para la detección de eventos de AH o, más precisamente, para el diagnóstico del SAHOS.

1.6. Descripción general

Los modelos considerados y desarrollados en la presente tesis se basan fundamentalmente en métodos de reconocimiento automático de patrones, los cuales tienen como único objetivo la clasificación de los datos (señales) de entrada en un conjunto de clases o categorías específicas usando información relevante proveniente de sus propiedades y características. Más precisamente, estos métodos permiten detectar en forma automática los eventos de AH mediante el análisis y procesamiento de las señales obtenidas a partir de la oximetría de pulso.

Los métodos propuestos en la presente tesis pueden pensarse como un conjunto de bloques (o etapas) generales relacionados en forma estratégica que permiten el análisis y procesamiento de las señales, donde se resalta su información discriminativa, y el reconocimiento de los patrones (eventos) correspondientes. La Figura 1.5 muestra un esquema general del procedimiento abordado para el diagnóstico sim-

plificado del SAHOS. Se puede observar que las señales de SaO_2 son inicialmente acondicionadas (bloque i)). El acondicionamiento de las señales se realiza mediante un filtrado adecuado seguido de un proceso de segmentado (más detalles en la Sección 3.3). El análisis basado en diccionarios discretos (bloque ii)) juega un papel

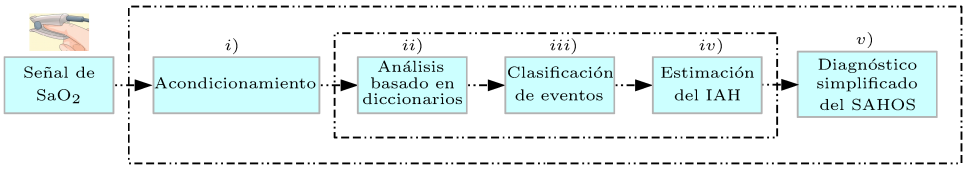


Figura 1.5: Procedimiento para el diagnóstico simplificado del SAHOS a partir de la detección automática de los eventos de AH.

crucial en el desarrollo de los métodos discriminativos propuestos en esta tesis. Si bien la representación rala de señales en términos de un diccionario discreto consiste en representar (en términos de su forma de onda) una señal particular mediante la combinación lineal de sólo unas pocas funciones base (llamadas átomos) del diccionario, en problemas de clasificación resulta de gran interés poder "cuantificar" el grado de información discriminativa que contiene cada uno de los átomos que lo conforman. Por esta razón, es necesario definir medidas que permitan cuantificar la "discriminabilidad" de los átomos. Las Secciones 2.4 y 3.2 introducen las medidas utilizadas en los experimentos desarrollados en esta tesis y la nueva medida propuesta para cuantificar el grado de discriminabilidad de los átomos, respectivamente. Durante la tarea de clasificación (bloque iii)), se considera un subconjunto de características discriminativas, seleccionadas apropiadamente a partir del análisis basado en diccionarios discretos, como entrada de un clasificador entrenado para detectar la presencia de eventos de AH en segmentos de señales de SaO_2. Además, a partir de los eventos de AH detectados, se procede a estimar el IAH (bloque iv)) y, por lo tanto, realizar el diagnóstico simplificado del SAHOS (bloque v)).

Esta tesis aborda el desarrollo de nuevos métodos que permiten resolver problemas de clasificación tanto binaria como multi-clase. En particular, en el contexto de problemas de clasificación binaria, se desarrollaron dos métodos para la selección de átomos discriminativos a partir de diccionarios redundantes previamente aprendidos (Capítulo 3). Además, en el contexto de problemas de clasificación multi-clase, se desarrolló un método iterativo para el aprendizaje de diccionarios estructurados

discriminativos (Capítulo 4). En cuanto a los dos primeros métodos, la Figura 1.6 muestra un esquema simplificado del proceso de construcción de sub-diccionarios mediante la selección de los átomos más discriminativos de un diccionario redundante. En este caso, el análisis basado en diccionarios discretos consiste inicialmente

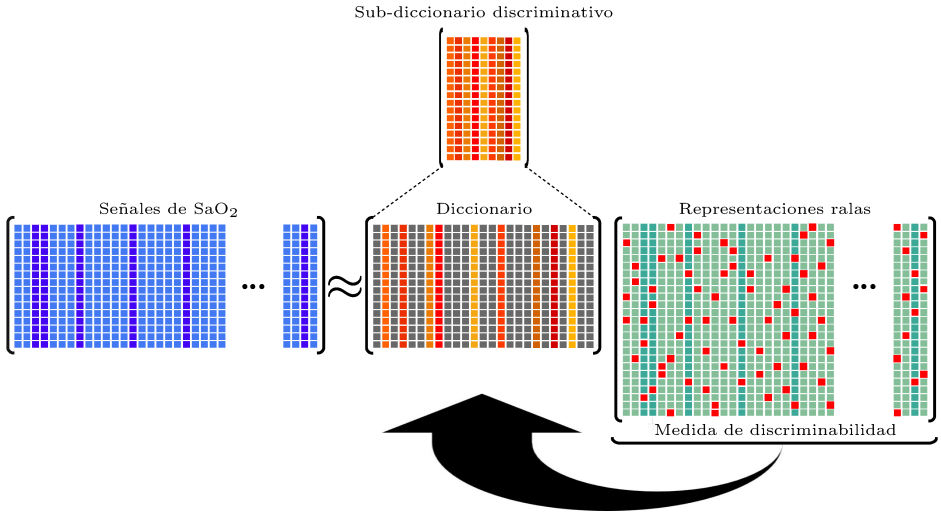


Figura 1.6: Esquema del proceso de selección de átomos discriminativos mediante el uso de medidas de discriminabilidad.

en obtener las representaciones ralas de todas las señales en términos de un diccionario que fue aprendido (o construido) previamente a partir de los mismos datos. Notar que, dado un conjunto de señales de SaO_2 acondicionadas de forma apropiada y con etiquetas de eventos de AH (resaltadas en color azul fuerte) y normales (sin eventos), el análisis propuesto consiste inicialmente en obtener todas las representaciones ralas de las señales en términos de un diccionario aprendido (o construido) previamente a partir de los datos disponibles. Además, a partir de las representaciones ralas de las señales y teniendo en cuenta la clase a la que éstas pertenecen, se procede a cuantificar el grado de información discriminativa que posee cada uno de los átomos del diccionario mediante el uso de "medidas de discriminabilidad". Finalmente, se procede a construir un sub-diccionario cuyos átomos son justamente los que aportan mayor información relevante respecto a las clases en los datos.

Por otro lado, la Figura 1.7 muestra el esquema propuesto para el aprendizaje de diccionarios estructurados. La idea principal de este nuevo método es encontrar un diccionario estructurado que este conformado por sub-diccionarios de clase espe-

cífica, es decir, que cada sub-diccionario contenga átomos que aporten información relevante para la clasificación sólo para una clase en particular. Para ello, en cada iteración del método se selecciona un grupo pequeño balanceado de señales etiquetadas y se aprende un diccionario. La detección de los átomos discriminativos (un átomo por clase) se realiza haciendo uso de una nueva medida de discriminabilidad combinada. Esta medida analiza la representación rala de los datos y determina cuando un átomo contiene información relevante de una clase en particular respecto a las otras. Es importante notar que, los átomos discriminativos del diccionario se seleccionan y se apilan en uno nuevo de forma estructurada. Finalmente, el proceso se repite realizando un nuevo muestreo de las señales etiquetadas y aprendiendo un nuevo diccionario. Para más detalles acerca de la nueva medida y del nuevo método propuestos en esta tesis, se remite al lector al Capítulo 4.

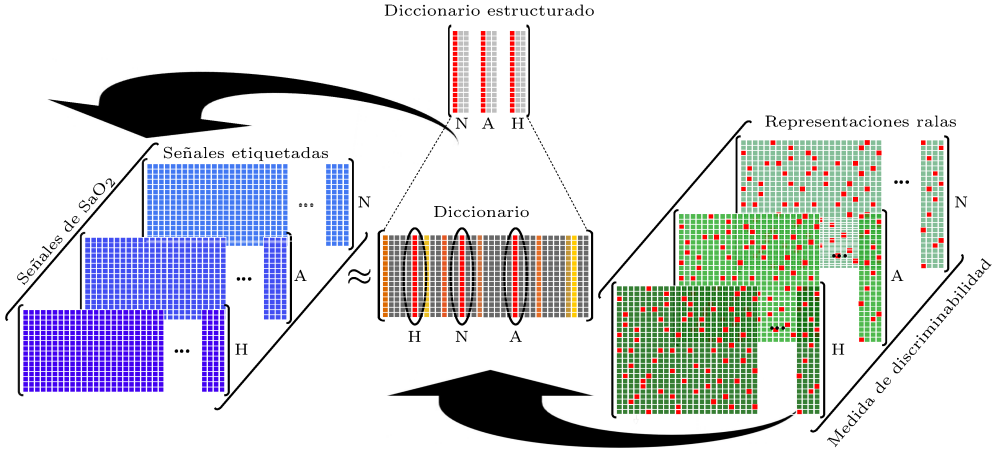


Figura 1.7: Esquema del proceso iterativo de aprendizaje de diccionarios estructurados discriminativos.

1.7. Organización de la tesis

La presente tesis se encuentra organizada en 5 capítulos mediante el formato de "Tesis por Compilación" de la siguiente manera:

- En el presente Capítulo se describió la patología que constituye la problemática a abordar. Además, se introdujeron los dos tipos de eventos respiratorios de interés, la clasificación de los métodos de diagnóstico del SAHOS y se detallaron

los objetivos perseguidos en esta tesis. Además, se presentaron los antecedentes de la investigación que incluyen un resumen de técnicas, métodos y resultados del estado del arte, una descripción de la metodología utilizada y finalmente se presenta la estructura de los capítulos.

- En el Capítulo 2 se describen los materiales y métodos que se utilizarán para lograr la detección automática de eventos respiratorios y, por lo tanto, poder estimar el grado de severidad del SAHOS. En primer lugar, se presentan brevemente los métodos de aprendizaje maquinal utilizados durante las fases experimentales. Luego se introduce la base de datos utilizada tanto para entrenar como para probar los métodos desarrollados en esta tesis y las medidas de desempeño usadas para evaluar el rendimiento del clasificador. Finalmente se introducen las medidas de complejidad.

- En el Capítulo 3 se presenta brevemente la técnica de Representaciones Ralas de señales en términos de diccionarios. Además, en este contexto, se introduce una nueva medida de discriminabilidad que permite cuantificar el grado de información discriminativa que tienen los átomos de un diccionario en problemas de clasificación binaria. Asimismo, introducen los dos métodos propuestos para el reconocimiento de eventos de AH que usan tal medida para la selección de características.

- En el Capítulo 4 se presenta una extensión de la medida de discriminabilidad binaria a problemas de clasificación multi-clase. Además, se introduce un nuevo método iterativo de aprendizaje de diccionarios estructurados. En una primer instancia el método se evalúa con una base de datos muy conocida de dígitos manuscritos y, posteriormente, se lo prueba para la detección de eventos de apnea y de hipopnea en forma separada.

- En el Capítulo 5 se presentan las conclusiones generales y particulares de los métodos desarrollados, la lista de publicaciones científicas logradas como fruto de esta tesis y los trabajos futuros.

2 Materiales y métodos

2.1. Técnicas de aprendizaje maquinal

A continuación se describen brevemente las técnicas de aprendizaje maquinal evaluadas durante el desarrollo de los métodos introducidos en la presente tesis. Estas técnicas se seleccionaron principalmente debido a su gran capacidad de modelado y su bajo costo computacional. Esta última característica es muy importante ya que permite una ejecución en tiempo real.

Perceptrón multi-capa

El perceptrón multi-capa (MLP, del inglés *Multilayer Percetron*), es un tipo de red neuronal artificial que puede operar con datos no separables linealmente [43]. Consiste en varias capas de nodos (perceptrones simples), donde cada capa está completamente conectada a la siguiente, pero no existen conexiones entre los nodos de una misma capa. La salida de un nodo (y) es la suma ponderada (w_i) de las entradas (x_i) más un término de bias (θ), y está afectada por la función de activación (f) de la forma:

$$y = f\left(\sum_{i=1}^{n} w_i x_i + \theta\right). \tag{2.1}$$

Generalmente $f(\cdot)$ es una función lineal para la capa de entrada, y no lineal (función sigmoidea o tangente hiperbólica) para las capas siguientes.

A diferencia de un árbol de decisión convencional, una red neuronal tiene fronteras de decisión más suaves, dando lugar a mejores modelos para la toma de decisiones. El costo computacional de la tarea de clasificación con una red neuronal entrenada es fija y requiere de un número determinado de multiplicaciones y sumas que están dados por el número de capas, la cantidad de neuronas en cada capa, la dimensión de la característica de entrada y el número de salidas.

Máquinas de soporte vectorial

Una máquina de soporte vectorial (SVM, del inglés *Support Vector Machine*), ha probado ser una de las mejores técnicas de aprendizaje maquinal para problemas de clasificación binaria [44, 45]. Primero, las características de entrada no están mapeadas linealmente en un espacio de características de alta dimensionalidad. El mapeo se realiza haciendo uso de una función de *núcleo*, entonces se construye una superficie de decisión lineal en este nuevo espacio de características. La superficie de decisión es un hiperplano y este está localizado en forma de maximizar el margen de separación entre las clases [46].

Las SVM se parecen a una función de base radial (RBF, del inglés *Radial Basis Function*), en el sentido de que esta mapea vectores de características en un espacio de dimensionalidad más alta, pero difiere en el aprendizaje de la frontera de decisión. Las SVM eligen sus vectores de soporte de una manera supervisada para definir las fronteras de decisión, en tanto una RBF define sus fronteras de decisión de una manera no-supervisada, encontrando los centroides RBF.

Máquinas de aprendizaje extremo

Una máquina de aprendizaje extremo (ELM, del inglés *Extreme Learning Machine*), es un tipo de red neuronal recientemente propuesta que cuenta con un buen desempeño de generalización y un corto período de entrenamiento [47]. Su arquitectura es similar a la de un MLP pero con diferencias sobre el típico número de unidades ocultas y sobre el modo de entrenamiento de la red. Los pesos entre la capa de entrada y la capa oculta (W_1) son inicializados aleatoriamente y nunca actualizados durante la vida de la red. Esto es similar a las unidades RBF, las cuales aprenden de una forma no supervisada y nunca se actualizan. Los pesos de los perceptrones de salida de una ELM (W_2) son aprendidos en un paso simple resolviendo el problema inverso regularizado:

$$W_2 = \left[\frac{1}{C}I + (W_1X)^T(W_1X)\right]^+ (W_1X)^TY, \tag{2.2}$$

donde Y es un vector con todas las salidas entrenadas, X es la matriz con todos los ejemplos entrenados y C el parámetro de regularización [48]. El número

de unidades ocultas es generalmente grande para mapear las entradas en un espacio de dimensión más grande, así se requiere más calculo que un MLP durante la clasificación.

2.2. Bases de datos

El grupo de señales biomédicas utilizadas para desarrollar y validar los métodos propuestos en esta tesis se obtuvo a partir de una base de datos muy conocida denominada SHHS-2 (del inglés *Sleep Heart Health Study*)[49, 50]. Esta base de datos fue originalmente creada con el fin de estudiar las correlaciones entre los trastornos respiratorios del sueño y las enfermedades cardiovasculares. La base de datos completa consta de 995 estudios de PSG. Asimismo, durante el estudio de cada PSG, se registraron un gran número de señales fisiológicas como por ejemplo señales electroencefalográficas (EEG), electrocardiográficas (ECG), flujo respiratorio y SaO_2, entre otras. Además, se incluyen anotaciones de los estados del sueño y de eventos de apnea y de hipopnea realizadas por los médicos expertos. Sin embargo, en esta tesis se tienen en cuenta únicamente las señales de SaO_2 con sus respectivas anotaciones de los eventos de apnea y de hipopnea. Por otro lado, es importante señalar que la base de datos utilizada en esta tesis se encuentra disponible en forma gratuita en el sitio `https://physionet.org/physiobank/`.

Para poder validar los métodos de procesamiento de señales propuestos en esta tesis, se requiere de un adecuado acondicionamiento previo de las señales de SaO_2, el cual consta básicamente de dos partes: *i*) el filtrado y *ii*) el segmentado. A continuación se describen brevemente cada una de ellas.

Filtrado: el oxímetro de pulso aporta una estimación no invasiva de la SaO_2 de la hemoglobina, variable que está directamente relacionada con el contenido de oxígeno en sangre arterial. Sin embargo, el registro de la señal de SaO_2 suele degradarse debido a los movimientos del paciente, la tendencia del nivel de la media, las desconexiones y la resolución limitada del oxímetro de pulso, entre otros factores. Además, como consecuencia de las desconexiones (ausencia de señal del sensor), en ese intervalo de tiempo se produce una importante pérdida de información de los valores de SaO_2. Si bien existen muchas formas de solucionar esta falla, en esta tesis se soluciona mediante la unión de los valores de SaO_2 antes y después de cada

desconexión mediante una linea recta, es decir, se realiza una interpolación lineal.

El filtrado de la señal de SaO_2 se realiza mediante la descomposición diádica de la DWT [51]. Durante su filtrado, la señal completa es segmentada mediante la aplicación de una ventana tipo "Boxcar" de longitud 256 que avanza progresivamente a un paso de 32 muestras. La eliminación del ruido de alta frecuencia se logra descartando el coeficiente de aproximación en el nivel 8 de la descomposición DWT con la ondita madre Daubechies 2. Asimismo, se elimina la tendencia del nivel medio y el ruido de baja frecuencia descartando los tres primeros coeficientes de detalle. Este proceso tiene el efecto de aplicar un filtro pasa-banda donde no sólo se elimina el ruido de baja y alta frecuencia, sino también el ruido de cuantificación y la tendencia del nivel de la media de la señal de SaO_2.

Segmentado: luego de ser filtrada mediante la descomposición diádica de la DWT, la señal de SaO_2 es segmentada apropiadamente con el fin de poder utilizar la técnica de codificación rala. Para ello, se extraen segmentos de longitud $N = 128$ con un grado de solapamiento de 75 % entre dos segmentos consecutivos y se apilan en forma de vectores columna $\mathbf{x}_i$ (ver Figura 2.1). Notar que la longitud de los segmentos representa 128 segundos del registro de la oximetría de pulso durante el sueño. Debido a que los eventos de AH pueden durar desde 10 segundos hasta más de 1 minuto, se fijo la dimensión de los segmentos en 128, es decir, cada segmento contiene 128 segundos del registro de la señal de SaO_2.

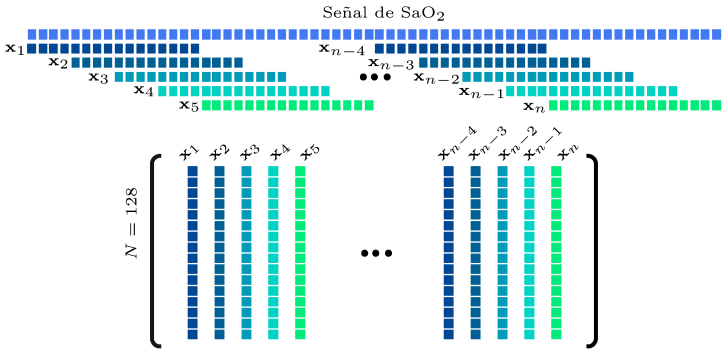


Figura 2.1: Esquema del proceso de segmentado de la señal de SaO_2.

2.3. Medidas de desempeño

A continuación se presentan las medidas de desempeño utilizadas en esta tesis para evaluar la capacidad del clasificador propuesto tanto en el reconocimiento de eventos de AH como en la detección de pacientes sospechosos de padecer un SAHOS moderado. Sin embargo, antes de introducirlas es necesario recordar las diferentes posibilidades que pueden darse al evaluar la presencia (o no) de los eventos de AH y del SAHOS moderado. Las Tablas 2.1 y 2.2 muestran un resumen de las distintas posibilidades que se dan al evaluar un clasificador en la detección de eventos de AH y de SAHOS moderado, respectivamente. Las columnas de estas tablas representan el

Tabla 2.1: Indicadores de desempeño del clasificador en el reconocimiento de los eventos de AH.

Detección	Evento de AH	
	Presente	Ausente
Positivo	Verdadero positivo (TP)	Falso positivo (FP)
Negativo	Falso negativo (FN)	Verdadero negativo (TN)

verdadero estado del evento de AH (o de la patología), el cual se ha determinado (sin errores) mediante el método patrón de oro o *gold standard*. Por otro lado, las filas de ambas tablas representan las detecciones (positivas y negativas) del clasificador propuesto.

Tabla 2.2: Indicadores de desempeño del clasificador en la detección del SAHOS moderado.

Prueba diagnóstica	SAHOS moderado	
	Presente	Ausente
Positivo	Verdadero positivo TP)	Falso positivo (FP)
Negativo	Falso negativo (FN)	Verdadero negativo (TN)

A raíz de las distintas posibilidades que pueden ocurrir al evaluar los métodos desarrollados en esta tesis para el reconocimiento de eventos de AH, o bien para la detección del SAHOS moderado, se define la medida de sensibilidad (SE), también

conocida como valor predictivo positivo o “recall”, como:

$$\mathrm{SE}(\,\%) \doteq \frac{\mathrm{TP}}{\mathrm{TP+FN}} \times 100\,\%. \tag{2.3}$$

De forma análoga se define la medida de especificidad (SP), la cual es conocida como valor predictivo negativo, como:

$$\mathrm{SP}(\,\%) \doteq \frac{\mathrm{TN}}{\mathrm{TN+FP}} \times 100\,\%. \tag{2.4}$$

Finalmente, se define la medida de exactitud (Acc) de la prueba como:

$$\mathrm{Acc}(\,\%) \doteq \frac{\mathrm{SE}+\mathrm{SP}}{2} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP+FN+TN+FP}} \times 100\,\%. \tag{2.5}$$

Durante las últimas décadas, el análisis basado en curvas ROC (*Receiver Operating Characteristic*) se ha convertido en una de las técnicas más populares para evaluar el desempeño de los métodos de diagnóstico médico [52, 53]. En particular, el análisis de curvas ROC se utiliza en la epidemiología clínica con el fin de cuantificar el grado de precisión con el que las pruebas (o métodos) de diagnóstico médico pueden discriminar entre los dos estados más comunes del paciente, típicamente denominados *enfermos* y *sanos* [54]. Conceptualmente, una curva ROC se construye mediante la variación de los valores de umbrales (o límites) de decisión para determinar cuando un paciente se encuentra enfermo. De acuerdo a cada uno de los valores fijados de esos umbrales, se produce una variación en los valores de las medidas de SE y SP de la prueba de diagnóstico. Más precisamente, la curva ROC se genera mediante la representación gráfica de todos los posibles valores de las medidas de SE vs 1-SP a través de todos los puntos de corte considerados en el análisis. Las pruebas de diagnóstico más discriminativas generan curvas ROC cuyas representaciones gráficas se localizan progresivamente cada vez más cerca de la esquina superior izquierda, es decir, el punto (0,1) del *espacio ROC*. La Figura Figura 2.2 muestra dos curvas ROC resaltadas en colores rojo y azul construidas a partir de las pruebas A y B, respectivamente. Se puede observar que la prueba A tiene una mayor capacidad discriminativa que la prueba B. Además, la curva ROC correspondiente a la prueba A tiene un valor de AUC mayor que la curva ROC generada a

partir de la prueba B.

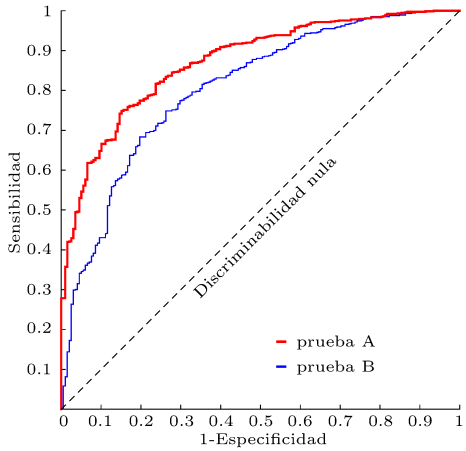


Figura 2.2: Curvas ROC construidas al evaluar dos métodos de diagnóstico distintos (prueba A vs prueba B).

En general, el punto de corte óptimo del análisis ROC es aquel umbral de decisión en el que se maximizan los valores de las medidas de SE y SP simultáneamente. Además, el punto de corte óptimo se obtiene al minimizar la distancia Euclídea entre cada uno de los puntos de la curva ROC y el punto (0,1) de la gráfica. El área bajo la curva (AUC, del inglés *Area Under the Curve*), determina la ubicación completa de la curva ROC en lugar de depender de un punto de corte específico [52, 55]. El AUC es una medida eficaz y combinada de valores de SE y SP que describe la validez inherente de las pruebas de diagnóstico [56].

2.4. Medidas de complejidad

Si bien hoy en día es ampliamente utilizada y aceptada, la noción de "complejidad" muy a menudo carece de una formalización rigurosa. Por lo tanto, no ha de sorprender que aún no exista una medida universalmente aceptada que sea capaz de cuantificar tal concepto. En particular, en el contexto de teoría de la información (*Information Theory*), se sabe que la complejidad de cualquier elemento de un código, o de cualquier característica de una representación de la señal en el contexto del procesamiento de la señal, está estrechamente relacionada con la información que transporta o, más precisamente, con el valor de su entropía [57]. Sin embargo, es

importante señalar que, en el contexto de la clasificación de señales, las características más informativas (en términos de clasificación) no son necesariamente las de mayor entropía. Por lo tanto, se necesitan más medidas *ad-hoc*. De hecho, cualquier medida de complejidad apropiada que corresponda a una característica dada debería, en cambio, estar fuertemente relacionada con la cantidad de información sobre la pertenencia a la clase que proporciona dicha característica. Uno podría entonces pensar en usar como medida de complejidad la entropía condicional de la clase dada la característica. Sin embargo, las características que proporcionan la información más discriminativa con respecto a una clase son casi siempre aquellas con valores de entropía condicional más bajas, y por lo tanto las características más útiles para la clasificación serán las menos complejas.

Si bien los elementos a_j de un vector aleatorio $\mathbf{a}$ son números reales, por razones prácticas resulta conveniente discretizarlos. Este proceso de discretización de la variable aleatoria continua se puede realizar de diferentes maneras. Una de esas formas consiste en particionar la linea real $\mathbb{R}$ en M intervalos igualmente espaciados $I_k \doteq (a_{min}, a_{max}]$, donde $a_{min} = \left(k - \frac{1}{2}\right)\Delta$ y $a_{max} = \left(k + \frac{1}{2}\right)\Delta$, $k \in \mathbb{Z}$, de longitud Δ y la variable aleatoria discreta asociada es $\mathcal{K}_j \doteq \sum_{k\in\mathbb{Z}} k\,\chi_{I_k}(a_j)$. La función de masa de probabilidad (PMF, del inglés *Probability Mass Function*), esta dada por: $p_{\mathcal{K}_j}(k) = P(a_j \in I_k) = \int_{I_k} \pi(a_j)\, da_j$, donde $\pi(a_j)$ es la función de densidad de probabilidad continua.

La entropía de Shannon, también conocida como entropía de la información, es una medida que permite cuantificar el grado de incertidumbre de una fuente de información [57]. El objetivo principal de esta teoría fue el de obtener cotas en operaciones de procesamiento de señales tales como compresión de datos, almacenamiento y comunicación, entre otras. Sea $\mathcal{K} = [\mathcal{K}_1\, \mathcal{K}_2\, \mathcal{K}_3\, \cdots\, \mathcal{K}_M]$ un vector aleatorio discreto y $P = [p_1\, p_2\, p_3\, \cdots\, p_M]$ su correspondiente vector de probabilidades, la entropía se define como:

$$\mathcal{H}_{\mathbf{a}}(P) \doteq -\sum_{m=1}^{M} p_m \log(p_m), \tag{2.6}$$

donde $P = \{p_1\, p_2\, p_3\, \cdots\, p_M\}$, donde p_m es la probabilidad de que el elemento $a_j \in I_k = (a_{min}, a_{max}]$ y M representa, en este caso, el número de particiones resultantes del proceso de discretización del vector aleatorio continuo. Si existe la certeza de que a_j pertenece a un determinado intervalo I_k, entonces $p_m = 1$ y $p_m \log(p_m) = 0$

resultando en pérdida de información. De forma análoga, el producto $p_m \log(p_m) = 0$ si $p_m = 0$.

La divergencia de Kullback-Leibler (KL) [58] es probablemente la medida de información más conocida y usada en numerosos problemas de clasificación [59, 60, 61]. Para comparar las PMFs condicionales correspondientes a los coeficientes de representación asociados al átomo ϕ_j, se utilizó la divergencia de KL de la siguiente manera:

$$\mathrm{KL}\left(p_{\mathcal{K}_j}(\cdot|c_1), p_{\mathcal{K}_j}(\cdot|c_2)\right) \doteq \sum_{k\in\mathbb{Z}} p_{\mathcal{K}_j}(k|c_1) \log\left(\frac{p_{\mathcal{K}_j}(k|c_1)}{p_{\mathcal{K}_j}(k|c_2)}\right), \tag{2.7}$$

Además de las propiedades teóricas y de cómputo que provee la divergencia de KL, lo que usualmente dificulta su uso en muchos problemas reales de clasificación es su falta de simetría. Se puede ver fácilmente que alternando los órdenes de los argumentos en (2.7), el valor de salida puede cambiar. Para solucionar este inconveniente, se puede usar la divergencia de Jeffrey (J) [62]. Si bien no se creó como una versión simétrica de la divergencia de KL, esta nueva medida es la suma de todas las distancias de KL posibles entre distribuciones de probabilidad. La divergencia de J se define como:

$$\mathrm{J}\left(p_{\mathcal{K}_j}(\cdot|c_1), p_{\mathcal{K}_j}(\cdot|c_2)\right) \doteq \mathrm{KL}\left(p_{\mathcal{K}_j}(\cdot|c_1), p_{\mathcal{K}_j}(\cdot|c_2)\right) + \mathrm{KL}\left(p_{\mathcal{K}_j}(\cdot|c_2), p_{\mathcal{K}_j}(\cdot|c_1)\right). \tag{2.8}$$

Otra versión simétrica suavizada de la divergencia de KL es la divergencia de Jensen-Shannon (JS) [63]. Para el problema de comparar las PMFs condicionales asociadas a cada clase, la divergencia de JS se define como:

$$\mathrm{JS}\left(p_{\mathcal{K}_j}(\cdot|c_1), p_{\mathcal{K}_j}(\cdot|c_2)\right) \doteq w_1 \mathrm{KL}\left(p_{\mathcal{K}_j}(\cdot|c_1), q_{\mathcal{K}_j}(\cdot)\right) + w_2 \mathrm{KL}\left(p_{\mathcal{K}_j}(\cdot|c_2), q_{\mathcal{K}_j}(\cdot)\right), \tag{2.9}$$

donde $q_{\mathcal{K}_j}(\cdot) = w_1 p_{\mathcal{K}_j}(\cdot|c_1) + w_2 p_{\mathcal{K}_j}(\cdot|c_2)$ y w_1 y w_2 son los pesos asociados a cada una de las MPFs condicionales, con $w_1, w_2 \geq 0$ y $w_1 + w_2 = 1$.

Dentro de los problemas de clasificación de señales, la distancia de Fisher (F), es una de las medidas más utilizadas [64]. A diferencia de las otras medidas presentadas en esta sección, que requieren estimaciones de las PMF condicionales, la distancia de F utiliza sólo dos parámetros de las distribuciones (los valores medios y las desviaciones estándar). Esto hace que esta medida sea mucho menos costosa desde el punto de vista computacional, pero asume implícitamente ciertas características de

la distribución en estudio (es decir, atributos de segundo orden). La distancia de F se define como:

$$\mathrm{F}(p_{\mathcal{K}_j}(\cdot|c_1), p_{\mathcal{K}_j}(\cdot|c_2)) \doteq \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}, \tag{2.10}$$

donde μ_ℓ y σ_ℓ^2 representan el valor medio y la desviación estándar de $p_{\mathcal{K}_j}(\cdot|c_\ell)$.

3 Selección de átomos discriminativos

Muchas veces resulta de gran utilidad representar un grupo de señales teniendo en cuenta sólo un número pequeño de elementos descriptores (átomos) extraídos de un conjunto grande (diccionario). Los modelos que producen este tipo de representaciones, conocidas como "representaciones ralas", pueden ser considerados óptimos, en el sentido de su completitud, complejidad y capacidad de generalización, entre otros criterios para la caracterización de un modelo. Si bien existen métodos que producen representaciones ralas robustas ante la presencia de ruido (distorsiones) y falta de datos (desconexiones), tales representaciones resultan a menudo inadecuadas si el objetivo final es la clasificación. Más aún, no es posible garantizar que todos (o la mayoría) de los átomos que conforman el diccionario tengan información útil para tareas de clasificación. Por otro lado, es deseable que el clasificador pueda aprovechar la información inherente sólo de un pequeño grupo de átomos para la clasificación. Esto permite tener clasificadores más sencillos y con mayor capacidad de generalización. Por este motivo, en este capítulo se describen los dos métodos desarrollados para la selección de átomos discriminativos a partir de un diccionario dado.

3.1. Representaciones ralas

La representación rala de una señal también es conocida como "codificación rala" en el campo de la "Teoría de la Información". Es importante notar que este tipo de representaciones son simplemente un caso particular dentro de un amplio espectro que se extiende desde las representaciones "representaciones locales" hasta las "representaciones completamente distribuidas" (o densas) [65].

Las representaciones locales, como su nombre lo indica, tienen sólo un elemento descriptor distinto de cero. Su principal ventaja es la fácil y directa aplicación en problemas de clasificación. Sin embargo, son extremadamente sensibles ante la presencia de ruido en los datos lo que dificulta su capacidad de generalización. Por otro lado, las representaciones completamente distribuidas utilizan todos, o la mayoría de

sus elementos para representar a una señal o un patrón dado. Estas representaciones poseen en general menos dimensiones, pero a costa de sacrificar la facilidad de clasificación y sin garantizar independencia entre las dimensiones. Una de las principales ventajas de estas representaciones es su tolerancia al ruido, debido a su gran redundancia. Sin embargo, se ha visto que un menor grado de redundancia es suficiente para producir un comportamiento robusto dando origen a las representaciones ralas que, en muchos casos prácticos, pueden resultar inclusive más robustas que las representaciones completamente distribuidas. Además, las representaciones ralas no son afectadas por la explosión combinatoria y son capaces de lograr representaciones "separables", en el sentido de la clasificación. Desde el punto de vista probabilístico, las representaciones completamente distribuidas suelen asociarse con distribuciones de probabilidad Gaussianas (de sus dimensiones o coeficientes), o incluso uniformes, ya que sus elementos están mayormente activos. Las representaciones locales, en cambio, tienen distribuciones de probabilidad tipo delta de Dirac. Por otro lado, las representaciones ralas tienen distribuciones de probabilidad con picos pronunciados en cero y colas largas, como por ejemplo las distribuciones "Laplacianas" o las de "Cauchy". La Figura 3.1 muestra un ejemplo de cada una de estos tres tipos de representaciones junto con sus correspondientes distribuciones de probabilidad.

El problema de "representación" de una señal puede plantearse en términos de hallar un modelo adecuado de la misma. A partir de este modelo, el cual se denota por la letra $\mathcal{M}$, es posible "generar" tal señal teniendo en cuenta un conjunto de elementos de entrada relacionados con ciertas características de la misma (modelo directo). Además, es posible invertir el modelo (modelo inverso) para obtener, ahora a través de $\mathcal{M}^{-1}$, las características significativas partiendo de una señal en particular. Sin embargo, lograr una "buena" estimación de todos los parámetros de un cierto modelo no es una tarea simple y, en general, requiere de una gran cantidad de datos del entorno. Asi mismo, es importante tener en claro que la formulación del problema directo o inverso constituye un aspecto relativo a la forma en que se plantea el modelo en relación con las características causa-efecto del fenómeno bajo estudio.

Si bien existen distintas formas de considerar un modelo, en este trabajo aborda un enfoque discreto. Además, se tendrá en cuenta que la señal $\mathbf{x}$ se genera mediante un modelo $\mathcal{M}$ con parámetros Φ y $\mathbf{a}$, y es afectada por un ruido ε en forma aditiva,

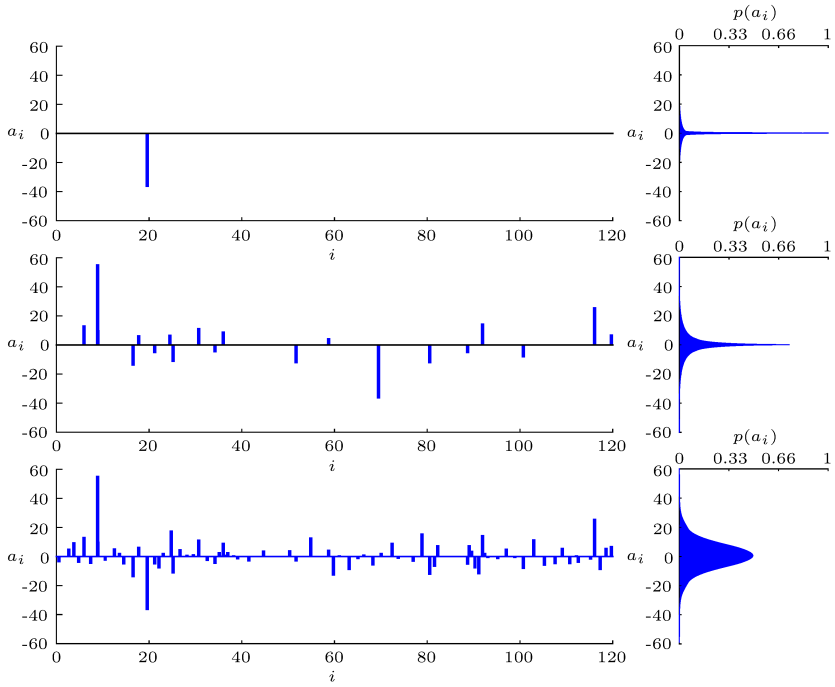


Figura 3.1: Tipos de representaciones de señales: *locales* (arriba), *ralas* (medio) y *completamente distribuidas* (abajo).

es decir:

$$\mathbf{x} = \mathcal{M}(\Phi, \mathbf{a}) + \varepsilon, \tag{3.1}$$

por lo que esta expresión se denomina ecuación del "modelo generativo". La Figura 3.2 muestra un esquema del modelo generativo abordado en la presente tesis. Se puede observar que una señal $\mathbf{x}$ se genera mediante la combinación lineal de solo unas pocas características bien definidas (columnas) de Φ.

El problema de hallar todas las representaciones ralas de un grupo de señales puede separarse en dos sub-problemas, conocidos como: *i*) codificación rala y *ii*) aprendizaje del diccionario. A continuación se procede a describir en detalle cada uno de esos sub-problemas. Para ello, sea $\mathbf{x} \in \mathbb{R}^N$ una señal discreta y sea $\Phi \in \mathbb{R}^{N \times M}$ (generalmente con $M \geq N$) un diccionario cuyas columnas $\boldsymbol{\phi}_j \in \mathbb{R}^N$ son formas de onda básicas llamadas "átomos" que se utilizan para obtener representaciones de la forma $\mathbf{x} = \Phi\mathbf{a}$. A partir de ahora, nos referiremos al vector $\mathbf{a} = [a_1 \; a_2 \; \cdots \; a_M]^T \in \mathbb{R}^M$ como una "representación" of $\mathbf{x}$ en términos de Φ tal que $\mathbf{x} = \Phi\mathbf{a}$. La noción de "raleza" consiste esencialmente en obtener representaciones con el menor número de

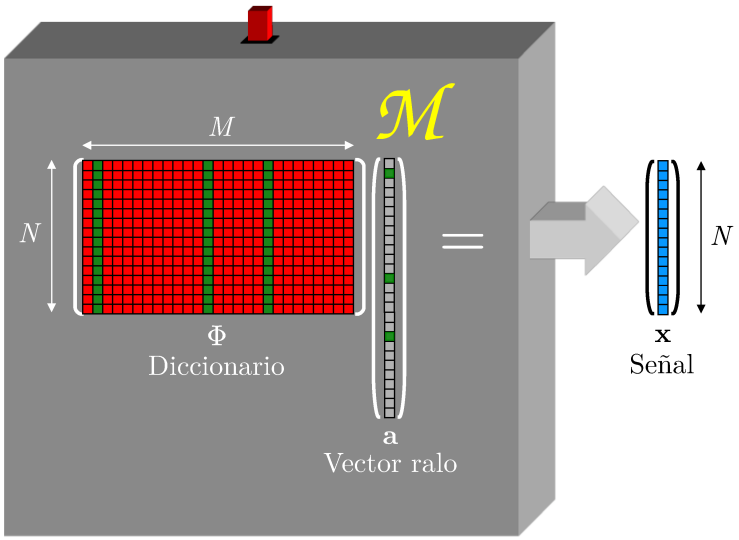


Figura 3.2: Modelo generativo.

elementos distintos de cero como sea posible. Una forma de lograr tales representaciones consiste en hallar la solución del siguiente problema:

$$(P_0): \qquad \min_{\mathbf{a}\in\mathbb{R}^M} ||\mathbf{a}||_0 \qquad \text{sujeto a} \qquad \mathbf{x} = \Phi\mathbf{a},$$

donde $||\mathbf{a}||_0$ denota la pseudo-norma l_0, definida como el número de elementos distintos de cero de $\mathbf{a}$. Resulta que imponer una representación exacta de $\mathbf{x}$ es una restricción demasiado restrictiva, lo que convierte a (P_0) en un problema NP-complejo [66, §1.8], lo que hace que el enfoque no sea adecuado para la mayoría de las aplicaciones prácticas.

Por lo tanto, la restricción de exactitud en la representación $\mathbf{x} = \Phi\mathbf{a}$ es a menudo relajado al permitir pequeños errores de representación e imponer un límite superior en la pseudo-norma l_0 de las representaciones. Por lo tanto, una versión de (P_0) tolerante a errores en la representación se define de la siguiente manera:

$$(P_0^q): \qquad \min_{\mathbf{a}\in\mathbb{R}^M} ||\mathbf{x} - \Phi\mathbf{a}||_2^2 \qquad \text{sujeto a} \qquad ||\mathbf{a}||_0 \leq q,$$

donde q es un parámetro entero prescrito. Esta formulación considera la presencia de posibles términos de ruido aditivo. En otras palabras, se supone que $\mathbf{x} = \Phi\mathbf{a}+\mathbf{e}$ donde

$\mathbf{e} \in \mathbb{R}^N$ es un pequeño término de energía del ruido. Por lo tanto, este enfoque resulta más apropiado en una amplia variedad de aplicaciones reales (como el procesamiento de señales biomédicas) donde las señales crudas capturadas están (en gran medida) contaminadas por ruido. Lamentablemente, los problemas de optimización (P_0) y (P_0^q) son muy difíciles de resolver (de hecho se tratan de problemas NP-completo) y sus costos computacionales resultan a menudo prohibitivos para muchas aplicaciones prácticas [67]. Por esta razón, se han planteado varias alternativas que permiten hallar soluciones aproximadas a estos problemas.

En [68], Chen *et al.* propuso un método de búsqueda de bases (BP, del inglés *Basis Pursuit*). El Método BP permite resolver el problema de optimización (P_0) reemplazando la norma ℓ_0 por la norma ℓ_1. Este nuevo problema (minimizar respecto a la norma ℓ_1) puede convertirse en uno de programación lineal tradicional y, por lo tanto, puede resolverse eficaz y exactamente con los métodos de punto interior de la programación lineal. Asimismo, se han propuesto nuevas estrategias voraces que permiten obtener soluciones aproximadas a los problemas (P_0) y (P_0^q). El método de búsqueda por coincidencias (MP, del inglés *Matching Pursuit*), es uno de los primeros métodos, y quizás el más representativo, que usa la estrategia voraz para hallar las soluciones aproximadas a dichos problemas [69]. La idea principal del método MP es elegir en forma iterativa el mejor átomo del diccionario basado en una cierta medida de similitud para obtener en forma aproximada la solución rala. Por otro lado, el método de búsqueda de coincidencias ortogonales (OMP, del inglés *Orthogonal Matching Pursuit*), es una mejora del método MP [38, 70]. En particular, este método de aproximación voraz emplea el proceso de ortogonalización para garantizar la dirección ortogonal de la proyección en cada una de las iteraciones. Además, se ha comprobado que el algoritmo OMP puede converger en no más de q iteraciones [38].

Si bien los diccionarios pre-construidos (o fijados de antemano) como, por ejemplo, los paquetes wavelet [71] conducen generalmente a una rápida codificación rala, se ha observado que se encuentran condicionados a ciertas clases de señales. Por lo tanto, debido a su falta de generalización, han surgido nuevos enfoques que introducen técnicas de aprendizaje de diccionario (DL, del inglés *Dictionary Learning*), a partir de los datos. Un problema de aprendizaje de diccionario asociado a los datos: q, M, $N \in \mathbb{N}$, $M \geq N$ y n señales en $\mathbb{R}^N$, $\mathbf{x}_1, \cdots, \mathbf{x}_n$, puede ser formalmente escrito

como:

$$(DL): \quad \min_{\substack{\Phi \in \mathbb{R}^{N \times M} \\ \mathbf{a}_i \in \mathbb{R}^M, ||\mathbf{a}_i||_0 \leq q, 1 \leq i \leq n.}} \sum_{i=1}^{n} ||\mathbf{x}_i - \Phi \mathbf{a}_i||_2^2$$

La solución de este problema produce por un lado un diccionario Φ y, por el otro lado, representaciones de todas las señales en términos de ese diccionario cumpliendo con la restricción de raleza para cada una de las señales involucradas $\mathbf{x}_1, \cdots, \mathbf{x}_n$. Es importante notar que en este proceso se minimiza el error total de representación.

Los primeros métodos de aprendizaje de diccionario a partir de los datos se desarrollaron originalmente hace casi dos décadas atrás [72, 73, 74]. Algunos de ellos han surgido al considerar marcos probabilísticos teniendo en cuenta los datos observados como realizaciones de ciertas variables aleatorias [72, 73]. En [73], por ejemplo, los autores desarrollaron un algoritmo denotado por NOCICA (del inglés, *Noise Overcomplete ICA*), que permite encontrar un diccionario redundante que maximice la función de probabilidad de la distribución de probabilidad de los datos. En ese trabajo, se derivó una expresión analítica para la función de verosimilitud aproximando la distribución posterior de las funciones Gaussianas. Por otro lado, un enfoque iterativo para el aprendizaje con diccionarios, conocido como el método de direcciones óptimas (MOD, del inglés *Method of Optimal Directions*), se presentó en [74]. La etapa de codificación rala de este método utiliza el algoritmo OMP [38] seguido de una simple regla de actualización del diccionario.

Más recientemente, Aharon *et al.* propuso un nuevo método iterativo denotado por KSVD (del inglés *K Singular Value Decomposition*) [39]. Este nuevo método consiste esencialmente en actualizar en forma iterativa tanto la representación rala como el diccionario. El método OMP se utiliza para obtener la representación rala mientras que el proceso de actualización del diccionario se realiza de a un átomo a la vez. Por otro lado, los coeficientes de representación pueden cambiar con el fin de minimizar el error total de representación.

3.2. Criterio de discriminabilidad binaria

Como se mencionó anteriormente, la técnica de representaciones ralas de señales consiste esencialmente en obtener una buena aproximación de todas las señales, en

términos de sus formas de onda, por medio de la combinación lineal de sólo unos pocos átomos de un diccionario dado. Sin embargo, en aplicaciones donde el objetivo final es la "clasificación", el interés principal no está centrado sólo en obtener buenas aproximaciones, sino más bien en su poder discriminativo. Esto es su capacidad de identificar la pertenencia de las señales a diferentes clases.

Teniendo en cuenta lo anterior, se plantea en esta tesis como una de las principales hipótesis que es posible extraer información discriminativa a partir del análisis de las activaciones de los átomos de un diccionario. Es decir, si un átomo ϕ_j se usa para representar una señal $\mathbf{x}$ dada, entonces su correspondiente coeficiente de representación a_j contiene importante información discriminativa que puede ser de gran utilidad en tareas de clasificación (notar que para la formulación de ésta hipótesis, se considera un problema de clasificación binario). Más precisamente, si ϕ_j se usa muchas más veces para representar señales pertenecientes a una clase en particular que para la otra, entonces ϕ_j es discriminativo para todas las señales pertenecientes a esa clase.

En base a la hipótesis establecida, se define la "frecuencia de activación condicional" η_ℓ^j, para $\ell = \{1, 2\}$, como el número de veces que se usa ϕ_j para representar todas las señales de la clase ℓ. Además, la "probabilidad de activación condicional" de ϕ_j dada (que una señal $\mathbf{x}$ pertenezca a) la clase ℓ se define como $p_\ell^j \doteq P(a_j \neq 0 | \mathbf{x} \in \ell)$. Dado un grupo de n_ℓ señales clase ℓ, la probabilidad de activación condicional se define como:

$$p_\ell^j = \frac{\eta_\ell^j}{n_\ell}. \tag{3.2}$$

Notar que si el problema de clasificación binario es balanceado, es decir si el número de señales disponibles de cada una de las clases el el mismo, por ejemplo $\hat{n}$, entonces $\eta_\ell^j \propto p_\ell^j$, más precisamente $\eta_\ell^j = 2n_\ell p_\ell^j$, para toda clase ℓ y átomo j.

Se considera que un átomo ϕ_j contiene importante información discriminativa para señales clase ℓ si $p_\ell^j > p_m^j$, para todo $m \neq \ell$. Por lo tanto, si ϕ_j es discriminativo para la clase ℓ, las activaciones del coeficiente de representación a_j estarán asociadas (o darán un indicio de que corresponden) a señales clase ℓ. Teniendo en cuenta estos conceptos, se define la función $\text{DCAF} : \{1, 2, \cdots, M\} \to \mathbb{R}_0^+$ como

$$\text{DCAF}(j) \doteq |p_1^j - p_2^j|, \tag{3.3}$$

a partir de ahora se referirá a DCAF$(\cdot)$ como la medida de información discriminativa. Además, DCAF$(\cdot)$ es simétrica, simple, su valor es siempre positivo, $0 \leq \text{DCAF}(j) \leq 1$, para todo $j = 1, 2, \cdots, M$, y es económica, en términos de su costo computacional.

Se dice que ϕ_j es discriminativo (para una de las dos clases) si DCAF$(j) > 0$. Además, el valor de DCAF(j) puede pensarse como una medida del grado de discriminabilidad de ϕ_j basado solamente en su probabilidad de activación condicional. Puesto que cada átomo ϕ_j de Φ tiene asociado su correspondiente grado de discriminabilidad DCAF(j), es posible ordenarlos en forma decreciente de acuerdo a DCAF(j). La Figura 3.3 muestra una curva que representa el grado de discriminabilidad DCAF(j^*) de cada uno de los átomos ϕ_{j^*}. Notar que en este proceso de jerarquización de los átomos, se ha reemplazado el valor del índice correspondiente a j por j^*. Esto se debe a que no necesariamente existe una correspondencia uno a uno entre los átomos más discriminativos de Φ y su "ubicación" en el diccionario. Por ejemplo, el átomo más discriminativo correspondiente a $j^* = 1$ no necesariamente corresponde al primer átomo ϕ_1 de Φ, sino que puede ser cualquiera de los demás.

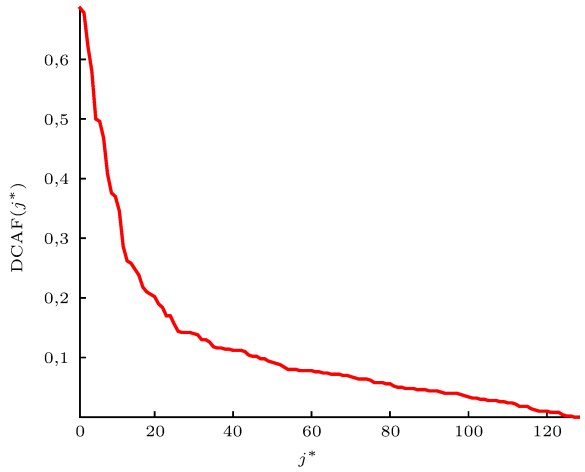


Figura 3.3: Representación gráfica del grado de discriminabilidad DCAF(j^*) de cada uno de los átomos $\phi_1, \phi_2, \cdots, \phi_{128}$ en un diccionario Φ ordenado en forma decreciente de acuerdo a DCAF(j^*).

Se puede observar que p_ℓ^j, para $\ell = 1, 2$, definida por (3.2) puede ser aproximada por el estimador de máxima verosimilitud de la probabilidad de activación

condicional, es decir:

$$p_m^j \approx p_{\mathcal{K}_j}(k \neq 0 | \ell = m), \qquad (m = 1, 2)\,. \tag{3.4}$$

Luego, teniendo en cuenta (3.4), la expresión (3.3) se puede reescribir de la siguiente manera,

$$\begin{aligned} \text{DCAF}(j) &\approx |p_{\mathcal{K}_j}(k \neq 0 | \ell = 1) - p_{\mathcal{K}_j}(k \neq 0 | \ell = 2)| \\ &\approx |\big(1 - p_{\mathcal{K}_j}(k = 0 | \ell = 1)\big) - \big(1 - p_{\mathcal{K}_j}(k = 0 | \ell = 2)\big)| \\ &\approx |p_{\mathcal{K}_j}(k = 0 | \ell = 2) - p_{\mathcal{K}_j}(k = 0 | \ell = 1)|, \end{aligned} \tag{3.5}$$

donde se ha expresado la medida DCAF(j) en base a las probabilidades de activación condicionales complementarias, es decir, las probabilidades de que el átomo ϕ_j no participe en la representación de los datos pertenecientes a ambas clases. Se puede observar que, a excepción de la medida F, el resto de las medidas de información introducidas en la Sección 2.4 pueden expresarse en términos de sumatorias, donde sólo uno de sus términos corresponde a los valores de las probabilidades condicionales para $k = 0$. Sin embargo, debido a la gran raleza de las representaciones, los términos asociados con $k = 0$ son particularmente importantes. Este hecho permite esperar cierta correlación entre los resultados obtenidos con las diferentes medidas de información existentes y la nueva medida DCAF(j).

La parte izquierda de la Figura 3.4 muestra una representación de las PMFs condicionales para los átomos ϕ_j (arriba) y ϕ_i (abajo). Al tratarse de representaciones ralas, las distribuciones de activación condicional de los átomos tienen picos agudos en $k = 0$ con colas pronunciadas. En la parte central de la figura se ilustran las mismas funciones en las que se descartaron los picos centrados en cero ($k = 0$). Se puede observar que tal exclusión produce una reducción significativa en la magnitud de la escala del eje de las abscisas, lo cual resalta la importancia de las activaciones de los átomos en las representaciones ralas. Sin embargo, la discrepancia entre las distribuciones no sólo se da en la probabilidad de activación de los átomos ($k = 0$), sino que además se observan ligeras diferencias entre los valores de probabilidad para todos (o la mayoría de) los valores de $k \neq 0$ (región ampliada). Asimismo, los valores absolutos de estas diferencias se representan por las regiones sombreadas

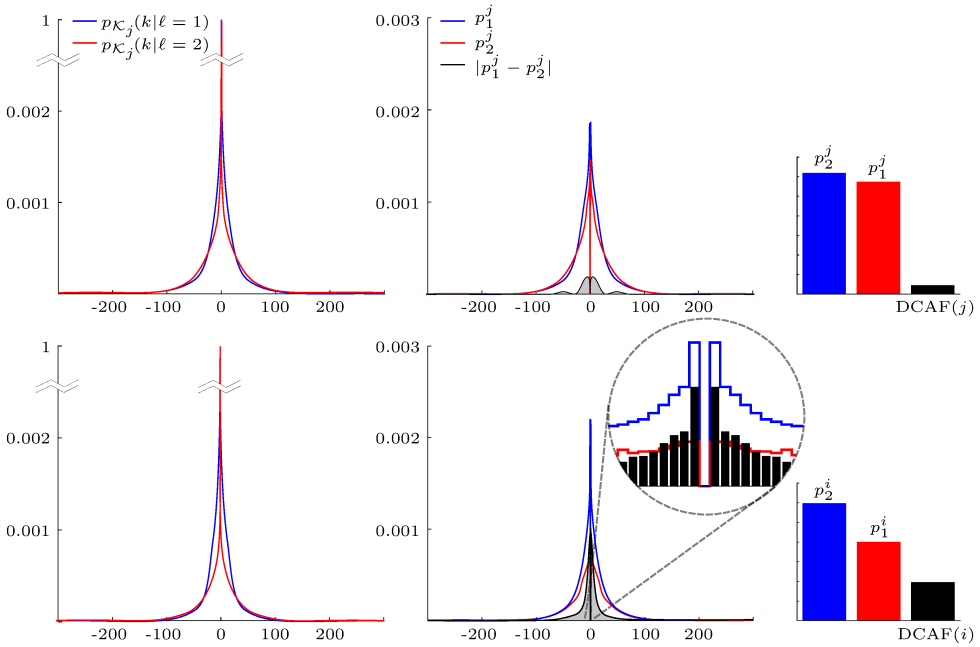


Figura 3.4: PMFs condicionales correspondientes a las activaciones de 2 átomos diferentes (izquierda), las mismas funciones excluyendo el pico centrado en cero ($k = 0$) y el valor absoluto de sus diferencias (centro) y una interpretación gráfica de la medida DCAF (derecha). Primera fila: un átomo no discriminativo ϕ_j. Segunda fila: un átomo discriminativo ϕ_i.

en color gris. También es importante señalar que estos valores de área mostrados en gris ($\sum_{k\neq 0} |p_{\mathcal{K}_j}(k|\ell = 2) - p_{\mathcal{K}_j}(k|\ell = 1)|$) no son necesariamente iguales a los correspondientes a los valores DCAF(j). Sin embargo, para las PMFs simétricas con curtosis alta y colas agudas (como es el caso de las PMFs utilizadas en este trabajo), las distribuciones condicionales y a-priori suelen ser similares y, por lo tanto, ambos valores de área están cerca uno del otro.

Los métodos que se introducen a continuación denotados por MDAS y MDCS tienen en cuenta dos formas distintas de emplear la información útil de los átomos más discriminativos con el fin de clasificar los eventos de AH.

3.3. Métodos propuestos

Los dos métodos discriminativos introducidos en esta sección pueden pensarse como métodos de selección de características significativas (en el contexto de la clasificación de señales) y se desarrollaron específicamente para un fin común que es la clasificación de eventos de AH. En primer lugar se desarrolló un método para la selección de átomos más discriminativos denotado por MDAS (del inglés *Most Discriminative Atoms Selection*). Luego, en segundo lugar, se propuso un nuevo método para la construcción de sub-diccionarios discriminativos denominado MDCS (del inglés *Most Discriminative Columns Selection*). A continuación se introduce una breve descripción de cada uno de los métodos propuestos. Para más detalles de los mismos, se remite al lector al artículo incluido en el Anexo B.

El método MDAS: luego del filtrado y segmentado de todas las señales de SaO_2 incluidas en el conjunto de entrenamiento ($\mathbf{X}_{trn}$), se aprende un diccionario Φ a partir de la aplicación del método NOCICA [73]. Luego, mediante el uso del algoritmo OMP, se obtienen los códigos ralos $\mathbf{a}_i \in \mathbb{R}^M$, para $i = 1, 2, \cdots, n$, de los segmentos $\mathbf{x}_i$ en términos de Φ a través de $\mathbf{x}_i = \Phi\mathbf{a}_i$ y se define la matriz de códigos ralos $\mathbf{A}_{trn} \doteq [\mathbf{a}_1\, \mathbf{a}_2\, \cdots\, \mathbf{a}_n]$. Claramente la matriz $\mathbf{A}_{trn}$ es tal que $\mathbf{X}_{trn} = \Phi\mathbf{A}_{trn}$. A partir de $\mathbf{A}_{trn}$, se utiliza la medida de información discriminativa d con el fin de cuantificar el grado de información discriminativa que contiene cada uno de los átomos ϕ_j, para $j = 1, 2, \cdots, M$, de Φ. La originalidad de este método es que utiliza el valor de los primeros F coeficientes $\{a_{j^*}\}$, para $j^* = 1, 2, \cdots, F$, $F \leq M$, correspondientes a los valores de las activaciones de los F átomos más discriminativos ϕ_{j^*} de Φ. Este proceso de selección de características discriminativas da origen a un vector de características cuya dimensión es menor que el código ralo original, que es utilizado directamente como entrada de un clasificador del tipo MLP entrenado para la detección de eventos de AH.

El método MDCS: Si bien este método discriminativo es similar al método descrito anteriormente, MDCS se diferencia claramente de MDAS en la forma en que selecciona las características discriminativas. Una vez que se identifican los F átomos más discriminativos ϕ_{j^*} de Φ, para $j^* = 1, 2, \cdots, F$, a diferencia del método anterior, se procede a construir un nuevo sub-diccionario Φ_N de tamaño $N \times F$ cuyos átomos son los más discriminativos de acuerdo a d. Luego, mediante el uso del

algoritmo OMP, se obtienen nuevos códigos ralos $\mathbf{a}_i \in \mathbb{R}^F$, para $i = 1, 2, \cdots, n$, de los segmentos $\mathbf{x}_i$ en términos de Φ_N a través de $\mathbf{x}_i = \Phi_N \mathbf{a}_i$. Finalmente, este nuevo vector de características es usado como entrada del clasificador entrenado para la detección de eventos de AH.

3.4. Experimentos realizados

Puesto que las personas que padecen un SAHOS severo requieren de un tratamiento especial y de forma urgente, los hospitales, clínicas y centros privados de medicina respiratoria tienen como principal propósito identificar a la mayor parte de la población que lo sufra. No obstante, la medicina de atención primaria es determinante en la identificación de los pacientes que padezcan un SAHOS moderado y requiere de una adecuada coordinación con las unidades de sueño de referencia, para el estudio de los casos difíciles o dudosos. Por esta razón, se necesita de equipos de diagnóstico simplificados que sean sencillos y de bajo costo. Un adecuado uso de estos equipos simplificados correctamente validados permitiría, seleccionando los casos, descentralizar el diagnóstico de los centros de referencia habitualmente saturados, hacia unidades de diagnóstico más pequeñas equipadas con oxímetros.

La base de datos utilizada para llevar a cabo los experimentos en esta tesisevaluar los métodos discriminativos propuestos, denominada estudio de la salud del corazón y del sueño (SHHS) de sus siglas del inglés “Sleep Heart Health Study”, ha surgido a partir de un estudio poblacional sobre las consecuencias cardiovasculares de los trastornos del sueño [49, 50]. La base de datos completa incluye 995 estudios que contienen varias señales biomédicas como señales EEG, ECG, flujo respiratorio y SaO_2, entre otras. Además, se incluyen anotaciones de las distintas etapas del sueño y de los eventos de apnea-hipopnea. Si bien es posible analizar “todas” las señales de la base de datos, los métodos discriminativos desarrollados en esta tesis permiten detectar en forma automática los eventos de apnea-hipopnea mediante el análisis y procesamiento exclusivo de señales de SaO_2.

El conjunto de datos disponible cuenta con 995 estudios, de los cuales 41 se descartaron debido a información inconsistente. Entre los 954 estudios restantes, se seleccionaron aleatoriamente 667 (equivalente a 70 %) estudios y se fijaron como datos de entrenamiento. En todos los casos, la prueba final se realizó utilizando los

287 (30 %) estudios restantes de la base de datos no utilizados durante la etapa de entrenamiento.

Si bien en experimentos preliminares se utilizaron distintos tipos de clasificadores, entre ellos se consideró un clasificador lineal, una SVM con núcleo Gaussiano, una ELM y un MLP, su elección final se ha basado fundamentalmente en aquel que maximiza la tasa de reconocimiento de eventos de AH, teniendo en cuenta además que el método en su conjunto tenga un bajo costo computacional.

El aprendizaje del diccionario puede llevarse a cabo mediante el uso de i) datos mezclados, es decir, sin tener en cuenta información acerca de las clases o ii) datos etiquetados, es decir, considerando información concerniente a las clases. Además, otro factor importante en el aprendizaje de los diccionarios es su tamaño, es decir, el número de átomos que lo conforman (el factor de redundancia (r_f) es una medida que evalúa el grado de completitud que tiene un diccionario r_f se define como el cociente M/N y el diccionario se denomina sub-completo, completo o sobre-completo si r_f es menor, igual o mayor a 1). En este esquema, la presente tesis aborda no sólo el aprendizaje de diccionarios a partir de datos mezclados y de datos etiquetados, sino también explora el efecto que produce su factor de redundancia.

A continuación se presenta un resumen de los resultados obtenidos al comparar la capacidad de la nueva medida DCAF y de las medidas de información introducidas en el Capítulo 2.4 en el reconocimiento de los eventos de AH. Además, se compara el desempeño uno de los métodos propuestos con otros tres métodos del estado del arte.

3.5. Resultados y discusión

Esta sección presenta los resultados obtenidos al evaluar el desempeño de los métodos propuestos en el reconocimiento de los eventos de AH y en el diagnóstico del SAHOS moderado. En un principio, se muestran los resultados obtenidos al evaluar la capacidad de "detectar" átomos discriminativos de la nueva medida DCAF y el resto de las medidas de información presentadas en la Sección 2.4. Luego, se presentan los resultados logrados por el método MDCS-OD (que hace uso de la medida DCAF para detectar átomos discriminativos) y se los compara con los obtenidos por otros tres métodos tradicionales [23, 30, 31].

3.5.1. Comparación con otras medidas

Esta sección presenta los resultados de los experimentos exploratorios, los que se realizaron con el objetivo de evaluar la capacidad que tiene cada una de las medidas de discriminabilidad no sólo para detectar los átomos que aportan mayor información para la clasificación, sino también para cuantificar su grado de discriminabilidad. Asimismo, se exploró el efecto que tienen algunos de los parámetros más importantes (en el contexto de las Representaciones Ralas) en el desempeño de una ELM entrenada para el reconocimiento de los eventos de AH. En los experimentos realizados, el número de neuronas en la capa oculta de la ELM corresponde a cuatro veces la dimensión del vector de entrada. Además, se hizo uso de la función *sigmoid* como función de activación de las neuronas.

El factor de redundancia de los diccionarios y la raleza en las representaciones son dos factores muy importantes a tener en cuenta y son precisamente los que se estudiaron en esta tesis. En particular, los experimentos realizados en este contexto abordan el aprendizaje de diccionarios con factores de redundancia 1, 2 y 4. Asimismo, se consideraron representaciones con distintos grados de raleza, las cuales varían desde el 5 % al 25 %. Por ejemplo, una representación con un grado de raleza del 5 % tiene sólo ese porcentaje de sus coeficientes diferentes de cero. Además, para cada una de las clases de diccionarios, se evaluó el efecto que produce la construcción de sub-diccionarios (donde sus átomos son discriminativos) en el desempeño del clasificador. Para ello se consideraron tamaños de sub-diccionarios que varían desde el 10 % hasta el 100 %.

En primer lugar, se aprendieron tres clases de diccionarios denotados por $\Phi 1$, $\Phi 2$ y $\Phi 4$ con factores de redundancia 1, 2 y 4, respectivamente, mediante el algoritmo K-SVD [39]. Estos tres diccionarios se aprendieron a partir de n segmentos de señales de SaO_2, $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$, incluidas en $\mathbf{X}_{trn}$ sin tener en cuenta información respecto a las etiquetas en los datos. Puesto que la dimensión de los segmentos es 128, los diccionarios $\Phi 1$, $\Phi 2$ y $\Phi 4$ están compuestos por 128, 256 y 512 átomos, respectivamente.

Luego, a partir de los diccionarios $\Phi 1$, $\Phi 2$ y $\Phi 4$ ya aprendidos, se hizo uso de las medidas de discriminabilidad para identificar los átomos discriminativos y de cuantificar su correspondiente grado de discriminabilidad. La Figura 3.5 muestra

una representación gráfica de los 7 átomos más discriminativos de $\Phi 1$ de acuerdo a la nueva medida DCAF (primer fila) y al resto de las medidas de información evaluadas en esta tesis (filas 2 a 5). Se puede observar que el átomo más discriminativo de $\Phi 1$ de acuerdo a la medida DCAF (color azul) brinda información relevante asociada a dos desaturaciones (bien marcadas) en la señal de SaO_2. Es importante señalar que este átomo resulta ser también el más discriminativo cuando se utiliza la divergencia de Jeffrey, o eventualmente cuando se utiliza la divergencia de Jensen-Shannon. Por otro lado, al usar la divergencia de Kullback-Leibler, tal átomo resulta ser el cuarto más discriminativo. Si se considera el segundo átomo más discriminativo de acuerdo a la medida DCAF (color rojo), ese átomo es el cuarto más discriminativo de acuerdo a la divergencia de Jensen-Shannon. Por último, se puede notar que la distancia de Fisher no identifica átomos discriminativos.

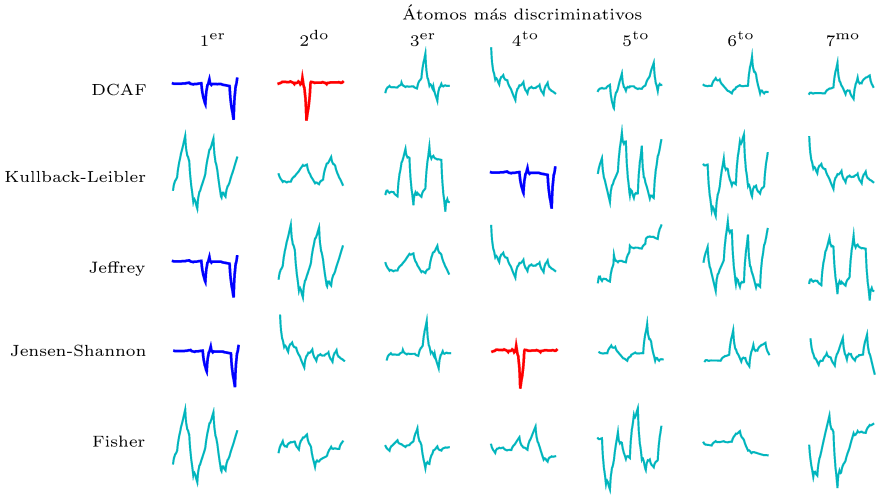


Figura 3.5: Representación gráfica de los 7 átomos más discriminativos determinados por la nueva medida DCAF (primer fila) y por las otras medidas de información (filas 2 a 5).

En base a los átomos más discriminativos (de acuerdo a cada una de las medidas) de cada uno de los diccionarios $\Phi 1$, $\Phi 2$ y $\Phi 4$, se procedió a construir 10 tipos de sub-diccionarios con diferentes tamaños que varían en porcentaje desde el 10 % hasta el 100 %. Es importante notar que, por ejemplo, un sub-diccionario cuyo tamaño corresponde al 10 % del diccionario original, está compuesto por el 10 % de los átomos más discriminativos de ese diccionario. Finalmente, para cada clase de dic-

cionario y teniendo en cuenta cada tipo de sub-diccionario, se evaluó el desempeño del clasificador para distintos niveles de raleza en las soluciones. Para más detalles respecto a los resultados obtenidos, observar las figuras 8, 9, 10 y 11 del artículo incluido en el Anexo A.

Además, se realizó un análisis comparativo entre las tasas de reconocimiento de eventos de AH promedio y máximas alcanzadas por el clasificador a partir de representaciones ralas obtenidas tanto en términos de los sub-diccionarios como de los diccionarios originales. La Tabla 3.1 muestra un resumen de los resultados alcanzados por el clasificador sólo para sub-diccionarios construidos con el 10 % de los átomos del diccionario original, permitiendo que los niveles de raleza varíen desde el 5 % hasta el 25 %. Se puede notar que, en todos los casos, los sub-diccionarios discriminativos superan a los diccionarios completos en la detección de eventos de AH. Asimismo, el factor de redundancia unitario, es decir, diccionarios con igual cantidad de átomos que su dimensión, mejora el desempeño del clasificador en la detección de eventos de AH.

Tabla 3.1: Tasas de reconocimiento de eventos de AH para sub-diccionarios de tamaño equivalente al 10 % de Φ1, Φ2 y Φ4.

Medida	Φ1		Φ2		Φ4	
	Máx.	Media	Máx.	Media	Máx	Media
DCAF	**72,62**	64,68	65,20	63,15	65,19	64,21
Kullback-Leibler	**73,20**	64,91	65,44	63,53	65,42	63,66
Jeffrey	**72,82**	64,88	64,50	62,82	65,39	63,68
Jensen-Shannon	**72,55**	64,10	65,02	63,18	65,87	64,01
Fisher	**72,23**	65,21	64,57	63,04	65,64	62,71
Diccionario original	66,39	59,77	68,13	59,57	69,28	69,21

Finalmente, se evaluó la capacidad del método propuesto (que hace uso de un sub-diccionario con el 10 % de los átomos más discriminativos de Φ_1) para el diagnóstico del SAHOS moderado, es decir, se fijó el umbral de detección de la patología en 15 eventos de AH por hora de sueño. Para ello, se realizó un análisis de curvas ROC del desempeño del método tanto con el uso de la nueva medida DCAF como de las otras medidas de información. Las representaciones gráficas de las curvas ROC se pueden

ver en detalle en el artículo incluido en el Anexo A. La Tabla 3.2 muestra las medidas de desempeño del método para cada una de las medidas extraídas a partir de las curvas ROC. Las columnas de la tabla presentan la distancia mínima ($d_{\text{mín}}$) entre el punto de corte y el punto (0,1) de la gráfica, los porcentajes de SE, SP y Acc y los valores de AUC. Los resultados muestran que la medida DCAF supera a las otras medidas de información en términos de $d_{\text{mín}}$, la sensibilidad y tasa de reconocimiento de pacientes con SAHOS moderado. Además, la divergencia de Jeffrey obtuvo en el máximo valor de AUC mientras que la mayor especificidad se obtuvo al hacer uso de la divergencia de Jensen-Shannon.

Tabla 3.2: Medidas de desempeño para el diagnóstico del SAHOS moderado mediante el uso de un sub-diccionario de tamaño equivalente al 10 % de $\Phi 1$.

Medida	$d_{\text{mín}}$	SE(%)	SP(%)	Acc(%)	AUC
DCAF	**0,2211**	**81,88**	87,32	**84,60**	0,9250
Kullback-Leibler	0,2242	81,46	87,39	84,43	0,9271
Jeffrey	0,2311	80,86	87,04	83,95	**0,9283**
Jensen-Shannon	0,2267	80,75	**88,03**	84,39	0,9244
Fisher	0,2280	80,66	87,91	84,29	0,9252

3.5.2. Comparación con otros métodos

Para comparar el desempeño del método propuesto con otros métodos del estado del arte, se decidió trabajar con redes neuronales del tipo perceptrón multi-capa (MLP). El ajuste de los principales parámetros del MLP se realizó mediante el método de búsqueda por grilla considerando: *i*) coeficiente de aprendizaje, *ii*) número de entradas (tamaño del vector de características) y *iii*) número de neuronas en la capa oculta. Claramente, el ajuste de estos parámetros se realizó con una partición de validación no utilizada luego para el entrenamiento y testeo de los algoritmos discriminativos propuestos. El entrenamiento del MLP se realizó a través de una estrategia de entrenamiento por mini-lotes llamada "Mini-Batch Training". Puesto que cada mini-lote esta formado por 1000 segmentos balanceados, esto evita un "sobre-ajuste" de la red. Además, la evidencia empírica establece que 4 iteraciones

del algoritmo gradiente conjugado escalado evita un "sobre-entrenamiento" de la red. Por último, con el fin de minimizar el sesgo en la clasificación de eventos, el esquema de entrenamiento recientemente descrito se ha repetido 455 veces.

En primer lugar, se aprendió un diccionario completo (CD, del inglés *Complete Dictionary*) Φ_{CD}, de tamaño 128×128 mediante datos mezclados, es decir, a partir de la matriz completa $\mathrm{X}_{\mathrm{train}}$. Luego se aprendió un diccionario sobre-completo (OD, del inglés *Overcomplete Dictionary*) Φ_{OD}, de tamaño 128×256. En particular, este diccionario ha sido construido mediante la unión de dos diccionarios completos Φ^{c1} y Φ^{c2}, que fueron aprendidos con los datos etiquetados $\mathrm{X}^{\mathrm{c1}}_{\mathrm{train}}$ y $\mathrm{X}^{\mathrm{c2}}_{\mathrm{train}}$, respectivamente. En todos los casos, el diccionario fue aprendido mediante el método NOCICA [73].

Si bien en la presente tesis se ha abordado el aprendizaje de diccionarios completos (CD) y sobre-completos (OD), en ambos casos las pruebas de laboratorio arrojaron resultados prometedores, logrando altas tasas de reconocimiento de eventos de AH. Sin embargo, se ha observado una leve mejoría en los resultados al utilizar información de clase en el proceso de construcción de diccionarios sobre-completos (redundantes). Por lo tanto, se logró establecer que los códigos ralos obtenidos a partir de diccionarios redundantes (que fueron aprendidos teniendo en cuenta información de clase en los datos) mejoran el rendimiento del clasificador en la detección de eventos de AH. Para más detalles sobre los resultados alcanzados por las dos clases de diccionarios estudiados, se remite al lector a la Sección 3 del artículo incluido en el Anexo B.

Los métodos discriminativos MDAS-OD y MDCS-OD propuestos en esta tesis y el método original FULL-OD fueron comparados al ser evaluados sobre el mismo conjunto de datos de prueba (no utilizados en el proceso de entrenamiento). La Tabla 3.3 muestra un resumen comparativo del desempeño que obtuvo el método FULL-OD y del alcanzado por ambos métodos discriminativos propuestos (MDAS-OD y MDCS-OD). Las medidas de desempeño presentadas en esta tabla han sido extraídas a partir del análisis de las curvas ROC, las cuales han sido construidas al evaluar cada uno de los tres métodos para la detección del SAHOS moderado. Puesto que el punto de corte (umbral de detección) elegido en estas curvas es aquel en el que los valores de sensibilidad (SE) y de especificidad (SP) se maximizan en simultáneo, las pruebas de laboratorio mostraron que el método MDCS-OD ha logrado obtener el máximo valor de las medidas AUC, SE y Acc, los cuales corresponden a 0,937,

85,65 % y 85,78 %, respectivamente, superando significativamente a los otros dos métodos analizados. Además, se observa que el método FULL-OD ha alcanzado un valor de SP de 87,32 %, el cual representa el valor máximo de esa medida entre todos los métodos evaluados.

Tabla 3.3: Medidas de desempeño para el diagnóstico del SAHOS moderado-severo para distintas combinaciones de los métodos MDCS y FULL.

Método	AUC	SE(%)	SP(%)	Acc(%)
FULL-OD	0,923	83,33	**87,32**	85,33
MDAS-OD	0,891	81,02	83,10	82,06
MDCS-OD	**0,937**	**85,65**	85,92	**85,78**

Además, se realizó una comparación entre el desempeño del método discriminativo MDCS-OD y otros tres métodos conocidos del estado del arte para el diagnóstico del SAHOS moderado. Los métodos que fueron considerados para realizar la comparación son los propuestos por Schlotthauer *et al.*, Vázquez *et al.* y Chiner *et al.* [23, 31, 30]. La Figura 3.6 muestra las curvas ROC construidas al evaluar cada uno de los métodos para el diagnóstico del SAHOS moderado. En particular, los resultados arrojaron que la curva ROC construida a partir del método discriminativo MDCS-OD es la que obtuvo el punto de corte (punto negro) más próximo al punto "óptimo" (punto rojo) de la gráfica. Este hecho indica que el método discriminativo MDCS-OD ha alcanzado el máximo valor de Acc en la detección del SAHOS moderado, superando a los demás métodos en este sentido.

Por otro lado, la Tabla 3.4 muestra un resumen comparativo de los resultados logrados por el método discriminativo MDCS-OD y de los alcanzados por todos los demás métodos evaluados para la detección del SAHOS moderado. Estos resultados muestran que el método MDCS-OD supera a los demás en cuanto a las medidas de desempeño AUC, SE y Acc, para las cuales se obtuvieron valores de 0,937, 85,65 % y 85,78 %, respectivamente. Asimismo, el método propuesto por Vázquez *et al.* alcanza una SP de 87,50 %, la cual representa el máximo valor de SP entre todos los métodos en cuestión. Sin embargo, los médicos especialistas utilizan generalmente el valor del AUC como medida estándar para la evaluación los distintos métodos de diagnóstico del SAHOS. En este sentido, el método discriminativo MDCS-OD resulta ser, entre

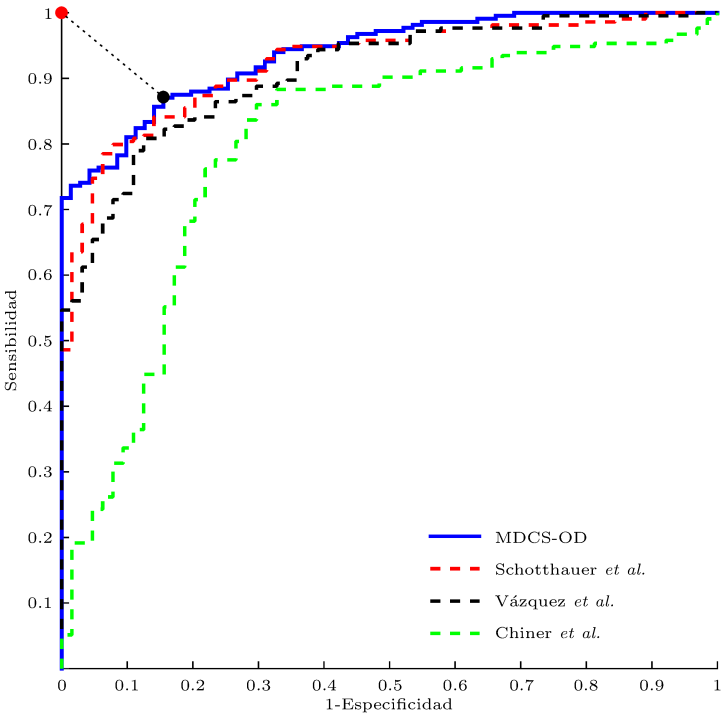


Figura 3.6: Curvas ROC para los 4 métodos de diagnóstico evaluados.

todas las alternativas analizadas, el más adecuado para el diagnóstico del SAHOS moderado. Además, se realizó una comparación de los costos computacionales de los algoritmos para poder analizar su posible ejecución en tiempo real. Los resultados obtenidos se detallan en la Tabla 5 del artículo incluido en el Anexo B.

Tabla 3.4: Medidas de desempeño para el diagnóstico del SAHOS moderado-severo para el método MDCS-OD y otros tres métodos de detección.

Método	AUC	SE(%)	SP(%)	Acc(%)
MDCS-OD	**0,937**	**85,65**	85,92	**85,78**
Schlotthauer *et al.* [23]	0,922	84,11	85,94	85,02
Vázquez *et al.* [31]	0,909	80,84	**87,50**	84,17
Chiner *et al.* [30]	0,795	76,17	78,12	77,15

Existen aplicaciones donde el punto de corte de la curva ROC suele desplazarse a lo largo de toda la curva logrando así maximizar la SP (desplazamiento hacia el

origen de coordenadas) o maximizar la SE (desplazamiento del punto de corte hacia el punto $(1, 1)$ de la gráfica). Si el objetivo principal de la prueba es la detección temprana de la enfermedad en un gran número de individuos aparentemente sanos (o *screening*), se suele elegir un valor de SE alto. Teniendo en cuenta esto, si se fija el valor de la SE en 98 % en las curvas ROC de la Figura 3.6, el método discriminativo MDCS-OD alcanza una especificidad del 46,48 %, seguido por el método propuesto por Schlotthauer *et. al.* que alcanza el 34,37 %.

Un aspecto adicional muy valioso de los métodos discriminativos el hecho de que se ha logrado establecer una fuerte correlación entre los vectores de características discriminativas y los eventos de apnea-hipopnea. Esta relación se puede ver en Figura 3.7. En la parte superior de esta figura se muestra una parte de la señal de oximetría de pulso filtrada con las marcas de los eventos de apnea-hipopnea etiquetados por el médico experto. Inmediatamente debajo se muestra una curva (en verde) que representa la suma absoluta de las activaciones de los 16 coeficientes más discriminativos y las etiquetas de los eventos de apnea-hipopnea (en rojo). La imagen que aparece en la parte inferior de la Figura 3.7 muestra el valor absoluto de los 15 coeficientes más discriminativos que obtuvo el método MDCS-OD. Se puede observar con claridad la fuerte correlación existente entre las etiquetas de eventos de apnea determinadas por los médicos expertos y los coeficientes más discriminativos.

Por otro lado, se realizó un análisis comparativo teniendo en cuenta la representación gráfica (en el dominio del tiempo) de los segmentos de señales de SaO_2 y las formas de onda de aquellos átomos más discriminativos que fueron utilizados en sus representaciones ralas correspondientes. La parte superior de la Figura 3.8 muestra (en color azul) la representación gráfica de un segmento de la señal de SaO_2. Se puede observar claramente que durante ese periodo de tiempo se produjeron dos eventos de AH (bandas verticales grises). Además, se observa que esos eventos respiratorios no solo ocurrieron con distinta duración, sino que también produjeron desaturaciones fácilmente "identificables" (marcas en color verde) en la señal de SaO_2. Por otro lado, las tres curvas en color rojo representan las formas de onda de tres de los 15 átomos más discriminativos que se activaron, es decir, que fueron utilizados para obtener la representación rala del segmento de señal bajo estudio. Uno podría entonces pensar que la forma de onda del átomo más discriminativo del diccionario Φ, es decir ϕ_1, contiene importante información relacionada al primer evento de AH.

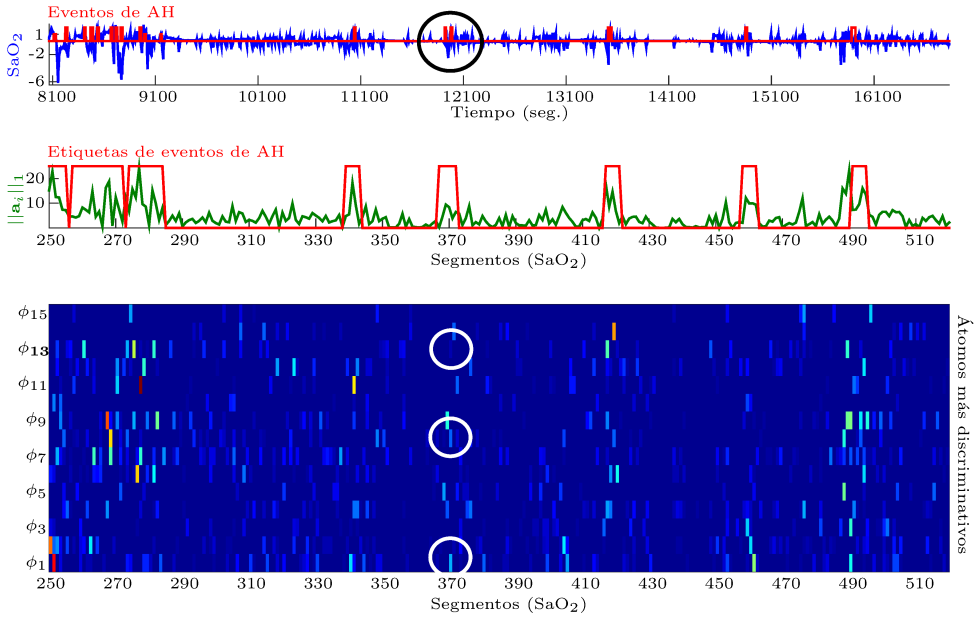


Figura 3.7: Correlación entre eventos de apnea-hipopnea y vectores de características finales.

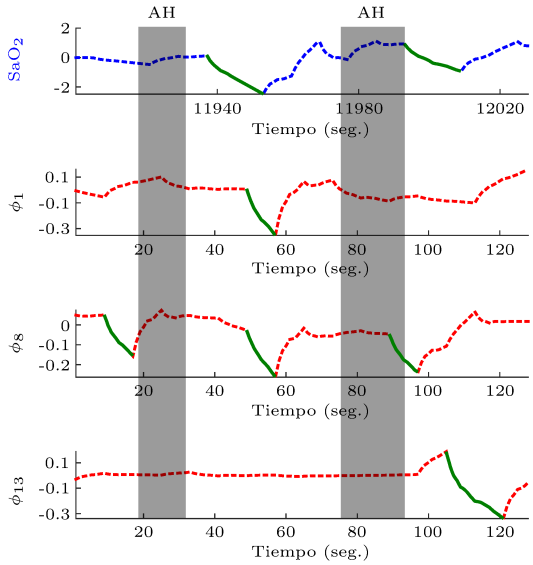


Figura 3.8: Formas de onda de *i*) un segmento de señal de SaO_2 (azul) y *ii*) tres de los átomos más discriminativos usados para obtener su representación rala (rojo). Marcas de los eventos de AH se han resaltado en color gris.

En forma análoga, uno podría suponer que la forma de onda del átomo ϕ_{13} tiene información relevante asociada al segundo evento de AH mientras que el átomo ϕ_8 aporta información compartida de ambos eventos respiratorios.

3.6. Conclusiones de este capítulo

En este capítulo se describió brevemente la técnica de Representaciones Ralas de las señales en términos de diccionarios discretos. Además, se presentó una nueva medida de discriminabilidad llamada DCAF, la cual no sólo es capaz de detectar cuando un átomo es discriminativo, sino también de cuantificar su grado de discriminabilidad para la clasificación. La nueva medida DCAF fue comparada con otras medidas de información muy conocidas del estado del arte logrando muy buenos resultados. Además, es importante señalar que la nueva medida DCAF es mucho más económica desde el punto de vista computacional ya que no requiere la estimación de las distribuciones condicionales de probabilidad.

Por otro lado, se comparó satisfactoriamente el nuevo método MDCS-OD propuesto en esta tesis con otros tres métodos del estado de arte logrando obtener muy buenos resultados. Más aún, el método MDCS-OD es el que logró obtener el máximo valor de AUC entre todos los métodos evaluados. Por último, se estableció una fuerte correlación entre las etiquetas de los eventos de AH (determinadas por los médicos expertos) y las activaciones de los átomos más discriminativos seleccionados por la nueva medida DCAF.

4 Aprendizaje de diccionarios estructurados

En la última década, la mayoría de los métodos de aprendizaje de diccionarios se usaron en varias tareas de clasificación de patrones logrando resultados muy prometedores [75]. Si bien estos métodos producen representaciones ralas de señales que son robustas ante distorsiones y desconexiones en el proceso de adquisición de las señales, tales representaciones resultan a menudo insatisfactorias si el objetivo final es la clasificación. Para superar (o al menos atenuar) esta debilidad, en los últimos años han surgido nuevos métodos que incorporan "información discriminativa" en los modelos que inducen raleza en las representaciones. En particular, se ha visto que los métodos que incorporan información discriminativa en el proceso de aprendizaje del diccionario logran mejores tasas de acierto que los métodos tradicionales, los cuales sólo se enfocan en minimizar el error total en la representación.

Recientemente, se ha observado un interés creciente en el desarrollo de nuevos métodos basados en representaciones ralas de señales para la clasificación [76, 77, 78]. Por ejemplo, una versión discriminativa del método estándar K-SVD aplicado al reconocimiento de rostros fue presentado por Zhang Q. *et al.* [76]. En ese trabajo, los autores incluyeron un término de discriminabilidad en la función objetivo del método estándar K-SVD. Los resultados muestran que esta modificación permite aprender diccionarios que resultan en bajo error en la reconstrucción y alta tasa de acierto. Además, Pham D. *et al.* [77] propuso un método iterativo que optimiza en simultáneo un diccionario y un clasificador lineal. El método desarrollado fue usado con éxito en un problema de categorización de imágenes. Más recientemente, un nuevo enfoque llamado "Label Consistent KSVD" (LC-KSVD) para aprendizaje de diccionarios discriminativos fue propuesto en [78]. En ese trabajo se integró eficientemente un término discriminativo y un clasificador lineal en la función objetivo.

No obstante, además de los métodos de aprendizaje de diccionarios en forma supervisada, se han propuesto nuevas opciones [79, 80, 81]. Estas nuevas alternativas se basan principalmente en la búsqueda de discriminabilidad en representaciones ralas

a través del desarrollo de nuevos métodos de aprendizaje de diccionarios "estructurados". El marco introducido en [79] aborda el aprendizaje de múltiples diccionarios que son simultáneamente reconstructivos y discriminativos, y el uso de los errores de reconstrucción de estos diccionarios en parches de imágenes que deriva en una clasificación por píxeles. Este algoritmo ha demostrado ser robusto, especialmente para tareas de clasificación de imágenes locales. En [80] se propuso un método para aprender múltiples diccionarios (no redundantes) para la categorización de objetos complejos. Este método se evaluó tanto en la categorización de objetos visuales como en los problemas relacionados con la clasificación de imágenes de documentos logrando muy buenos resultados. En [81], se introdujo un método que optimiza simultáneamente un diccionario estructurado (palabras visuales de categoría específica por cada característica) y un clasificador. Este método arrojó buenas tasas de reconocimiento, mostrando una mejora significativa con respecto a los métodos de clasificación de objetos más avanzados. Un nuevo método para el aprendizaje estructurado de diccionarios fue propuesto recientemente por Sun *et al.* [82]. En ese trabajo, el diccionario aprendido se particionó en sub-diccionarios específicos para cada clase para la clasificación, la cual se realiza midiendo el error mínimo de reconstrucción entre todas las clases. El método se probó utilizando tanto los datos sintéticos como los datos del mundo real que muestran un buen rendimiento.

En este capítulo se presentan una nueva medida de discriminabilidad de los átomos de un diccionario dado en el contexto de problemas de clasificación multi-clase y un nuevo método iterativo para la construcción de diccionarios estructurados. Esta nueva medida no sólo tiene en cuenta la probabilidad de activación condicional de los átomos y la magnitud de su correspondiente coeficiente de representación, sino también el efecto que dicho átomo tiene en el error total de representación. Asimismo, la nueva medida es capaz de cuantificar eficazmente el grado de información discriminativa que tienen los átomos en este contexto. Por otro lado, el nuevo método produce diccionarios estructurados muy adecuados para tareas de clasificación multi-clase.

4.1. Criterio de discriminabilidad multi-clase

En el capítulo anterior se presentó una nueva medida que permite cuantificar el grado de información discriminativa que tienen los átomos de un diccionario dado

en el contexto de problemas de clasificación binaria. En esta sección, se introduce una extensión de esa medida a problemas de clasificación multi-clase. Esta generalización consiste en definir y usar una nueva función multi-objetivo que permita cuantificar el grado de información discriminativa de los átomos de un diccionario en este contexto. Esta función se definirá como una combinación convexa de tres términos discriminativos, todos ellos basados en representaciones ralas afín de las señales. A continuación se presenta una descripción detallada de cada uno de estos términos, así como una definición formal de la función.

Frecuencia de activación condicional: conceptualmente, la frecuencia de activación condicional puede considerarse como un punto de partida "razonable" para determinar el grado de información que "aportan" los átomos de un diccionario para la clasificación de señales. Por esta razón, este nuevo enfoque comienza por calcular la frecuencia de activación η_ℓ^j de cada átomo ϕ_j, $j = 1, 2, \cdots, M$, dada la clase ℓ, para $\ell = 1, 2, \cdots, k$. Asimismo, la probabilidad de activación condicional de ϕ_j dada la clase ℓ se define como $p_\ell^j \doteq P(a_j \neq 0 | \mathbf{x} \in \ell)$. Además, dado un conjunto de n_ℓ señales clase ℓ, esta probabilidad condicional puede aproximarse mediante el cociente η_ℓ^j / n_ℓ.

Para un cierto j, $1 \leq j \leq M$, se denotará por ℓ_j^+ a la clase que maximiza todas las probabilidades de activación condicionales p_ℓ^j, para todo $\ell = 1, 2, \cdots, k$, es decir tal que

$$p_{\ell_j^+}^j = \max_{1 \leq \ell \leq k} p_\ell^j. \tag{4.1}$$

En caso de haber más de un valor de ℓ que maximice p_ℓ^j, se define ℓ_j^+ mediante la selección aleatoria de una de ellas, por ejemplo la "menor" (notar que el orden en que se definen las clases es completamente irrelevante). Del mismo modo, para un j particular, $1 \leq j \leq M$, se define ℓ_j^* como la clase que dar lugar al segundo valor mayor de p_ℓ^j, es decir

$$p_{\ell_j^*}^j = \max_{\substack{1 \leq \ell \leq k \\ \ell \neq \ell_j^+}} p_\ell^j. \tag{4.2}$$

Nuevamente, en caso de existir más de un valor de ℓ_j^* que satisfaga (4.2), ℓ_j^* se selecciona en forma aleatoria entre todas las clases candidatas.

Luego, se define la función $m_{af} : \{1, 2, \cdots, M\} \to \mathbb{R}_0^+$ por

$$m_{af}(j) \doteq \frac{p^j_{\ell_j^+} - p^j_{\ell_j^*}}{p^j_{\ell_j^+}}, \tag{4.3}$$

se referirá a $m_{af}(\cdot)$ como la "medida de frecuencia de activación".

Notar que $0 \leq m_{af}(\cdot) \leq 1$. Se dice que el átomo ϕ_j es "discriminativo" (para la clase ℓ_j^+) si y sólo si $m_{af}(j) > 0$. Claramente, en este contexto, si ϕ_j es discriminativo, éste lo será sólo para la clase ℓ_j^+, de lo contrario no lo será para ninguna de ellas. Más aún, el valor de $m_{af}(j)$ puede pensarse como una "medida" del grado de discriminabilidad de ϕ_j basado solamente en la información de su frecuencia de activación.

Con el fin de fortalecer la interpretación conceptual de la frecuencia de activación condicional de los átomos, en esta tesis se incluye una representación gráfica de todas las representaciones ralas, matriz $\mathbf{A} \doteq [\mathbf{A}_1\ \mathbf{A}_2\ \mathbf{A}_3\ \mathbf{A}_4]$, de $\mathbf{X} \doteq [\mathbf{X}_1\ \mathbf{X}_2\ \mathbf{X}_3\ \mathbf{X}_4]$ en términos de Φ a través de $\mathbf{X} \approx \Phi\mathbf{A}$. La Figura 4.1 muestra todas las representaciones ralas $\{\mathbf{A}_\ell\}$ (para cada una de las 4 clases) del conjunto de señales $\{\mathbf{X}_\ell\}$ en términos de Φ. En la imagen se distinguieron en color naranja por un lado el átomo ϕ_j (la j-

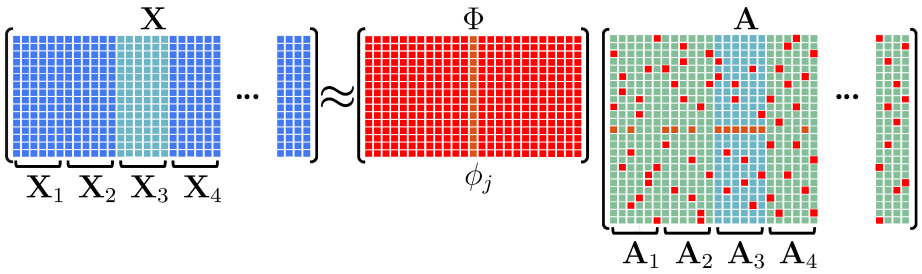


Figura 4.1: Ilustración de las activaciones del átomo ϕ_j para datos pertenecientes a cuatro clases diferentes.

ésima columna) de Φ y, por el otro, sus correspondientes activaciones a_j (la j-ésima fila) de $\mathbf{A}$. Se puede observar que ϕ_j se activa (usa para representar) con mayor frecuencia para datos clase $\ell = 3$ que para el resto de las clases. Más aún, dado que ϕ_j se activa en la totalidad de los casos para representar señales clase $\ell = 3$, la probabilidad de activación de ϕ_j dada $\ell = 3$ es máxima, es decir, $p^j_3 = 1$. Por lo tanto, de acuerdo a la medida de frecuencia de activación, ϕ_j es discriminativo para

señales clase $\ell = 3$. Por otro lado, la Figura 4.2 ilustra un gráfico de barras de los distintos valores de p_ℓ^j, para las clases $\ell = 1, 2, 3, 4$. Las barras del gráfico muestran claramente que ϕ_j es discriminativo para señales clase $\ell = 3$. Además, en este caso en particular, $\ell_j^+ = 3$ y $\ell_j^* = 2$.

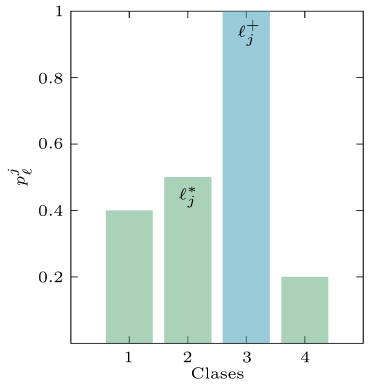


Figura 4.2: Magnitudes de las frecuencias de activación condicionales de ϕ_j para cada una de las 4 clases.

Magnitud del coeficiente: las representaciones ralas de señales no sólo brindan información relevante respecto a la "activación" de los átomos, sino también éstas pueden enfatizar características importantes inmersas en ciertas formas de onda asociadas a eventos particulares en señales o imágenes. Con esta observación en mente, se procede a definir una nueva (segunda) medida que tiene en cuentas magnitudes de los coeficientes de representación. Para ello, dado un átomo ϕ_j, sean ℓ_j^+ y ℓ_j^* las clases definidas en (4.1) y (4.2), respectivamente, y sean $\mathbf{A}_{\ell_j^+}$ y $\mathbf{A}_{\ell_j^*}$ las matrices que proveen las representaciones ralas de $\mathbf{X}_{\ell_j^+}$ y $\mathbf{X}_{\ell_j^*}$, respectivamente, en términos de Φ, es decir, $\mathbf{X}_{\ell_j^+} = \Phi\mathbf{A}_{\ell_j^+}$ y $\mathbf{X}_{\ell_j^*} = \Phi\mathbf{A}_{\ell_j^*}$. Adicionalmente, se denota q_ℓ^j al cociente $||\left[\mathbf{A}_\ell\right]_{j,:}||_1/n_\ell$, donde $\left[\mathbf{A}_\ell\right]_{j,:}$ representa la j-ésima fila de la matriz $\mathbf{A}_\ell$. La medida de la magnitud del coeficiente es la función $m_{cm} : \{1, 2, \cdots, M\} \to \mathbb{R}_0^+$ definida por

$$m_{cm}(j) \doteq \frac{q_{\ell_j^+}^j - q_{\ell_j^*}^j}{q_{\ell_j^+}^j}. \tag{4.4}$$

Nuevamente $0 \leq m_{cm}(\cdot) \leq 1$. Considerando esta nueva medida, un átomo ϕ_j se considera discriminativo (para la clase ℓ_j^+) si y sólo si $m_{cm}(j) > 0$ y, en ese caso, el

valor de $m_{cm}(j)$ cuantifica el correspondiente grado de discriminabilidad de ϕ_j para la clase ℓ_j^+.

Error en la representación: por último se procede a describir la tercer medida introducida en esta tesis para cuantificar el grado de discriminabilidad de los átomos de un diccionario. En particular, esta nueva medida considera la contribución de cada átomo ϕ_j al error total de representación. Sea $\mathbf{A}_\ell \doteq [\mathbf{a}_1\ \mathbf{a}_2\ \cdots \mathbf{a}_{n_\ell}]$ la matriz que provee la representación rala de $\mathbf{X}_\ell \doteq [\mathbf{x}_1\ \mathbf{x}_2\ \cdots \mathbf{x}_{n_\ell}]$, tal como en la medida presentada anteriormente. Claramente, la contribución de la clase ℓ al error total de representación se puede escribir como [39]

$$\begin{aligned}
\sum_{i=1}^{n_\ell} ||\mathbf{x}_i - \Phi\mathbf{a}_i||_2^2 &= \|\mathbf{X}_\ell - \Phi\mathbf{A}_\ell\|_F^2 \\
&= \left\|\mathbf{X}_\ell - \sum_{j=1}^{M} \phi_j \left[\mathbf{A}_\ell\right]_{j,:}\right\|_F^2 \\
&= \left\|\left(\mathbf{X}_\ell - \sum_{i\neq j} \phi_i \left[\mathbf{A}_\ell\right]_{i,:}\right) - \phi_j \left[\mathbf{A}_\ell\right]_{j,:}\right\|_F^2 \\
&\doteq \left\|\mathbf{E}_\ell^j - \phi_j \left[\mathbf{A}_\ell\right]_{j,:}\right\|_F^2, \qquad (4.5)
\end{aligned}$$

donde $\mathbf{E}_\ell^j$ denota el error total de representación para todas las señales clase ℓ cuando se descarta ϕ_j. Por lo tanto, un valor grande de $\mathbf{E}_\ell^j$ indica que la contribución de ϕ_j a la representación de señales clase ℓ es importante. Se define entonces la medida de error en la representación $m_{re}:\ \{1, 2, \cdots, M\} \to \mathbb{R}_0^+$ por

$$m_{re}(j) \doteq \frac{r_{\ell_j^+}^j - r_{\ell_j^*}^j}{r_{\ell_j^+}^j}, \qquad (4.6)$$

donde $r_\ell^j \doteq \mathbf{E}_\ell^j / n_\ell$, para $\ell = 1, 2, \cdots, k$, $j = 1, 2, \cdots, M$.

Nuevamente $0 \leq m_{re}(\cdot) \leq 1$, y se dice que ϕ_j es discriminativo (para la clase ℓ_j^+) respecto a esta medida si y sólo si $m_{re}(j) > 0$. En ese caso, el valor de $m_{re}(j)$ cuantifica el correspondiente grado de discriminabilidad.

Medida combinada a raíz de las distintas formas de cuantificar la discriminabilidad de los átomos, es razonable pensar en una medida que las combine apro-

piadamente. Teniendo en cuenta esto, dados dos parámetros positivos α y β, con $\alpha + \beta \leq 1$, se define la función $m_{\alpha,\beta} : \{1, 2, \cdots, M\} \to \mathbb{R}_0^+$ como

$$m_{\alpha,\beta}(j) \doteq \alpha\, m_{af}(j) + \beta\, m_{cm}(j) + (1 - \alpha - \beta)\, m_{re}(j). \tag{4.7}$$

En esta tesis se referirá a $m_{\alpha,\beta}$ como la "medida de discriminabilidad combinada". Claramente, como α y β varían entre 0 y 1, (4.7) evalúa todas las combinaciones convexas posibles de las tres medidas individuales m_{af}, m_{cm} y m_{re}. Sin embargo, hallar el par de parámetros óptimos (α^*, β^*) que conducen al mejor desempeño del clasificador, es un problema muy difícil de resolver.

En la próxima sección se introduce el nuevo método de aprendizaje de diccionarios estructurados, el cual se basa en la medida de discriminabilidad combinada para la selección de átomos discriminativos en el contesto de problemas de clasificación multi-clase.

4.2. Método propuesto

Los métodos supervisados para el aprendizaje de diccionarios han despertado un gran interés en los últimos años. Originalmente, esos métodos se enfocaron sólo en el aprendizaje eficiente de diccionarios discretos simples (no estructurados) que incorporan información "discriminativa" (en términos de la clasificación de señales) en su proceso de optimización. Esta información se puede introducir al modelo de aprendizaje mediante la incorporación de diferentes criterios discriminativos [83, 84, 85]. Entre los criterios más comúnmente usados se puede nombrar a la función de coste "softmax", el criterio de Fisher y el error de clasificación predictiva lineal, por nombrar sólo algunos de ellos [76, 77, 80, 75].

Si bien existen distintas formas de optimizar en simultáneo un diccionario y un clasificador, una estrategia utilizada muy a menudo consiste simplemente en dividir ese problema en dos sub-problemas [79, 80]. De esta manera, es posible utilizar todos los métodos de aprendizaje de diccionarios tradicionales disponibles, tales como MOD y KSVD, y por lo tanto poder entrenar el clasificador en una etapa posterior. El método iterativo que se introduce en esta tesis no sólo se basa en esta estrategia sino también, en cada iteración, selecciona los átomos más discriminativos de cada

clase para la construcción del diccionario estructurado final. Para ello, se propone un nuevo método de aprendizaje de diccionario estructurado multi-clase denotado por DAS-KSVD (del inglés *Discriminant Atom Selection KSVD*) en el que se utiliza la medida de discriminabilidad combinada $m_{\alpha,\beta}$ para seleccionar eficientemente los átomos discriminativos específicos de la clase a partir de algunos diccionarios "auxiliares" dados para construir uno estructurado de forma iterativa.

Si bien en esta tesis no se profundiza en la descripción de la implementación del método DAS-KSVD, a continuación se procederá a describir los aspectos más importantes del mismo. Para más detalles, se refiere al lector a la Sección 3.3 del articulo incluido en el Anexo D.

Una de las principales hipótesis planteadas para el desarrollo del método DAS-KSVD es que se pueden aprender (y detectar) átomos discriminativos en forma iterativa que modelen características importantes de cierto grupo de señales que no han sido modeladas por ninguno de los demás átomos. Asimismo, una de las formas de generar átomos consiste en aprenderlos teniendo en cuenta "distintos" grupos de señales en cada iteración. Por esta razón, el método DAS-KSVD tiene como objetivo principal aprender un diccionario estructurado denotado por $\Phi_D^{(I)} \doteq [\Phi_1\ \Phi_2\ \cdots\ \Phi_k]$, el cual esta formado por k sub-diccionarios Φ_ℓ (apilados uno al lado del otro), $\ell = 1, 2, \cdots, k$, cada uno de ellos construido a partir de I átomos discriminativos, de acuerdo a $m_{\alpha,\beta}$, para señales clase ℓ.

A partir de este momento se considerarán los vectores $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ como realizaciones de un vector aleatorio $\mathcal{X}$ de dimensión N. Asimismo, para el aprendizaje del diccionario estructurado, el método requiere de los siguientes parámetros: la matriz de señales denotada por $\mathbf{X}_{trn} \doteq [\mathbf{X}_1\ \mathbf{X}_2\ \cdots\ \mathbf{X}_k]$, de tamaño $N \times n$, compuesta por $n = \sum_{\ell=1}^{k} n_\ell$ muestras, el nivel de raleza requerido q, el factor de redundancia r_f, el número t de señales de entrenamiento pertenecientes a la clase ℓ, $t \ll n_\ell$, el número de iteraciones I y el vector de etiquetas de clases $\mathbf{c}$.

El algoritmo propuesto asigna una distribución de probabilidad uniforme inicial p_0 sobre $\mathbf{X}_{trn}$ de modo que $p_0(i) = 1/n$, para todo $i = 1, 2, \cdots, n$. El valor de $p_0(i)$ es la probabilidad de que $\mathbf{x}_i$ sea seleccionada de $\mathbf{X}_{trn}$ para formar parte de una nueva matriz de aprendizaje denotada por $\mathbf{X}_{lrn}$, para aprender el diccionario inicial Φ. Además, si una determinada señal de entrenamiento $\mathbf{x}_i$ se utiliza para aprender el diccionario Φ en una iteración en particular, entonces es deseable que dicha

señal tenga menor probabilidad de ser seleccionada que las otras en las siguientes iteraciones. Por lo tanto, promoviendo diversidad de esta manera, se podría pensar que los átomos aprendidos tienen la capacidad de modelar las diferentes propiedades intrínsecas de las señales.

El enfoque discriminativo propuesto consiste en optimizar y utilizar la nueva medida de discriminabilidad combinada $m_{\alpha,\beta}$ para seleccionar los átomos más discriminativos de Φ para cada una de las k clases. Como se explica en la sección anterior, el valor de $m_{\alpha,\beta}(j)$ cuantifica al grado de discriminabilidad del átomo ϕ_j para una (y sólo una) clase, la cual se indica con ℓ_j^+. No obstante, el proceso de selección de los átomos más discriminativos conlleva un problema muy complejo, puesto que encontrar el par óptimo de parámetros (α^*, β^*) es un problema muy difícil de resolver. Para más detalles sobre el ajuste de ese par de parámetros, se remite al lector al artículo incluido en el Anexo D. Además, la construcción del sub-diccionario consiste básicamente en tomar uno por uno los átomos más discriminativos de Φ para cada una de las clases de k y apilarlos lado a lado. En el caso de que haya más de un candidato relacionado con la clase que cumpla con el criterio discriminativo propuesto, ϕ_j se define como el átomo que maximiza todos los valores posibles de m_{α^*,β^*}. De lo contrario, en caso de que Φ no posea átomos discriminativos, se reinicia el proceso de muestreo de las señales.

4.3. Resultados y discusión

Finalmente se evaluó el desempeño del método DAS-KSVD en dos escenarios diferentes dentro del contexto de problemas de clasificación multi-clase. En un principio, se consideró un problema clásico de clasificación de dígitos manuscritos y, posteriormente, se tuvo en cuenta el problema del diagnóstico del SAHOS moderado a partir de la detección de los eventos de apnea y de hipopnea en forma separada. A continuación, se presentan los resultados obtenidos en cada uno de esos escenarios.

Análisis del desempeño de DAS-KSVD: con el objetivo de evaluar y comparar el desempeño del método propuesto se hizo uso de una de las bases de datos más utilizadas para validar métodos en el área de Visión Computacional (*Computer Vision*) y de Reconocimiento de Patrones (*Pattern Recognition*), denominada MNIST (del inglés *Modified NIST*) [86]. Esta base de datos clásica ha sido amplia-

mente utilizada para evaluar nuevos métodos incluyendo, por ejemplo, técnicas de Representaciones Ralas, Aprendizaje Profundo, ELM y varios tipos de redes neuronales [40, 87, 88, 89, 90]. La base de datos MNIST contiene un total de 70.000 imágenes normalizadas y centradas en escala de grises de dígitos manuscritos que varían desde el 0 (cero) hasta el 9 (nueve), cada uno de tamaño 28×28. Así mismo, el número de imágenes por clase varía de 5.421 a 6.742, correspondientes a las clases $\ell = 5$ y $\ell = 1$, respectivamente. Los resultados obtenidos al evaluar el método DAS-KSVD para el reconocimiento de dígitos manuscritos superan a los obtenidos por técnicas equivalentes del estado del arte. Para más detalles, se remite al lector al artículo incluido en el Anexo D.

Aplicación del método propuesto: por otro lado, se evaluó el desempeño del nuevo método para el diagnóstico simplificado del SAHOS. En este sentido, se planteó una extensión del problema de clasificación binaria (presencia o ausencia de eventos de AH) mencionado en el capítulo anterior a un problema de clasificación multi-clase. En particular, se tuvieron en cuenta tres estados (clases) bien diferenciados en el registro de las señales de SaO_2 denotados por respiración normal (N), evento de apnea (A) y evento de hipopnea (H) correspondientes a las clases $\ell = 1$, $\ell = 2$ y $\ell = 3$, respectivamente. Como ya se indicó anteriormente, desde el punto de vista clínico, la correcta estimación del IAH es muy importante para determinar la severidad del trastorno y, por lo tanto, tratar de forma adecuada al paciente [91]. Asimismo, poder identificar los eventos de apnea y de hipopnea en forma separada durante el estudio, brinda al médico especialista mayor información acerca de la dinámica del trastorno y su severidad. Por ejemplo, puede suceder que dos personas tengan el mismo IAH pero distintos índices de eventos de apnea (IA) e índices de eventos de hipopnea (IH). Claramente, si una persona que tiene un SAHOS severo y, además, casi la totalidad de los eventos son de apnea (IA alto), esta persona requiere un tratamiento y un seguimiento especial.

La Figura 4.3 muestra las representaciones gráficas de una pequeña porción de una señal de SaO_2 cruda (arriba) y filtrada mediante el uso de los filtros wavelet usados en esta tesis (medio). La parte inferior de esta figura ilustra las etiquetas de cada una de las tres clases correspondientes a los eventos N, A y H. Si se observa en detalle los dos últimos eventos de hipopnea y apnea (regiones sombreadas en color gris), se puede notar que el valor mínimo al que llega la señal es muy similar

en ambos eventos. Además, a pesar de que las desaturaciones en la señal tienen pendientes diferentes, se puede observar fácilmente que ambos eventos poseen una morfología muy parecida. Estos son sólo algunos de los principales factores que reflejan la dificultad en poder discriminar entre estas dos clases de eventos a partir de la oximetría de pulso.

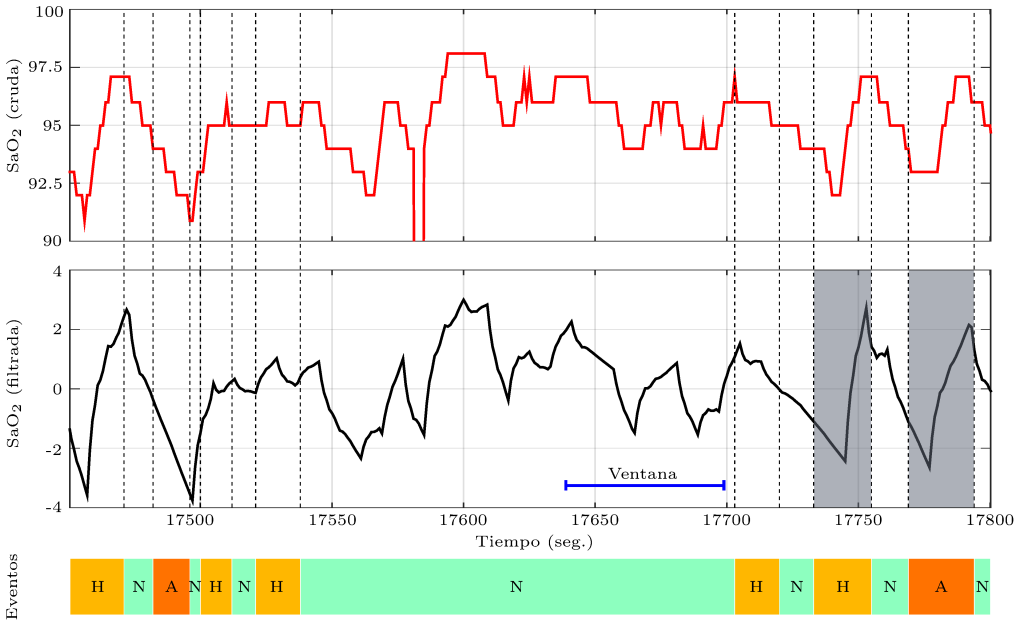


Figura 4.3: Registro de la señal de SaO_2 cruda (arriba), filtrada (medio) y etiquetas de los distintos eventos respiratorios que ocurren durante el sueño (abajo).

Claramente, la detección "individual" de los eventos de apnea y de hipopnea conlleva un problema mucho más difícil de resolver que el de clasificación binaria planteado en el capítulo anterior que sólo tiene en cuenta la presencia (o no) de tales eventos, independientemente del tipo que se trate. Con el objetivo de estimar de forma cualitativa el grado de solapamiento de las clases en los datos, se hizo uso del método de reducción de dimensionalidad denotado por *Mapeo de Sammon* [92]. La Figura 4.4 muestra la distribución de los datos de entrenamiento usados en esta tesis respecto a las etiquetas de respiración normal y de los eventos de apnea y de hipopnea. Es importante notar que la distribución de los datos en el espacio de los atributos se solapan entre sí, lo que dificulta la distinción entre las clases, incluso considerando representaciones en un espacio de alta dimensionalidad (que utilizan

un gran número de atributos). Asimismo, los datos clase N son los que parecerían tener el menor grado de solapamiento respecto a las clases A y H. Por otro lado, los datos clase H son los que presentan un mayor grado de solapamiento tanto con los datos clase N como con los datos clase A. Más aún, se observa que el mayor grado de solapamiento esta dado entre los datos pertenecientes a las clases A y H. Un análisis similar de la distribución de los eventos N, A y H en la señal de flujo respiratorio se puede encontrar en el trabajo propuesto por Koley *et al.* [93].

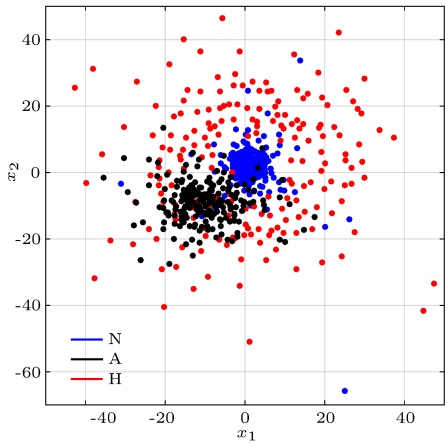


Figura 4.4: Distribución de los eventos N, A y H luego de aplicar el método de reducción de dimensionalidad denominado *Mapeo de Sammon*, en dos de sus atributos más importantes obtenidos a partir de la señal de SaO_2 (estimada a partir de 200 ejemplos de cada clase del conjunto de entrenamiento).

A continuación se mencionan brevemente los valores de los principales parámetros del método DAS-KSVD utilizados en los experimentos. En primer lugar, se fijó el número de iteraciones (I) del algoritmo en 64. Por lo tanto, el tamaño del diccionario estructurado final en 128×192. En otras palabras, $\Phi_D^{(I)} \doteq [\Phi_1\ \Phi_2\ \Phi_3]$, donde Φ_1, Φ_2 y Φ_3 corresponden a los sub-diccionarios discriminativos para las clases $\ell = 1$, $\ell = 2$ y $\ell = 3$, respectivamente. Para la detección de los átomos más discriminativos para cada una de las clases, se usó la medida de discriminabilidad combinada $m_{\alpha,\beta}$ con el par óptimo de parámetros (α^*, β^*) nulo tal como el obtenido en el artículo incluido en el Anexo D. Por lo tanto, los átomos más discriminativos no sólo serán aquellos con mayor probabilidad de activación condicional, sino también serán los que minimicen el error total en la representación de las señales. La Figura 4.5

muestra una representación gráfica de los 6 átomos más discriminativos, correspondientes a las primeras 6 iteraciones del algoritmo, seleccionados por la medida de discriminabilidad combinada para cada una de las 3 clases. En la parte superior de esta figura se pueden observar los átomos que tienen mayor información discriminativa para los segmentos de respiración normal. Además, uno podría pensar que la forma de onda de esos átomos representa variaciones cuasi-periódicas asociadas al ritmo respiratorio normal, típicas de la oximetría de pulso. Por otro lado, las partes central e inferior de la figura ilustran los átomos más discriminativos para los segmentos con eventos de apnea e hipopnea, respectivamente. Si bien las formas de onda de los átomos más discriminativos de ambos eventos parecen ser similares, las desaturaciones son más visibles y abruptas en los átomos correspondientes a los eventos de apnea. Este hecho se asocia claramente con lo que ocurre en la realidad ya que, como se mencionó en el Capítulo 1, estos eventos respiratorios se producen como consecuencia de la obstrucción total de la vía respiratoria durante el sueño.

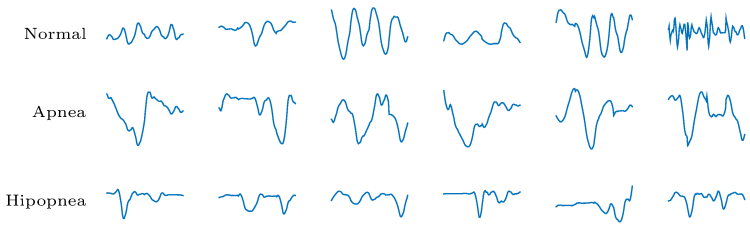


Figura 4.5: Representación gráfica de algunos de los átomos que conforman los subdiccionarios Φ_1 (arriba), Φ_2 (medio) y Φ_3 (abajo).

A partir del diccionario estructurado construido ($\Phi_D^{(I)}$), se entrenó y evaluó el desempeño de un clasificador del tipo MLP con una capa oculta con tres neuronas en la capa de salida para la detección de los eventos N, A y H. En particular, para desarrollar los experimentos de este capítulo, se usó un MLP con 500 neuronas en la capa oculta, función de activación "tansig" y el método de aprendizaje gradiente descendente conjugado. Para realizar esta tarea, se seleccionó al azar un grupo balanceado de 30.000 segmentos de señales de SaO_2 y se construyó una nueva matriz denotada por $\hat{\mathbf{X}} \doteq [\mathbf{X}_1\ \mathbf{X}_2\ \mathbf{X}_3]$, de tamaño 128×30.000, donde $\mathbf{X}_\ell$, para $\ell = 1, 2, 3$, es la matriz que contiene segmentos pertenecientes a la clase ℓ. Para entrenar el clasificador, la nueva matriz $\hat{\mathbf{X}}$ se particionó en 3 sub-matrices denotadas por $\hat{\mathbf{X}}_{trn}$,

$\hat{\mathbf{X}}_{val}$ y $\hat{\mathbf{X}}_{tst}$ para el entrenamiento (70 %, 21.000 ejemplos), validación (15 %, 4.500 ejemplos) y prueba (15 %, 4.500 ejemplos), respectivamente. La Figura 4.6 muestra las matrices de confusión obtenidas para el entrenamiento, la validación y la prueba del clasificador propuesto y una matriz de confusión que representa el promedio general. En particular, se procederá a analizar los resultados obtenidos con los datos de prueba (matriz inferior-izquierda de la figura). Las primeras 3 filas de esta matriz representan los eventos detectados por el clasificador, mientras que las primeras 3 columnas representan las etiquetas (verdadero valor) de los eventos. Los primeros 3 elementos de la diagonal principal de esta matriz (cuadros en color verde) representan el número total de eventos detectados correctamente (parte superior) y la tasa relativa de verdaderos positivos (parte inferior). Asimismo, la última fila de la matriz representa la tasa de reconocimiento (Acc) para cada una de las clases. En particular, se obtuvieron valores de Acc de 86,1 %, 63,2 % y 23,4 % correspondientes a las clases N, A y H. Por otro lado, se logró una tasa de reconocimiento promedio de 58,2 %. Claramente, y como era de esperar, el desempeño del clasificador no fue tan bueno en la detección de segmentos pertenecientes a la clase H. Esto se puede explicar desde el punto de vista de las distribuciones de las clases en los datos, ya que la distribución de los eventos de hipopnea se encuentran muy solapados tanto con los datos de la clase N como con los datos de la clase A. Sin embargo, si unificamos las clases A y H, la tasa de acierto global para la detección de eventos (independientemente de la clase de evento que se trate) es del orden del 64 %, la cual puede compararse con el desempeño (promedio) obtenido con los métodos presentados en el capítulo anterior (ver Tabla 3.1).

Finalmente, se evaluó el desempeño del método DAS-KSVD en el diagnóstico del SAHOS moderado, es decir, se fijó el umbral de detección de la patología en 15 eventos de AH por hora de estudio. Para realizar este experimento y con el fin de comparar este nuevo método con el introducido en el capítulo anterior, se evaluaron los mismos 287 estudios de prueba de la base de datos. Estos datos de prueba no se tuvieron en cuenta para aprender el diccionario estructurado y para el entrenamiento de la red neuronal. La Figura 4.7 muestra las curvas ROC construidas al evaluar el nuevo método DAS-KSVD y el método MDCS-OD en color rojo y azul, respectivamente, para el diagnóstico del SAHOS moderado. Puesto que el punto de corte (umbral de detección del método) elegido es aquel que maximiza la sensibilidad y la

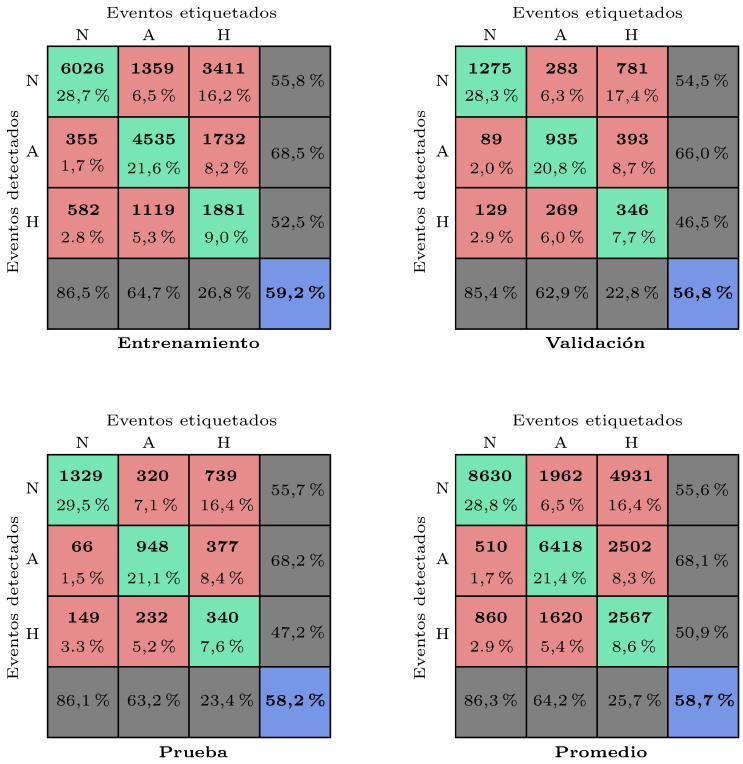


Figura 4.6: Matrices de confusión. Primera fila: entrenamiento (izquierda) y validación (derecha). Segunda fila: prueba (izquierda) y promedio general (derecha).

especificidad en simultáneo, las pruebas de laboratorio arrojaron que el nuevo método ha logrado obtener valores de AUC, sensibilidad y especificidad de 0,905, 87,04 % y 74,65 %, respectivamente. Estos resultados sugieren que las representaciones ralas de señales en términos de diccionarios estructurados proveen un marco apropiado no solo para el diagnóstico del SAHOS moderado, sino también para la detección de los eventos de apnea e hipopnea en forma separada.

Por otro lado, se comparó el desempeño logrado por cada uno de los métodos propuestos en esta tesis observando que ambos tienen un muy buen desempeño superiores al 0,9 de AUC. Se observa que el método MDCS-OD supera al método DAS-KSVD en cuanto a las medidas AUC y especificidad, logrando obtener un 0,937 y 85,65 %, respectivamente (ver Tabla 3.3). Sin embargo, se puede notar que el nuevo método DAS-KSVD supera al método MDCS-OD en términos de la sensi-

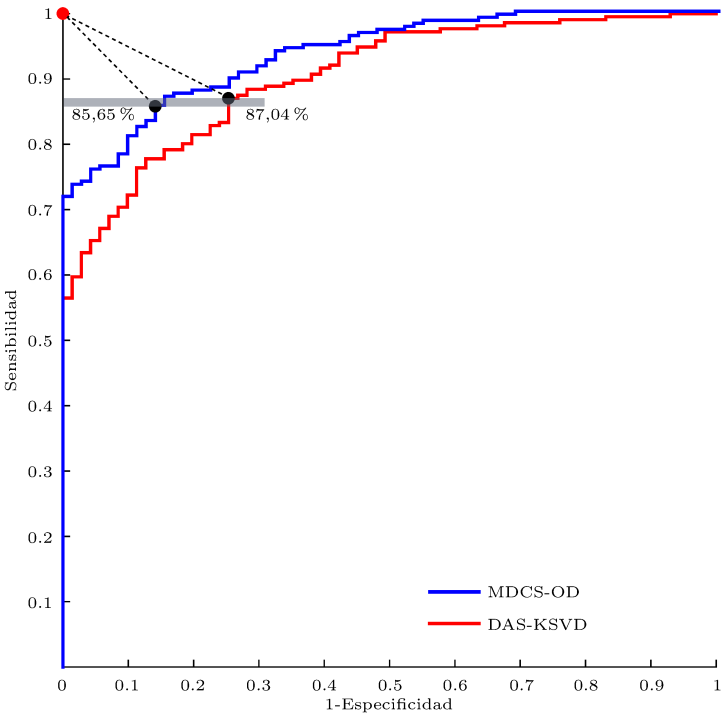


Figura 4.7: Curvas ROC construidas a partir de los métodos MDCS-OD (azul) y DAS-KSVD (rojo) para el diagnóstico del SAHOS moderado.

bilidad, logrando obtener un 87,04 %. Se hace notar también que esta sensibilidad es mayor que la obtenida por los otros métodos del estado del arte [23, 30, 31]. Este valor representa un incremento del 1,39 % en la tasa de detección de pacientes que realmente están enfermos (zona gris de la figura), lo cual mejora la sensibilidad del diagnóstico. Si bien estos resultados son preliminares quedando aún margen para mejoras, indicarían que el aprendizaje de diccionarios estructurados proporciona información muy valiosa relacionada con los eventos de apnea y de hipopnea en la señal de SaO_2.

4.4. Conclusiones de este capítulo

En este capítulo se introdujo una extensión del problema de clasificación binaria presentado en el capítulo anterior a un problema de clasificación multi-clase. En este

contexto, se propuso una generalización de la medida DCAF (la cual tiene en cuenta solo dos clases en los datos) a más de dos clases. En particular, la nueva medida de discriminabilidad combinada no sólo tiene en cuenta la probabilidad condicional de activación de los átomos de un diccionario y el valor de su correspondiente coeficiente de activación, sino también considera el efecto que éste produce en el error total de representación. Además, utilizando esta medida, se presentó un nuevo método iterativo llamado DAS-KSVD para el aprendizaje de diccionarios estructurados en el contexto de problemas de clasificación multi-clase. Este nuevo método usa la medida de discriminabilidad combinada para detectar los átomos más discriminativos para cada una de las clases en los datos.

En primer lugar, se realizó un análisis del desempeño del método DAS-KSVD con el uso de una base de datos de dígitos manuscritos muy utilizada en la literatura logrando obtener tasas de reconocimiento superiores a los obtenidos por técnicas semejantes del estado del arte. Por otro lado, se evaluó satisfactoriamente la capacidad del nuevo método para el diagnóstico del SAHOS logrando obtener un muy buen desempeño en la detección de la patología. Más precisamente, se mejoró la sensibilidad del diagnóstico del SAHOS moderado.

5 Conclusiones y trabajos futuros

5.1. Conclusiones

En esta tesis se abordó el diseño, desarrollo, implementación y evaluación de tres métodos para el reconocimiento automático de los eventos de apnea-hipopnea a partir del análisis y procesamiento de las señales de SaO_2. Para ello se desarrollaron e implementaron algoritmos que utilizan herramientas de aprendizaje maquinal, de procesamiento de señales, de reconocimiento de patrones y de problemas inversos.

Como primera aproximación a la solución del problema, se desarrollaron dos métodos de selección de características denominados MDAS y MDCS, los cuales se basan en representaciones ralas de señales de SaO_2 para el reconocimiento de eventos de AH. Además, en esta tesis se introdujo una nueva medida de discriminabilidad binaria denotada por DCAF, la cual es capaz de detectar átomos discriminativos en un diccionario. Asimismo, esta medida permite cuantificar eficientemente sus correspondientes grados de discriminabilidad, lo cual resulta útil a los efectos de la clasificación. Los métodos MDAS y MDCS hacen uso de la media DCAF para detectar los átomos más discriminativos de un diccionario dado y, a partir de ellos, realizan la selección de características. En particular, el método MDCS utiliza la medida DCAF para seleccionar los átomos más discriminativos y, a partir de ellos, construir un sub-diccionario. En base a los experimentos desarrollados en esta tesis, el desempeño de la nueva medida DCAF fue comparada con el de varias otras medidas de información del estado del arte. Los resultados muestran que DCAF logró un muy buen desempeño. Además, es importante señalar que, desde el punto de vista computacional, la nueva medida DCAF es mucho más económica ya que no requiere de la estimación de las distribuciones de probabilidad condicionales. Por otro lado, el nuevo método MDCS fue comparado con otros tres métodos del estado de arte, superando significativamente el desempeño de todos ellos. Más aún, el método MDCS logró obtener el máximo valor de AUC entre todos los métodos evaluados. Por último, se mostró que existe una muy fuerte correlación entre las etiquetas de

los eventos de AH (determinadas por los médicos expertos) y las activaciones de los átomos más discriminativos seleccionados por la nueva medida DCAF.

Con el fin de lograr un diagnóstico del SAHOS moderado más certero y de mejorar su tratamiento, se introdujo una extensión del problema de clasificación binaria a uno multi-clase. En este contexto, se propuso una generalización de la medida DCAF (la cual tiene en cuenta solo dos clases en los datos) a más de dos clases. En particular, la nueva medida de discriminabilidad combinada no solo tiene en cuenta la probabilidad condicional de activación de los átomos en un diccionario dada la clase y el valor de su correspondiente coeficiente de activación, sino que también incorpora el efecto que éste tiene sobre el error total de representación. En esta tesis se presentó además un nuevo método iterativo llamado DAS-KSVD para el aprendizaje de diccionarios estructurados en el contexto de problemas de clasificación multi-clase, que utiliza ésta medida. El nuevo método permite detectar los átomos más discriminativos para cada una de las clases. Utilizando una base de datos de dígitos manuscritos ampliamente utilizada en la literatura, se realizó un análisis del desempeño del método DAS-KSVD obteniéndose tasas de reconocimiento superiores a las obtenidas por técnicas semejantes del estado del arte.

También se utilizó el nuevo método DAS-KSVD en un problema de clasificación multi-clase asociado al SAHOS (clasificación de eventos en normal, de apnea y de hipopnea). Los resultados obtenidos muestran que éste método tiene un muy buen desempeño en la detección de la patología.

5.2. Artículos científicos

Los resultados de investigación logrados durante la realización de esta tesis han sido publicados, en un principio, en congresos nacionales e internacionales y, posteriormente, en revistas internacionales indexadas de amplio reconocimiento en el área. A continuación se enumeran tanto los artículos científicos publicados (y enviados para su publicación) en revistas internacionales como los trabajos publicados en actas de congresos.

En revistas internacionales:

1. R.E. Rolón, L.D. Larrateguy, L.E. Di Persia, R.D. Spies and H.L. Rufiner, *Dis-*

criminative methods based on sparse representations of pulse oximetry signals for sleep apnea hypopnea detection, Biomedical Signal Processing and Control, Elsevier, Vol. 33, pp. 358-367, 2017, DOI: `http://dx.doi.org/10.1016/j.bspc.2016.12.013`. (Anexo B)

2. R.E. Rolón, I.E. Gareis, L.E. Di Persia, R.D. Spies and H.L. Rufiner, *Complexity-based discrepancy measures applied to classification of apnea-hypopnea events*, Complexity, Special Issue: Measuring Complexity of Biomedical Signals, Wiley-Hindawi, pp. 1-18, 2018, DOI: `https://doi.org/10.1155/2018/1435203`. (Anexo A)

3. R.E. Rolón, L.E. Di Persia, R.D. Spies and H.L. Rufiner, *A multi-class structured dictionary learning method using discriminant atom selection*, Pattern Analysis and Applications, Springer, pp. 1-18, enviado, 2018. (Anexo D)

En actas de congresos:

1. R.E. Rolón, L.E. Di Persia, H.L. Rufiner and R.D. Spies, *Most discriminative atom selection for apnea-hypopnea events detection*, VI Latin American Congress on Biomedical Engineering, Online ISBN 978-3-319-13117-7, Vol. 4, pp. 572-575, 2014.

2. L.D. Larrateguy, R.E. Rolón, L.E. Di Persia, R.D. Spies and H.L. Rufiner, *Método de screening para la detección de SAHOS utilizando selección de funciones discriminativas*, 42^o Congreso Argentino de Medicina Respiratoria, Publicado por Revista Americana de Medicina Respiratoria, ISSN 1852-1630, pp. 99, 2014.

3. R.E. Rolón, L.E. Di Persia, H.L. Rufiner and R.D. Spies, *A method for atoms selection applied to screening for sleep disorders*, 1st Pan-American Congress on Computational Mechanics, ISBN 978-84-943928-2-5, pp 1155-1166, 2015.

4. R.E. Rolón, L.E. Di Persia, H.L. Rufiner and R.D. Spies, *Two alternatives for atoms selection applied to screening for sleep disorders*, V Congress on Industrial, Computational and Applied Mathematics, ISSN 2314-3282, Vol. 5, pp 437-440, 2015.

5. R.E. Rolón, L.E. Di Persia, R.D. Spies and H.L. Rufiner, *A method for discriminative dictionary learning with application to pattern recognition*, VI Congress on Industrial, Computational and Applied Mathematics, 2017. (Anexo C)

5.3. Trabajos futuros

Los trabajos futuros estarán dirigidos principalmente a mejorar y extender los resultados alcanzados con la segunda propuesta, es decir, con el método de aprendizaje de diccionarios estructurados. En lo referente al algoritmo, se pretende desarrollar una nueva estrategia que permita encontrar el valor del par óptimo de parámetros (α^*, β^*) en la medida de discriminabilidad combinada. Asimismo, se trabajará en el diseño de un nuevo procedimiento para el remuestreo de las señales con el fin de aumentar la capacidad de generalización del diccionario estructurado, incorporar información temporal de la historia de los eventos de apnea y de hipopnea y evaluar otros tipos de clasificadores, entre otras actividades.

Por otro lado, se explorará el uso de diccionarios "convolutivos" unidimensionales (1D) que sean capaces de encontrar representaciones invariantes ante desplazamientos temporales [94]. Se espera que este tipo de diccionarios agregue información relevante respecto a la dinámica temporal de las señales de SaO_2. Asimismo, se evaluará el uso de técnicas de aprendizaje de diccionarios profundos [40]. Estas novedosas técnicas permiten lograr representaciones que tienen una estructura jerárquica (por nivel) de los átomos del diccionario. De esta manera, se puede realizar un aprendizaje jerárquico de las características más importantes de las señales de SaO_2, como por ejemplo la respiración normal (primer nivel), los eventos de apnea y de hipopnea (segundo nivel), patología severa (tercer nivel), etc.

Puesto que el futuro de las técnicas de diagnóstico simplificado del SAHOS está fuertemente orientado al desarrollo de nuevas tecnologías de "no contacto" (o contacto mínimo) que utilicen un número reducido de señales fisiológicas (obtenidas mediante técnicas no invasivas), se buscará incorporar la señal de sonido traqueal con el fin de mejorar el desempeño de los clasificadores. En este contexto, se estudiará el uso de nuevas técnicas de aprendizaje de diccionarios acoplados capaces de incorporar exitosamente la información adicional derivada de los cambios tanto en

la dinámica temporal de ambas señales como en las secuencias de eventos de apnea y de hipopnea [95].

Anexos

Los Anexos A, B, C y D que se incluyen a continuación refieren a los siguientes artículos, que son producto de esta tesis:

- R.E. Rolón, I.E. Gareis, L.E. Di Persia, R.D. Spies and H.L. Rufiner, *Complexity-based discrepancy measures applied to classification of apnea-hypopnea events*, Complexity, Special Issue: Measuring Complexity of Biomedical Signals, Wiley-Hindawi, pp. 1-18, 2018, DOI: `https://doi.org/10.1155/2018/1435203`.
- R.E. Rolón, L.D. Larrateguy, L.E. Di Persia, R.D. Spies and H.L. Rufiner, *Discriminative methods based on sparse representations of pulse oximetry signals for sleep apnea hypopnea detection*, Biomedical Signal Processing and Control, Elsevier, Vol. 33, pp. 358-367, 2017, DOI: `http://dx.doi.org/10.1016/j.bspc.2016.12.013`.
- R.E. Rolón, L.E. Di Persia, R.D. Spies and H.L. Rufiner, *A method for discriminative dictionary learning with application to pattern recognition*, VI Congress on Industrial, Computational and Applied Math., 2017.
- R.E. Rolón, L.E. Di Persia, R.D. Spies and H.L. Rufiner, *A multi-class structured dictionary learning method using discriminant atom selection*, Pattern Analysis and Applications, Springer, pp. 1-18, enviado, 2018.

El postulante declara haber tenido un rol protagónico consistente principalmente en el análisis, diseño, implementación y evaluación de los métodos presentados y en los experimentos realizados para la obtención de los resultados que allí se muestran. Estas tareas fueron realizadas bajo el seguimiento y supervisión del director y del co-director de tesis, Dr. H. L. Rufiner y Dr. R. D. Spies, respectivamente. En cuanto a la escritura de los artículos, el tesista ha sido el autor principal, guiado por los comentarios, sugerencias y revisiones del director y co-director de tesis y los demás co-autores que colaboraron en cada publicación. Los abajo firmantes avalan esta declaración.

Dr. Hugo L. Rufiner	Dr. Rubén D. Spies
Director	Co-director

Anexo A

Complexity-based discrepancy measures applied to classification of apnea-hypopnea events

Artículo publicado en la revista *Complexity*, Special Issue: Measuring Complexity of Biomedical Signals, Wiley-Hindawi, pp. 1-18, 2018, DOI: DOI: https://doi.org/10.1155/2018/1435203.

Complexity-based discrepancy measures applied to classification of apnea-hypopnea events

R.E. Rolón[a], I.E. Gareis[a,c], L.E. Di Persia[a], R.D. Spies[b], H.L. Rufiner[a,c]

May 2, 2018

[a] Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, sinc(i), FICH−UNL/CONICET, Santa Fe, Argentina
[b] Instituto de Matemática Aplicada del Litoral, IMAL, FIQ−UNL/CONICET, Santa Fe, Argentina
[c] Laboratorio de Cibernética, Fac. de Ing., Univ. Nacional de Entre Ríos, Argentina

Abstract

In recent years an increasing interest in the development of discriminative methods based on sparse representations with discrete dictionaries for signal classification has been observed. It is still unclear, however, what is the most appropriate way for introducing discriminative information into the sparse representation problem. It is also unknown which is the best discrepancy measure for classification purposes. In the context of feature selection problems, several complexity-based measures have been proposed. The main objective of this work is to explore a method that uses such measures for constructing discriminative sub-dictionaries for detecting apnea-hypopnea events using pulse oximetry signals. Besides traditional discrepancy measures, we study a simple one called difference of conditional activation frequency (DCAF). We additionally explore the combined effect of over-completeness and redundancy of the dictionary as well as the sparsity level of the representation. Results show that complexity-based measures are capable of adequately pointing out discriminative atoms. Particularly DCAF yields competitive averaged detection accuracy rates of 72.57% at low computational cost. Additionally, ROC curve analyses show averaged diagnostic sensitivity and specificity of 81.88% and 87.32%, respectively. This shows that discriminative sub-dictionary construction methods for sparse representations of pulse oximetry signals constitute a valuable tool for apnea-hypopnea screening.

Keywords: Discriminative information, discrepancy measures, sparse representation, apnea-hypopnea events, pulse oximetry signal.

1 Introduction

Although it is widely used and accepted, the notion of complexity has very often avoided a rigorous formalization. It is therefore not surprising that no universally accepted measure exists yet for quantifying such a concept. In particular, within information theory, the complexity of any element of a code, or of any feature of a signal representation in the context of signal processing, is known to be strongly related to the information it carries or, more precisely, to the value of its entropy. It is important to point out however that, in the context of signal classification, the more informative features (in terms of classification) are not necessarily the ones with larger entropy. Hence more "ad-hoc" measures are needed. In fact, any appropriate complexity measure corresponding to a given feature should be instead, strongly related to the amount of information about class membership provided by such a feature. One could then think of using as measure of complexity the conditional entropy of the class given the feature. However, features providing the most discriminative information regarding a class are almost always those with lower conditional entropy values, and hence the best features for classification purposes will be the least complex ones.

Information theory was originally based on the engineering of noisy communication channels, and it is closely associated to a large number of disciplines such as signal processing, artificial intelligence, complex systems and pattern recognition, to name only a few. We are particularly interested in the latter. Pattern recognition is a discipline which is mainly oriented to the generation of algorithms or methods that can decide an action based upon certain recognized similarities (patterns) in the input data. Within signal classification, which is perhaps one of the most important subfields of pattern recognition, several discrepancy measures have been used in problems coming from a wide variety of areas such as machine

learning [1], image and speech processing [2], neural networks [3] and biomedical signal processing [4, 5], among others. Among them the most commonly used is probably the Kullback-Leibler (KL) divergence [6, 7]. This divergence, also known as relative entropy, was used as a discriminative measure for selecting, from a large collection of orthonormal bases, the one attaining maximum information [1]. A more recent approach was introduced by Gupta *et.al.* [8] who used this divergence as a discrepancy measure in the traditional k-nearest neighbor (k-NN) algorithm, yielding competitive classification performances in the context of raw electroencephalographic signal classification. Although it provides certain computational and theoretical advantages, the lack of symmetry of the KL divergence has motivated the development of several symmetric versions such as the so called J-divergence [9] and the well known and widely used Jensen-Shannon divergence [10].

Sparse representation of signals constitute a useful technique which has drawn wide interest in recent years due to its success in many applications such as signal and image processing [11]. This technique allows the analysis of the signals by means of only a few well-defined basic waveforms. Due to its advantages, such as robustness to noise and dimension reduction, among others, sparse representation has acquired a large popularity in the area of biomedical signal processing. For example, this technique has been successfully applied to several problems including the estimation of the human respiratory rate [12] and electrocardiographic signal processing, both for signal enhancement and QRS complex detection, for improving heart disease analysis and diagnosis [13]. It is timely to point out however that, up to our knowledge, no applications of discrepancy measures to sparse representation for signals classification are known yet.

All reconstructive methods, such as principal components analysis (PCA), independent components analysis (ICA) and the previously mentioned sparse representations [14], produce particular types of signal representations minimizing a given cost functional which usually involves both fidelity and regularization terms. These methods have been successfully applied in a wide variety of problems such as signal denoising, missing data and outliers, among many others. On the other hand, discriminative methods such as linear discriminant analysis (LDA) are oriented to find optimal decision boundaries to be used for classification tasks. It is well known that for signal classification, which is our main interest in this work, discriminative methods generally outperform reconstructive methods. It is mainly for this reason that several authors have recently developed supervised approaches based on sparse representation which are simultaneously reconstructive and discriminative [15, 16].

The obstructive sleep apnea-hypopnea (OSAH) syndrome [17] is one of the most common sleep disorders and more often that not it remains undiagnosed and therefore not treated. This syndrome is caused by repeated events of partial or total blockage of the upper airway during sleeping, which correspond to events of hypopnea and apnea, respectively. To evaluate the severity degree of the OSAH syndrome, medical physicians have created the so called apnea-hypopnea index (AHI), which is defined as the average number of apnea-hypopnea events per hour of sleep. In terms of this index OSAH is classified as normal, mild, moderate or severe depending on whether such an index falls in the interval $[0, 5)$, $[5, 15)$, $[15, 30)$, or $[30, \infty)$, respectively. The gold standard test for OSAH diagnosis is a study called polysomnography (PSG). However, PSG is both costly and lengthy and the accessibility to this type of study is limited. Additionally, PSG studies require of information coming from a variety of physiological signals such as electroencephalography (EEG), airflow, pulse oximetry (SaO_2), etcetera. It is known however that cessation of breathing associated with apnea-hypopnea events are always accompanied by a drop in the oxygen saturation level in the SaO_2 signal record, although quite often such a drop is very small and almost impossible to detect by a human observer.

The main objective of this work is precisely to develop a technique based on sparse representations and the use of appropriate discriminative information that be able to accurately and efficiently detect apnea-hypopnea events by using only the SaO_2 signal. Several ways exists for combining discriminative information and sparse representations within the context of signal classification. We shall follow one consisting of using the discriminative information for detecting those atoms having the most frequent activations in order to provide them as input for a classifier. This approach was initially introduced in [4] where two methods using the absolute value of the activation differences of the atoms as a measure of the discriminative information for the detection of OSAH were presented. In this work a rigorous formalization of such a measure is introduced and compared with several other discrepancy measures for classifying apnea-hypopnea events. Also, the combined effect of using different sizes of non-redundant dictionaries and different sparsity degree is explored in detail. Results show clearly that the proposed measure is capable of adequately pointing out discriminative atoms in a full dictionary, yielding competitive accuracy rates in the detection of individual apnea-hypopnea events. Additionally, this new approach is computationally very cheap. In fact, it has proved to be at least twice faster than those associated to all other discrepancy measures.

The rest of this article is organized as follows: in Section 2 the obstructive sleep apnea-hypopnea syndrome is explained. Sparse representation of signals is introduced in Section 3. In Section 4 several discriminative information measures are presented. Section 6 contains a detailed description about the performed experiments. Results and discussions are introduced in Section 7 while conclusions are presented in Section 8.

2 Sleep apnea-hypopnea

Apnea-hypopnea events occur as a consequence of a functional-anatomic disturbance of the upper airway producing its partial or total blockage. At the end of an apnea-hypopnea event, a pronounced desaturation of the blood hemoglobin commonly occurs. These desaturations generate characteristic patterns in the pulse oximetry record known as intermittent hypoxemias. The hypoxemia-reoxygenation cycles promote oxidative stress, angiogenesis and tumor growth, favor the sympathetic activation with increment of blood pressure and systemic and vascular inflammation with endothelial dysfunction which contributes to multi-organic chronic morbility, metabolic abnormalities and cognitive impairment [18]. Additionally, strong correlations between neoplastic diseases and the OSAH syndrome have been described in [19]. Also, a recent study among male mice suggests that OSAH's intermittent hypoxia can be associated to fertility reduction [20]. Currently this pathology affects more than 4% of the human population around the world [21]. Additionally, it was found that aging, male gender, snoring and obesity are all risk factors for OSAH syndrome [22].

Although very limited in many countries, overnight polysomnography (PSG) is currently the gold standard tool for diagnosing OSAH syndrome. As previously mentioned, a full PSG consists of the simultaneous measurement of several physiological signals such as EEG, electrocardiography (ECG), respiratory effort, airflow, SaO_2 and electrical activity produced by skeletal muscles (EMG), etc. Mainly due to its ease of acquisition, we are particularly interested in the SaO_2 signal. Figure 1 shows a typical temporal plot of just a few physiological signals coming from a full PSG. This figure also depicts a portion of an original raw airflow signal as well as the corresponding portion of the SaO_2 signal. The corresponding labels of apnea-hypopnea events (dashed lines) are also shown. Finally, at the bottom of this figure, the electrical activity of the heart as well as the sleep stages are shown. In a typical PSG study, after a normal period of sleep the recorded signals are provided to medical experts who analyze the whole record and mark the apnea-hypopnea events and sleep stages, needed for the posterior evaluation the AHI index. Due to its complexity and cost, a few alternatives to PSG have been adopted. One of the most popular ones is the so called home respiratory polygraphy (HRP) [23] which requires no neurophysiological signals. Although studies have shown that there exists a high correlation between AHI values generated by HRP and PSG studies [24], HRP still needs of several physiological signals, whose acquisition strongly affects the normal sleeping of the person. It is therefore highly desirable to develop a reliable OSAH screening system which makes use of as few as possible physiological signals. In this regard, pulse oximetry, being a cheap and non-invasive technique, has become a suitable alternative for screening purposes [25].

In this work we shall develop a method for the detection of apnea-hypopnea events that uses only the SaO_2 signals. Our approach leads to a binary classification problem whose main purpose is the detection of the presence (or not) of events of apnea and hypopnea. It is timely to point out that although our method does take into consideration an appropriate fidelity term, we are by no means interested in achieving accurate signal representation.

3 Sparse representations

As previously mentioned, one of the most popular reconstructive methods is based on sparse representations of the signals involved. Sparsity can be enforced by including upper bounds for the number of non-zero coefficients in the representation of the given signals in terms of atoms in a dictionary.

Formally, the problem of sparse representations of signals can be separated into two sub-problems, the so-called sparse coding problem and the dictionary learning problem. We shall now proceed to describe in detail each one of these sub-problems. To be more precise, let $\mathbf{x} \in \mathbb{R}^N$ be a discrete signal and let $\Phi \in \mathbb{R}^{N \times M}$ (generally with $M \geq N$) be a dictionary whose columns $\boldsymbol{\phi}_j \in \mathbb{R}^N$ are atoms that we want to use for obtaining a representation of $\mathbf{x}$ of the form $\mathbf{x} = \Phi\mathbf{a}$. Here, and in the sequel, we shall refer to the vector $\mathbf{a} = [a_1 \ a_2 \ \cdots \ a_M]^T \in \mathbb{R}^M$ as a "representation" of $\mathbf{x}$. Sparsity consists essentially of obtaining a representation with as few non-zero elements as possible. A way of obtaining such a representation

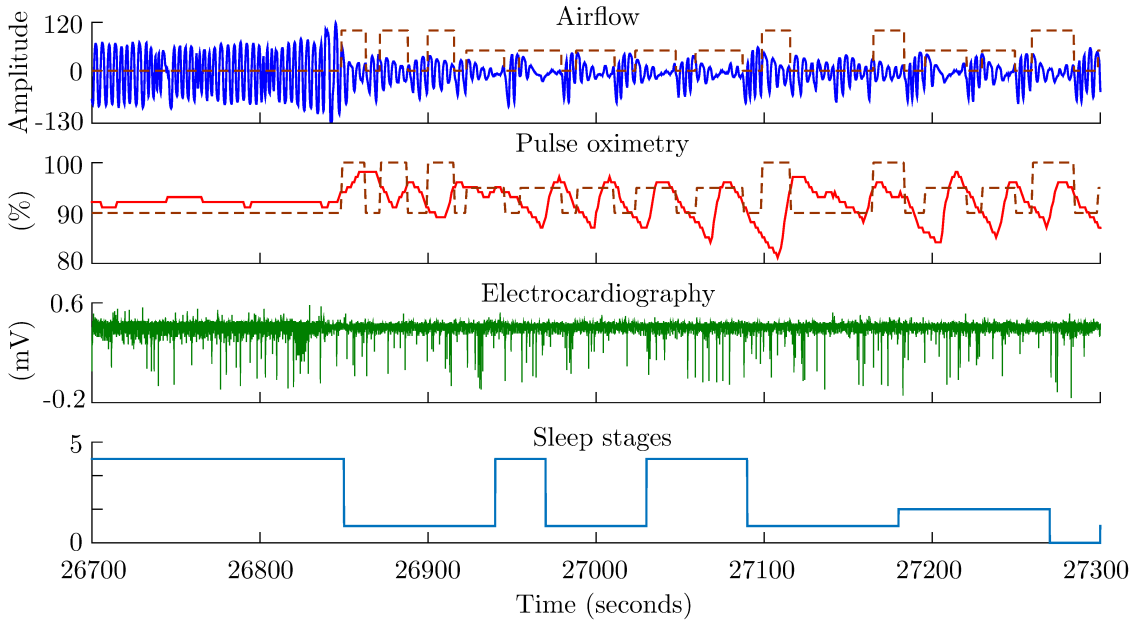


Figure 1: A portion of a few number of physiological signals coming from a full PSG. Dashed lines (brown) are apnea-hypopnea labels introduced by the medical expert.

consists of solving the following problem:

$$(P_0): \quad \min_{\mathbf{a}} ||\mathbf{a}||_0 \text{ subject to } \mathbf{x} = \Phi\mathbf{a},$$

where $||\mathbf{a}||_0$ denotes the l_0 pseudo-norm, defined as the number of non-zero elements of $\mathbf{a}$.

Several questions regarding problem (P_0) immediately arise. Among them: *i*) does there exist an exact representation $\mathbf{x} = \Phi\mathbf{a}$?, *ii*) if an exact representation exists, is it unique?, *iii*) in the case of non-uniqueness, how do we find the "sparsest" representation?, *iv*) how difficult is it, from the computational point of view, to solve problem (P_0)?. Although it is not an objective of this article to get into details about the answers to these questions, it turns out that imposing exact representation is most often a too restrictive and therefore inappropriate constrain and, on the other hand, solving (P_0) is generally an NP hard problem yielding this approach highly unsuitable for most applications. For more details we refer the reader to [26, §1.8].

In order to overcome some of the difficulties which entail solving problem (P_0), several relaxed versions of it have been considered. One of them consists of allowing a small representation error while imposing an upper bound on the l_0 pseudo-norm of the representation:

$$(P_0^q): \quad \min_{\mathbf{a}} ||\mathbf{x} - \Phi\mathbf{a}||_2 \text{ subject to } ||\mathbf{a}||_0 \leq q,$$

where q is a prescribed integer parameter. This formulation takes into account the existence of possible additive noise terms; in other words it assumes that $\mathbf{x} = \Phi\mathbf{a} + \mathbf{e}$ where $\mathbf{e} \in \mathbb{R}^N$ is a small energy noise term. Thus, this approach is particularly suitable in most real applications (such as biomedical signal processing) where measured signals are always contaminated by noise. Several greedy strategies have been proposed for solving problem (P_0^q) [27, 28]. Among them, orthogonal matching pursuit (OMP) [28] is perhaps the most commonly used strategy. This greedy algorithm guarantees convergence to the projection of $\mathbf{x}$ into the span of the dictionary atoms, in no more than q iterations. Figure 2 shows an example of the values of a particular coefficient a_{j^*} associated to the atom ϕ_{j^*} obtained by applying the OMP algorithm for a large number (almost half a million) of segments of SaO_2 signals and its corresponding activation histogram.

Although pre-constructed dictionaries, such as the well known wavelet packets [29], typically lead to fast sparse coding, they are almost always restricted to certain classes of signals. Is is mainly for this reason that new approaches introducing data-driven dictionary learning techniques emerged. A dictionary learning problem associated to the data: q, M, $N \in \mathbb{N}$, $M \geq N$ and n signals in $\mathbb{R}^N$, $\mathbf{x}_1, \cdots, \mathbf{x}_n$, can be formally written as:

$$(DL): \quad \min_{\substack{\Phi \in \mathbb{R}^{N \times M} \\ \mathbf{a}_i \in \mathbb{R}^N, ||\mathbf{a}_i||_0 \leq q, 1 \leq i \leq n.}} \sum_{i=1}^{n} ||\mathbf{x}_i - \Phi\mathbf{a}_i||_2$$

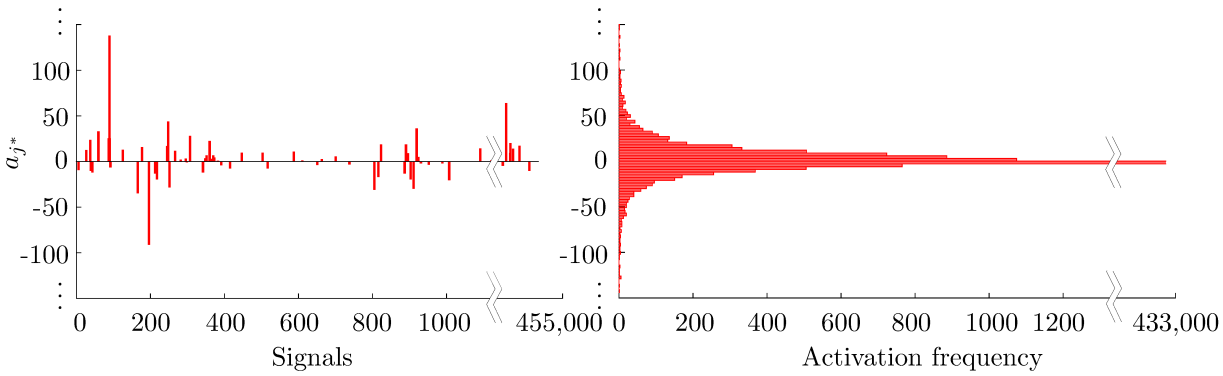


Figure 2: The values of the activations of a particular atom for each signal (left) and the corresponding histogram of activations (right).

This problem consists of simultaneously finding a dictionary Φ and representations of the n signals $\mathbf{x}_i$, $1 \leq i \leq n$, (in terms of atoms of such a dictionary) complying with the sparsity constraint for each one of the n signals, while minimizing the total representation error.

The first data-based dictionary learning algorithms were originally developed almost three decades ago [30, 31, 32]. Some of them have their roots in probabilistic frameworks by considering the observed data as realizations of certain random variables [30, 31]. In [31] for example, the authors developed an algorithm for finding a redundant dictionary that maximizes the likelihood function of the probability distribution of the data. In that work, an analytic expression for the likelihood function was derived by approximating the posterior distribution by Gaussian functions. An iterative approach for dictionary learning, known as the "method for optimal directions" (MOD), was presented in [32]. The sparse coding stage of this method makes use of the OMP algorithm followed by a simple dictionary updating rule. A new iterative algorithm was recently proposed by Aharon *et.al.* in [14]. This new approach, called "K singular value decompositions" (KSVD), consists mainly of two stages: a sparse coding stage and a dictionary learning stage. The OMP algorithm is used for the sparse coding stage, which is followed by a dictionary updating step where the atoms are updated one at a time and the representation coefficients are allowed to change in order to minimize the total representation error.

4 Discriminative sub-dictionary construction

Although data-driven dictionary learning algorithms produce sparse representations of signals which are robust against noise and missing data, such representations turn out to be unsuitable if the final objective is signal classification. This is mainly so because those algorithms do not take into account any a-priori or available information concerning class membership. In order to overcome this difficulty, some strategies which incorporate appropriate class information have been proposed [4, 33, 16]. In [33], for instance, the authors developed a discriminative dictionary learning method by efficiently integrating a single predictive linear classifier into the cost function of the KSVD algorithm. A method incorporating a discriminative term into the cost function of the standard KSVD algorithm was presented in [16]. This method finds an optimal dictionary which is simultaneously representative and discriminative for face recognition tasks. In this work, we make use of a simple approach for detecting discriminative atoms from a previously learned dictionary and using them to build a new sub-dictionary. This approach, which was originally presented in [4], consists of solving two problems, namely: *i*) the above mentioned full (DL) problem and *ii*) a discriminative sub-dictionary (DSD) construction problem. We shall now proceed to describe problem *ii*). One way to obtain discriminative sub-dictionaries consists of maximizing an appropriate discriminative value functional $G(\cdot)$. Given a data matrix $\mathbf{X} \in \mathbb{R}^{N \times n}$, a class label vector $\mathbf{c} \in \mathcal{C}^n$ (where $\mathcal{C}$ is the set of all classes; in the binary case $\mathcal{C} = \{c_1, c_2\}$), a dictionary $\Phi \in \mathbb{R}^{N \times M}$ and $p \in \mathbb{N}$ (with $p < M$), the most discriminative sub-dictionary $\hat{\Phi}^{\mathbf{d}} \in \mathbb{R}^{N \times p}$, according to an appropriate prescribed discriminative value functional $G_{\mathbf{X},\mathbf{c},\Phi} : \mathbb{R}^{N \times p} \to \mathbb{R}_0^+$ is defined as:

$$(DSD): \quad \hat{\Phi}^{\mathbf{d}} = \underset{\substack{\mathbf{d} \doteq [i_1\, i_2 \cdots i_p] \\ i_j \in \{1,2,\ldots,M\} \\ i_j \neq i_k \forall j \neq k}}{\operatorname{argmax}} \quad G_{\mathbf{X},\mathbf{c},\Phi}\left(\Phi^{\mathbf{d}}\right),$$

where for $\mathbf{d} \doteq [i_1\, i_2 \cdots i_p]$, $\Phi^{\mathbf{d}}$ denotes the $N \times p$ matrix whose j^{th}-column is the i_j^{th}-column of Φ. The function G, which must be provided, quantifies the discriminative power of each sub-dictionary $\Phi^{\mathbf{d}}$. Thus, large values of G correspond to highly discriminative sub-dictionaries while small values of G are

associated to sub-dictionaries with low discriminability.

Several questions concerning problem (DSD) clearly emerge. Among them: i) how do we find an appropriate discriminative value function G?, ii) given the functional G, does problem (DSD) have a solution?, iii) if it does, is it unique?, iv) in the case of non-uniqueness, how do we decide which sub-dictionary, among the optimizers, is the best for our classification purposes?, v) how difficult is it, in terms of computational cost, to solve problem (DSD)?. Although this problem has not been extensively studied, is it known that solving (DSD) is computationally very challenging for $p > 1$, mainly due to the combinatorial explosion problem. A way to overcome the computational complexities entailed by problem (DSD) consists of defining an appropriate discriminative value functional G for $p = 1$. In that way G is independently evaluated at each one of the atoms (columns) of Φ and the discriminative sub-dictionary $\Phi^{\mathbf{d}} \in \mathbb{R}^{N\times p^*}$ is constructed by stacking side-by-side the first p^* ranked columns of Φ with largest G values. This simplification is based on the assumption that each atom in the dictionary is used to model specific characteristics that are not completely modeled by the other atoms. Thus, the discriminative information provided by a particular atom will be different from the information contributed by other atoms.

5 Discriminative value functions for atom selection

Several ways for appropriately constructing discriminative value functions G exists. In this section we present two different approaches to define such a function. Namely i) using traditional discrepancy measures and ii) using a new discriminative measure to which we shall refer as the "difference of conditional activation frequency" (DCAF). We shall previously need to introduce an appropriate setting and terminology regarding probability density functions (PDFs) in the context of sparse representations for signal classification.

Here, and in the sequel, we shall consider the vectors $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ as realizations of a particular random vector $\mathcal{X}$. Any sparse representation of those vectors will result in the PDFs of each coefficient a_j (associated to the atom ϕ_j) showing a very concentrated peak at zero with heavy tails (as depicted in Figure 2). In the context of binary signal classification it is reasonable to think that if a given atom ϕ_{j^*} is highly discriminative, then the conditional PDFs $\pi(a_{j^*}|c_1)$ and $\pi(a_{j^*}|c_2)$ will be significantly different. Thus, if a dictionary Φ is poorly discriminative, then one should expect $\pi(a_j|c_1) \approx \pi(a_j|c_2)$ for all j.

Although the elements a_{j^*} of the representation vector $\mathbf{a}$ are in general real numbers, for practical reasons it is appropriate to discretize them. That can be done in the usual way by partitioning the real line $\mathbb{R}$ into intervals $I_k \doteq \left(\left(k-\frac{1}{2}\right)\Delta, \left(k+\frac{1}{2}\right)\Delta\right]$, $k \in \mathbb{Z}$, of length Δ and the associated discretized random variable $\mathcal{K}_j \doteq \sum_{k\in\mathbb{Z}} k\, \chi_{I_k}(a_j)$. The corresponding probability mass function (PMF) is $p_{\mathcal{K}_j}(k) = P(a_j \in I_k) = \int_{I_k} \pi(a_j)\, da_j$, $k \in \mathbb{Z}$. Figure 3 shows the estimated PMF and the corresponding conditional PMFs (given each one of the two classes), both for a non-discriminative and a discriminative atom using SaO_2 signals.

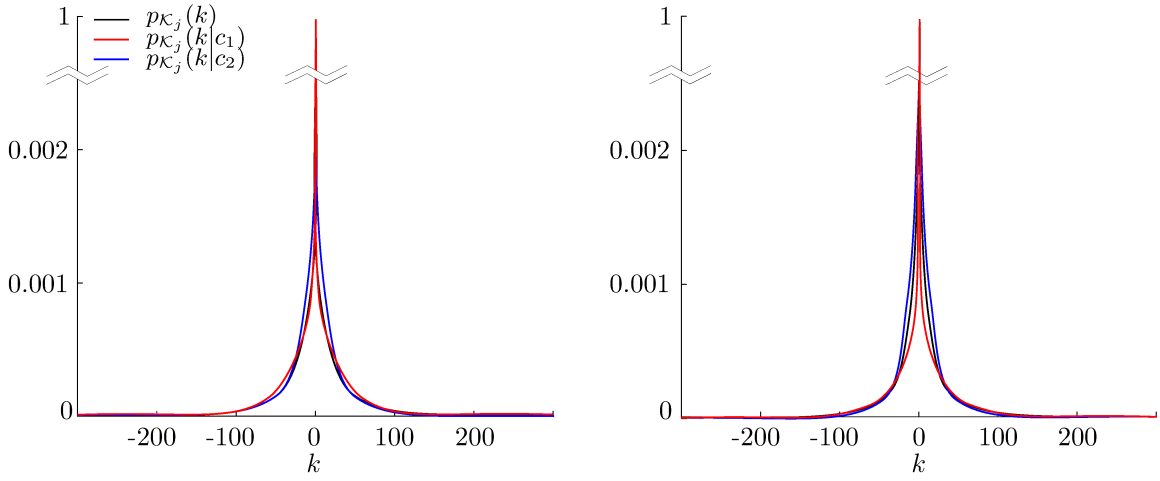


Figure 3: Estimated probability mass functions for a non-discriminative atom ϕ_j (left) and a discriminative one (right).

We shall now proceed to define how we compute the discriminative value function G. Given the data matrix $\mathbf{X} \in \mathbb{R}^{N\times n}$, the corresponding class label vector $\mathbf{c} \in \mathcal{C}^n$ and a full dictionary $\Phi \in \mathbb{R}^{N\times M}$, the first step consists of obtaining the sparse matrix $\mathbf{A} \doteq [\mathbf{a}_1\ \mathbf{a}_2\ \cdots\ \mathbf{a}_n] \in \mathbb{R}^{M\times n}$ by applying the OMP algorithm.

The j^{th}-row of this sparse matrix is then used for estimating the conditional PMFs $p_{\mathcal{K}_j}(\cdot|c_1)$ and $p_{\mathcal{K}_j}(\cdot|c_2)$. Finally, the value of G at the atom ϕ_j is computed as the discrepancy (as quantified by an appropriate discrepancy measure) between these two PMFs. In what follows, we introduce the discrepancy measures that we shall use in this work.

5.1 Traditional discrepancy measures

A great diversity of measures whose purpose is performing comparisons between probability distributions exists [34]. In this work the best known and more commonly used ones are compared in terms of their performance for selecting the most discriminative atoms in a dictionary. The KL, J and JS divergence measures were utilized, along with the Fisher score (F).

The KL divergence [7] is probably the most widely used information "distance" measure from a theoretical framework and it was successfully applied in numerous problems for signal classification [1, 35, 36]. To compare the two conditional PMFs associated with the activation of the j^{th}-atom the KL distance was used as follows:

$$\mathrm{KL}\left(p_{\mathcal{K}_j}(\cdot|c_1), p_{\mathcal{K}_j}(\cdot|c_2)\right) \doteq \sum_{k\in\mathbb{Z}} p_{\mathcal{K}_j}(k|c_1) \log\left(\frac{p_{\mathcal{K}_j}(k|c_1)}{p_{\mathcal{K}_j}(k|c_2)}\right), \tag{1}$$

assuming that $0\log(0) \doteq 0$.

Despite the computational and theoretical properties provided by KL distance, what usually becomes a trouble in many problems of signal classification is its lack of symmetry. It can be easily seen that altering the order of the arguments in (1) can change the output value. To solve this issue a symmetric version of the KL distance can be used such as the J-divergence [9], which, even though was not initially created as a symmetric version of the KL distance, is the sum of the two possible KL distances between probability distributions. In this article the J-divergence is defined as follows:

$$\mathrm{J}\left(p_{\mathcal{K}_j}(\cdot|c_1), p_{\mathcal{K}_j}(\cdot|c_2)\right) \doteq \mathrm{KL}\left(p_{\mathcal{K}_j}(\cdot|c_1), p_{\mathcal{K}_j}(\cdot|c_2)\right) + \mathrm{KL}\left(p_{\mathcal{K}_j}(\cdot|c_2), p_{\mathcal{K}_j}(\cdot|c_1)\right). \tag{2}$$

Another symmetric smoothed version of the KL distance is the JS divergence [10]. For the problem of comparing the two conditional probabilities associated to each class it is defined as:

$$\mathrm{JS}\left(p_{\mathcal{K}_j}(\cdot|c_1), p_{\mathcal{K}_j}(\cdot|c_2)\right) \doteq w_1\mathrm{KL}\left(p_{\mathcal{K}_j}(\cdot|c_1), q_{\mathcal{K}_j}(\cdot)\right) + w_2\mathrm{KL}\left(p_{\mathcal{K}_j}(\cdot|c_2), q_{\mathcal{K}_j}(\cdot)\right), \tag{3}$$

where $q_{\mathcal{K}_j}(\cdot) = w_1 p_{\mathcal{K}_j}(\cdot|c_1) + w_2 p_{\mathcal{K}_j}(\cdot|c_2)$ and w_1 and w_2 are the weights associated to each of the conditional PMFs, with $w_1, w_2 \geq 0$ and $w_1 + w_2 = 1$. An interesting feature of the JS-distance is the fact that different values of weights (w_1 and w_2) can be assigned to the probability distributions according to their importance. In this work $w_1 = P(c_1)$ and $w_2 = P(c_2)$ i.e. the weights are associated with the a-priori probabilities of the classes. Note that computing the JS-distance as defined here is the same as computing the mutual information between the class and the activations, i.e. $\mathrm{JS}\left(p_{\mathcal{K}_j}(\cdot|c_1), p_{\mathcal{K}_j}(\cdot|c_2)\right) = \mathrm{MI}(\mathcal{K}_j, \mathcal{C})$.

Within signal classification problems, F is a measure which has been extensively used. Unlike the other measures presented here, that require estimations of the conditional PMFs, F uses just two parameters of the distributions (the means and standard deviations). This makes this measure much less expensive computationally speaking, but implicitly assumes certain characteristics of the distribution under study (i.e. second order characteristics). In the case of univariate binary problem at hand the F can be defined as:

$$\mathrm{F}(p_{\mathcal{K}_j}(\cdot|c_1), p_{\mathcal{K}_j}(\cdot|c_2)) \doteq \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}, \tag{4}$$

where μ_j and σ_j^2 are the mean and standard deviation of $p_{\mathcal{K}_j}(\cdot|c_j)$ [37].

Although the above mentioned discrepancy measures provide, in a certain sense, "measures" of distance between two probability distribution functions, most of them (such as the KL divergence and those symmetric variants) are not strictly a metric. For instance, the KL divergence is a non-symmetric discrepancy measure where the triangular inequality is not satisfied. Nevertheless, $\mathrm{KL}(p_{\mathcal{K}_j}(\cdot|c_1), p_{\mathcal{K}_j}(\cdot|c_2))$ is a non-negative measure, i.e. $\mathrm{KL}(p_{\mathcal{K}_j}(\cdot|c_1), p_{\mathcal{K}_j}(\cdot|c_2)) \geq 0$ and $\mathrm{KL}(p_{\mathcal{K}_j}(\cdot|c_1), p_{\mathcal{K}_j}(\cdot|c_2)) = 0$ if and only if $p_{\mathcal{K}_j}(\cdot|c_1) = p_{\mathcal{K}_j}(\cdot|c_2)$.

5.2 Difference of conditional activation frequency

In a previous work a method called most discriminative column selection (MDCS) for the construction of a discriminative sub-dictionary was originally presented [4]. The sparse representations of the signals in

terms of sub-dictionaries constructed using MDCS provided good performance in the detection of apnea-hypopnea events. In the mentioned work, the most discriminative atoms were identified by comparing the difference of conditional activation frequency (DCAF).

The candidates to be considered as "most discriminative" according to [4] are those atoms with higher absolute difference between conditional activation probabilities given the class. That is, an atom is considered as highly discriminative if it is active, in proportion, more times for one of the classes. The use of this approach as a measure of discriminative power follows from the idea that one of the most expressive parameters regarding the importance of a given atom is its activation probability. Moreover, if certain atoms are active mostly for a given class, then it is assumed they represent features of importance in the description of that particular class.

Following this idea, DCAF is defined as:

$$\mathrm{DCAF}(\eta_1^j, \eta_2^j) \doteq |\eta_1^j - \eta_2^j|, \tag{5}$$

where:

$$\eta_\ell^j \doteq \frac{\text{number of activations of the } j^{\text{th}}\text{-atom for } c_\ell}{\text{number of } c_\ell \text{ samples}}. \tag{6}$$

The measure defined in (5) is symmetric, its value is always ≥ 0, and is inexpensive in terms of computing[1].

It can easily be seen that the definition of η_ℓ^j in (6) is equal to the maximum likelihood estimation of the conditional probability of activation, i.e.:

$$p_{\mathcal{K}_j}(k \neq 0 | c_\ell) \approx \eta_\ell^j. \tag{7}$$

Replacing this expression in (5) we can write,

$$\begin{aligned} \mathrm{DCAF}(\eta_1^j, \eta_2^j) &\approx |p_{\mathcal{K}_j}(k \neq 0|c_1) - p_{\mathcal{K}_j}(k \neq 0|c_2)|, \\ &\approx |\big(1 - p_{\mathcal{K}_j}(k = 0|c_1)\big) - \big(1 - p_{\mathcal{K}_j}(k = 0|c_2)\big)|, \\ &\approx |p_{\mathcal{K}_j}(k = 0|c_2) - p_{\mathcal{K}_j}(k = 0|c_1)|, \end{aligned} \tag{8}$$

finally expressing the DCAF in terms of the complementary conditional probabilities that the atoms will not be activated. With the exception of the F, all the measures presented in Section 5.1 can be expressed as summations, where only one of the terms is computed using the probabilities that $k = 0$. However, due to the high sparsity of the representations the terms associated with $k = 0$ are particularly important. This fact allows us to expect some correlation between the results obtained with the different discrepancy measures and the DCAF.

Figure 4 shows a graphical interpretation of measures obtained by using DCAF for examples of a discriminative atom (upper) and a non-discriminative atom (bottom). It can be seen that there are exist significant differences, in terms of magnitude orders, between the vertical scales of the graphics of the conditional PMFs (left) and the corresponding ones zeroing the probability values for $k = 0$ (center). This fact highlights the importance of the activation probability when working with sparse representations. However, the discrepancy between the distributions is not restricted to the activation probability, since there are also differences in the probability values for all other k, as shows the zoomed region of the graphic (middle bottom). It must be mentioned that the sum over the values of k of the difference of conditional probabilities (the value of the areas shown in gray in the figure) are not necessarily equal to the corresponding DCAFs. Nevertheless, for symmetric PMFs with high kurtosis and heavy tails, where the conditional distributions are similar to the a-priori distribution, the value of the gray areas can be reasonably associated with the value of the DCAF. This assumptions are well met by the PMFs of the sparse representations used in this work.

6 Experimental setup

This section presents the proposed system and its configuration settings, aimed at detecting patients suspected of suffering from moderate-severe OSAH syndrome. It also describes the database used for training and testing the method along with the measures selected for assessing its performance.

The main objective of our research is to explore the effect of using discrepancy measures to rank the atoms according to their discriminative power. Also, the experiments are designed to determine the effect of using dictionaries with different degree of over-completeness (redundant dictionaries) for the

[1] If the classes are balanced the DCAF can be replaced just by simply counting, without the necessity of dividing with the number of samples.

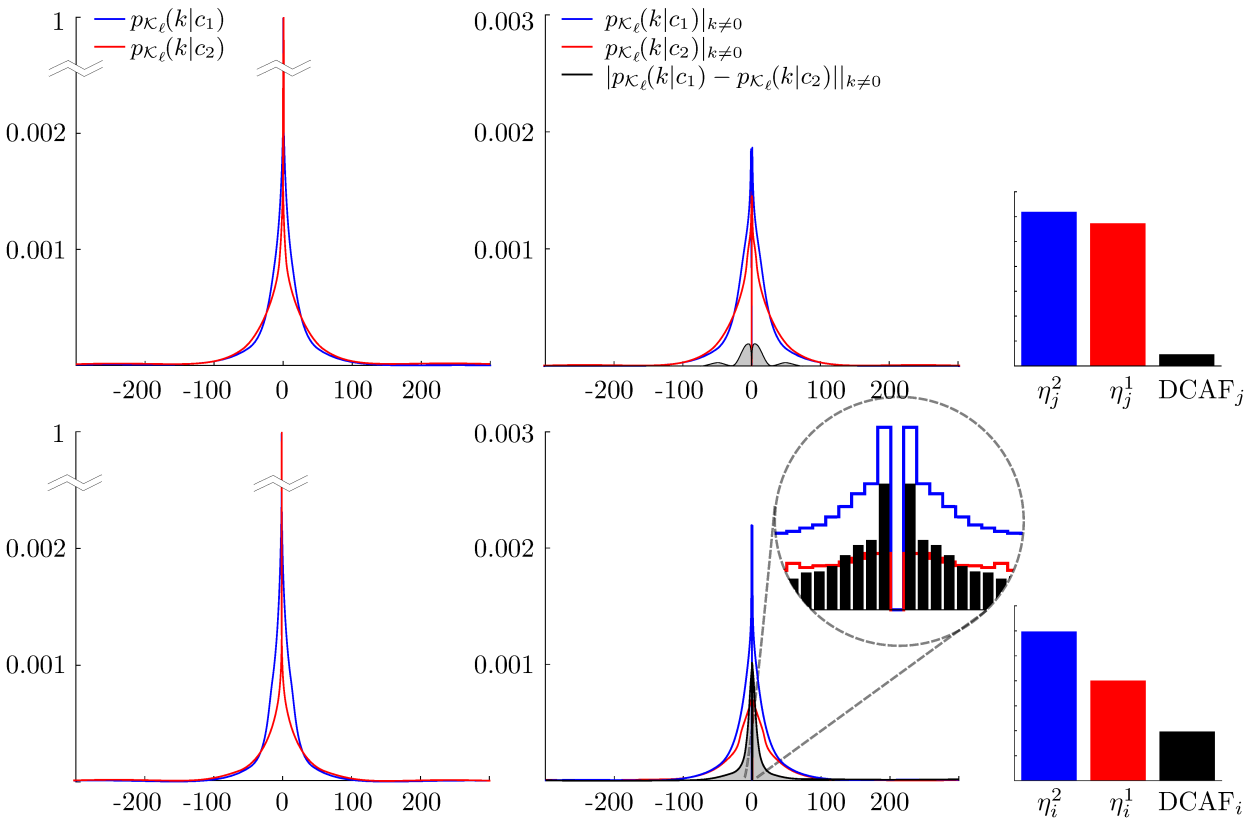


Figure 4: Several representations for the activations of a non-discriminative atom ϕ_j (upper) and a discriminative one (bottom) according to the DCAF. It can be seen the conditional PMFs (left). In the middle the conditional PMFs without the zero are shown with the absolute value of the difference in the probabilities for each k (without tacking into account the value of $k = 0$). To the right the values of η_1^ℓ and η_2^ℓ as well as the absolute value of their difference are depicted.

detection of apnea-hypopnea events. Additionally, the performance of the system for different sizes of sub-dictionaries and sparsity degrees is analyzed.

Figure 5 shows a simplified block diagram of the presented system. It can be observed that our system comprises a training phase (above) and a testing phase (below). To clarify the system's description, we divided it into three different stages, namely Stage I, Stage II and Stage III. It can be seen that stages I and II are included into training and testing phases while Stage III is only used during testing. Stage I is composed by a pre-processing block whose inputs are the raw SaO_2 signals and its outputs are filtered segments of such signals, as described in Section 6.1. At the training phase, Stage II receives segmented signals and finds an optimal discriminative sub-dictionary. During the testing phase, Stage II obtains a sparse matrix in terms of the previously found sub-dictionary. These processes are thoroughly described in Section 6.2. Finally, the obtained sparse codes are used as input of Stage III. This stage detects apnea-hypopnea events and estimates the AHI value, as described in Section 6.3.

6.1 Database and signal's pre-processing

The sleep heart health study (SHHS) dataset [38, 39] was originally designed to study correlations between sleep-disordered breathing and cardiovascular diseases. This dataset includes a large number of PSG studies, each of them containing several physiological signals such as EEG, ECG, nasal airflow, SaO_2, among others. Medical expert annotations of sleep stages, arousals and apnea-hypopnea events are also provided. In this work, only the SaO_2 signal (sampled at 1Hz) and its corresponding apnea-hypopnea labels are considered for performing the experiments. In this article, the first online version of such a database (SHHS-2) is used. This version of the database contains a total of 995 freely available PSG studies[2].

[2] `https://physionet.org/physiobank/`

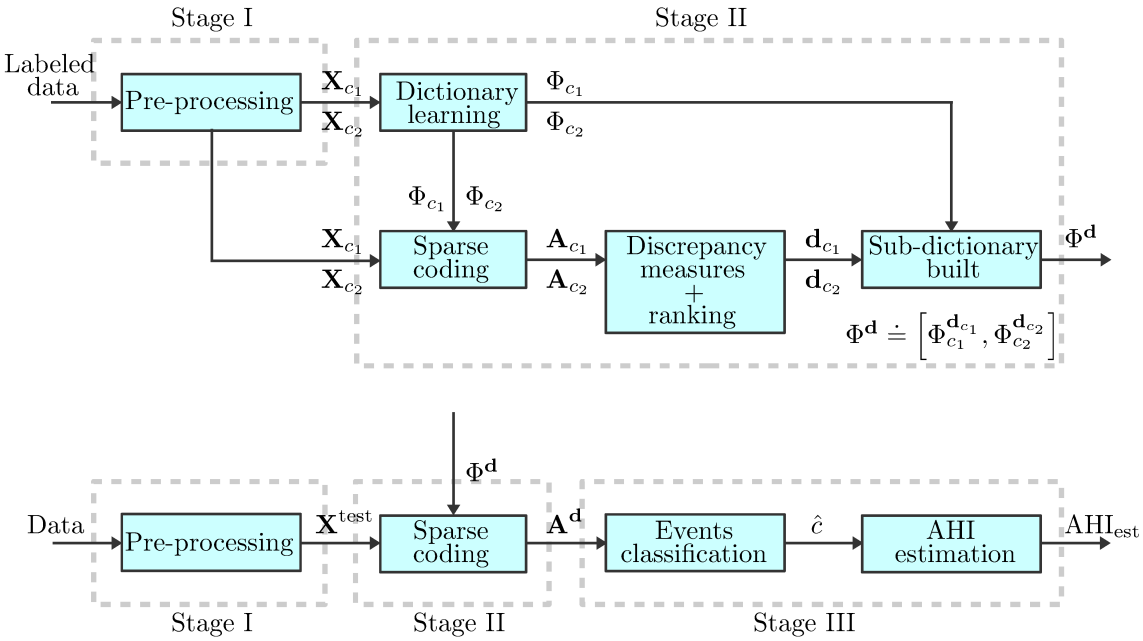


Figure 5: Block diagram of the proposed system during training (top) and testing (bottom).

The SaO_2 signals are mainly degraded by patient movements, baseline wander, disconnections and the limited resolution of pulse oximeters, among others factors. When a disconnection occurs, the recording during the time interval where the sensor signal is blocked is lost. In order to overcome this inconvenient, the values of blood oxygen saturation during such an interval are linearly interpolated. To denoise the signals, a wavelet processing technique [40] is used. The denoising process is performed by zeroing the approximation coefficients at level 8, as well as the coefficients of the first three detail levels of the discrete dyadic wavelet transform with mother wavelet Daubechies 2. The signals are then synthesized using the modified wavelet coefficients by inverse discrete dyadic wavelet transform. The application of this wavelet decomposition technique has the effect of a band-pass filter where the baseline wander and both the low frequency noise and the high frequency noise, as well as the quantization noise are eliminated. Figure 6 shows a small fragment of the original raw SaO_2 signal (top) and its wavelet-filtered version (bottom). Labels of apnea-hypopnea events (dashed lines) introduced by the medical experts are also added. These labels were generated by medical experts using the airflow information and thus are not aligned to the desaturations, i.e. there is a variable delay between the start time of an event and the corresponding desaturation.

The application of the sparse representation technique requires an appropriate segmentation of the signals. Segments of length $N = 128$ (corresponding to 128 seconds of the signal recording) with a 75% overlapping between two consecutive segments are taken. In this process, the time intervals where a disconnection occurs are not taken into account. The segments of pulse oximetry signals are then simultaneously arranged as column vectors $\mathbf{x}_i \in \mathbb{R}^N$ and labeled with ones (c_1) and minus ones (c_2), where a one corresponds to apnea-hypopnea events, and a minus one to the lack of it. Finally a signal matrix $\mathbf{X}$ is built by stacking side-by-side the column vectors $\mathbf{x}_i$, i.e. the signal matrix is defined as $\mathbf{X} \doteq [\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_n]$.

As mentioned above, the entire dataset used in this work contains 995 complete studies, 41 of which were not taken into account for performing the experiments since the size of the signal vectors differs from the corresponding vector of class labels. Among the remaining 954 studies, a subset of 667 (70%) studies were randomly selected and fixed for learning the dictionary and training the classifier. The remaining 287 (30%) studies were left out for the final test. The SaO_2 signals were filtered using wavelet filters and segmented as explained previously into column vectors of size 128. After performing the filtering and segmentation process, a signal matrix $\mathbf{X}^{\text{train}}$ of size 128×455515 is assembled by joining two previously constructed signal matrices, one for each class, $\mathbf{X}^{\text{train}} \doteq [\mathbf{X}_{c_1}^{\text{train}}\ \mathbf{X}_{c_2}^{\text{train}}]$, which contain 183163 and 272352 segments, respectively. On the other hand, for each study included into the testing dataset, a testing matrix $\mathbf{X}^{\text{test}}$ is built.

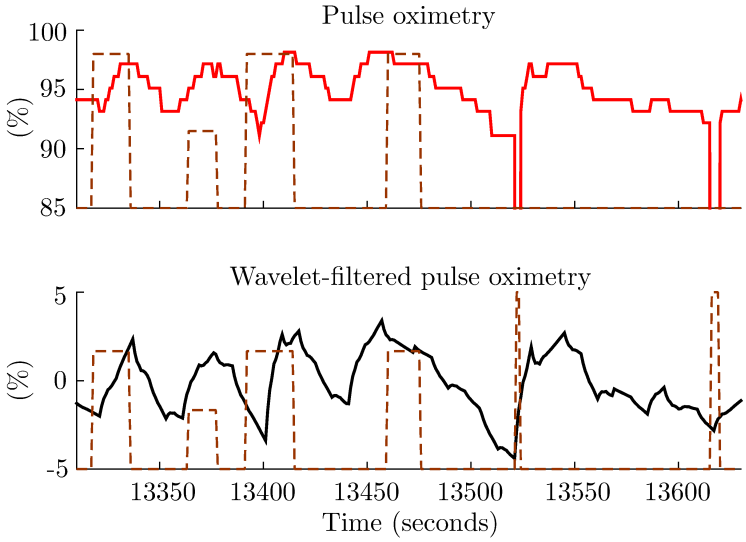


Figure 6: A small fragment of a pulse oximetry signal (top) and its wavelet-filtered version (bottom). Dashed lines represent labels of apnea-hypopnea events established by the medical expert.

6.2 Sparse coding and sub-dictionary construction

In our experiments, the learning of the dictionaries is performed by using the traditional KSVD method [14]. Optimized MATLAB codes for dictionary learning using KSVD as well as for sparse coding using the OMP algorithm are freely available for academic and personal use at the Ron Rubinstein's personal web page[3]. At the beginning, the atoms assigned to conform the initial dictionary are randomly selected from the input signal matrix for training without tacking into account any information about the classes. If the signal's space dimension is fixed, which should be the effect of constructing dictionaries with different over-completeness degree?. To answer this question, three types of dictionaries denoted by $\Phi 1$ of size 128×128, $\Phi 2$ of size 128×256 and $\Phi 4$ of size 128×512, corresponding to redundancy factors of 1, 2 and 4, respectively, were built. First the dictionary $\Phi 1$ was constructed by joining two sub-complete dictionaries of sizes 128×64 denoted by $\Phi 1_{c_1}$ and $\Phi 1_{c_2}$ learned using a large number of training segments (a total of 100,000 segments for each of the classes) belonging to the classes c_1 and c_2, respectively. Following the same idea, redundant dictionaries denoted by $\Phi 2$ (256 atoms) and $\Phi 4$ (512 atoms) were appropriately built. At the dictionary learning stage the number of non-zero elements was selected and fixed as a percentage value of 12.5 of the atoms conforming the dictionary. Also a total of 30 iterations of the KSVD algorithm were performed.

Once the dictionary has already been trained, the sparse representation vectors $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n$ corresponding to the input signals $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ are obtained by applying the OMP algorithm. In such a procedure, the nearest integer number to a percentage value of 12.5 of M is selected and fixed. The reason for having chosen this percentage value is because it presented the best trade-off between representativity and discriminability of the segments. Thus, sparsity values of $q = 16$, $q = 32$, and $q = 64$ are selected to represent the input signals for training in terms of the full dictionaries $\Phi 1$, $\Phi 2$, and $\Phi 4$, respectively.

Histograms are typically used to approximate data distributions. In this work we make use of histograms of the atom's activations to approximate the PDFs. The discretization process was performed by using a Δ value of 0.5. The detection of the most discriminative atoms is obtained by maximizing the discrepancy between the conditional PMFs of the atom's activations given the classes. This objective is achieved using the proposed DCAF measure as well as those denoted by KL, J, JS and F. The application of different discrepancy measures to the sparse vectors allows for the selection of different "discriminative atoms", which implies the construction of discriminative sub-dictionaries which are essentially different. The construction of sub-dictionaries, here denoted by $\Phi 1^{\mathbf{d}}$, $\Phi 2^{\mathbf{d}}$, and $\Phi 4^{\mathbf{d}}$, is performed by selecting atoms from $\Phi 1$, $\Phi 2$, and $\Phi 4$, respectively. Once the most discriminative atoms are detected, the sub-dictionary is built and consequently the feature vectors are obtained by applying the OMP algorithm. Finally each feature vector is assigned to be the input of the ELM classifier.

[3] `http://www.cs.technion.ac.il/~ronrubin/software.html`

6.3 Events detection and AHI estimation

Multilayer perceptron (MLP) neural networks trained for signal classification have proved to be a tool which provides quite good performances for OSAH syndrome detection [4], however, the process of training this class of neural network becomes very costly mainly in terms of time. For this reason, in this work we propose the use of extreme learning machine (ELM) [41] which is a type of single-hidden layer feedforward neural networks (SLFNs), instead of using MLP neural networks. Theoretically, this algorithm (ELM) results in providing a good generalization performance at extremely fast learning speed. The experimental results based on a few artificial and real benchmark function approximation and classification problems including large complex applications show that ELM can produce good generalization performance in most cases and can learn thousands times faster than conventional popular learning algorithms for feedforward neural networks [42].

Basic ELM classifier's MATLAB codes are available for download on the Guang-Bin Huang's web page[4]. To train such a classifier, the main parameters to be fixed are the number of neurons in the hidden layer as well as the activation function of the neurons. In our experiments, the number of neurons in the hidden layer of the ELM corresponds to four times the feature vector dimension. Also the well know sigmoid activation function, which is the most common activation function in the nodes of the hidden and/or output layer, is chosen.

In order to evaluate the performance of the proposed classifier in the detection of individual apnea-hypopnea events (a local approach), or more specifically, in the identification of persons suspected of suffering from moderate-severe OSAH syndrome (a global approach), three performance measures are used. For the identification of single segments containing apnea-hypopnea events, the sensitivity ($\mathrm{SE_{AH}}$) represents the total number of correctly classified segments of signals for which any apnea-hypopnea event occurred. Following the same idea, for the detection of individual segments of signals "not containing" any apnea-hypopnea event, the specificity ($\mathrm{SP_{AH}}$) is defined as the total number of correctly classified segments for which any apnea-hypopnea is not present. The accuracy ($\mathrm{AC_{AH}}$) is finally defined as follows:

$$\mathrm{AC_{AH}} \doteq \frac{1}{n}\sum_{i=1}^{n}\delta(c_i,\hat{c}_i), \tag{9}$$

where n represents the total number of segments, c_i and $\hat{c}_i$ denote the corresponding class label of the i^{th}-segment and the corresponding prediction of the classifier, respectively, and $\delta(x,y)$ represents the delta function whose output is true (one) if the condition $x=y$ is satisfied and false (zero) otherwise.

The differences in performance obtained for the event detection between each discrepancy measure were evaluated in order to test whether or not they are statistically significant. The test was performed assuming statistical independence of the classification errors for the different studies and approximating the error's Binomial distribution by means of a normal distribution. This assumptions are reasonable due to the large number of SaO_2 signal segments available for each study (about 1100 segments per study, totaling 301306 segments).

The estimated AHI index ($\mathrm{AHI_{est}}$) is defined as the average number of predicted events per hour of study. This new index is used for OSAH syndrome detection. In this case, the sensitivity ($\mathrm{SE_{OSAH}}$) is defined as the ratio of persons with OSAH syndrome for whom the final test is positive, and the specificity ($\mathrm{SP_{OSAH}}$) is defined as the ratio of health patients for whom the final test is negative. Also the area under the ROC curve (AUC) derived from a receiver operating characteristic (ROC) analysis [43] is used. A ROC analysis consists of computing the values of the sensitivity and specificity across all the possible detection threshold (DT) values. Then, the ROC curve is built by performing a plot of 1-specificity versus sensitivity values. This curve has been widely used by medical physicians for evaluating diagnostic tests [44]. A comparison between two different methods can be effectively done by finding the optimal cut-off point, in certain sense, of the curve and evaluating their corresponding performances. Finally, the accuracy $\mathrm{AC_{OSAH}}$ is defined as follows:

$$\mathrm{AC_{OSAH}} \doteq \frac{1}{m}\sum_{i=1}^{m}\delta(\mathrm{AHI}_{\mathrm{est}}^{(i)} > \mathrm{DT}, \mathrm{AHI}^{(i)} > 15), \tag{10}$$

where m corresponds to the total number of studies coming from testing dataset and "DT" is the detection threshold value that results in the best cut-off point of the ROC curve. This point, which maximizes simultaneously sensitivity and specificity, corresponds to the minimum euclidean distance (d_{min}) to the point (0;1) of the ROC curve.

[4] `http://www.ntu.edu.sg/home/egbhuang/elm_codes.html`

7 Results and discussion

In this section results of the performed experiments are presented and discussed. This section is mainly separated into two sub-sections, namely *i*) the performance tunning section and *ii*) the optimal system performance section.

7.1 Performance tunning

This section presents results of the exploratory experiments performed to find optimal configurations of the proposed system. As explained in Section 6.2, three different full dictionaries called $\Phi 1$, $\Phi 2$ and $\Phi 4$ were learned by applying the standard KSVD algorithm. In this process, it is expected that most dictionary atoms would capture high frequency oscillations and normal respiration cycles in SaO_2 signals. It is important to point out however that, typical desaturations in signals associated to apnea-hypopnea events should be encoded by some atoms. Secondly, the sparse matrices $\mathbf{A}1$, $\mathbf{A}2$ and $\mathbf{A}4$ were obtained by applying the OMP algorithm. As described in Section 6.2 several measures were used to quantify the discriminative degree of individual atoms of each one of the studied dictionaries. Finally, the dictionary atoms were ranked in decreasing order of magnitude according to their discriminative power. Figure 7 shows the waveforms of the first seven ranked atoms of the dictionary $\Phi 1$ according to our measure (first row) as well as the first seven ranked atoms of such a dictionary according to all other discrepancy measures (rows from two to five). It can be seen that the most discriminative atom selected by DCAF (dashed waveform) provides information about two well-defined desaturations in the signal. It is also important to point out that, this atom corresponds to the most discriminative one when using J divergence, or eventually when using the JS divergence. Moreover, one can clearly note that no highly discriminative atoms were taken when using Fisher score.

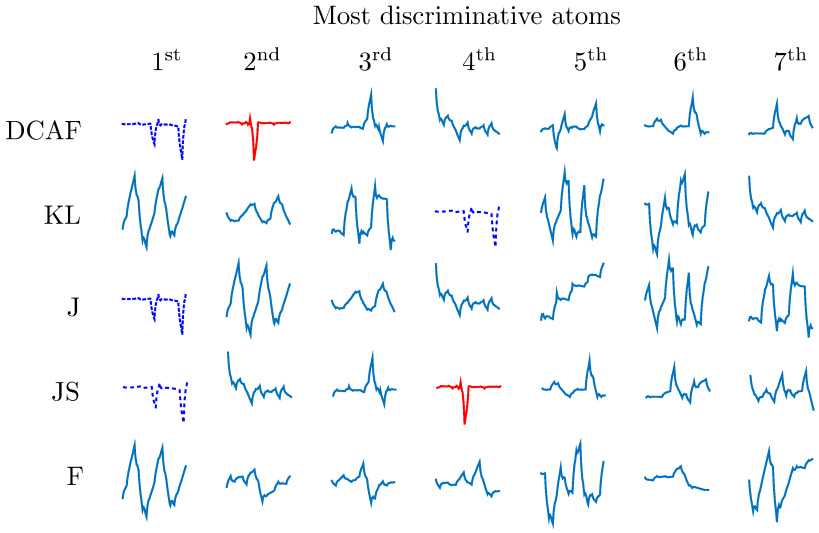


Figure 7: Waveforms corresponding to the first seven ranked atoms according to each one of the evaluated measures.

Discriminative sub-dictionaries called $\Phi 1^{\mathbf{d}}$, $\Phi 2^{\mathbf{d}}$ and $\Phi 4^{\mathbf{d}}$ were built by stacking side-by-side the first p ranked atoms from $\Phi 1$, $\Phi 2$ and $\Phi 4$, respectively, according to their discriminative degree. It is appropriate to mention that, the evaluation of several discrepancy measures leads to the construction of different discriminative sub-dictionaries. However, optimal values of p (sub-dictionary size) and q (sparsity level) are parameters that need to be tuned. In order to find optimal values of such hyper-parameters, a grid search was performed.

The performance of our system was first tested by performing a "random selection" of the dictionary atoms. The involved results were fixed and appropriately used as reference. The random selection of the atoms was performed ten times. Additionally, for each one of the atoms random selection, 60 iterations of the grid search were performed. Thus, the accuracy rate's variations introduced by the classifier were minimized. Figure 8 shows three images corresponding to averaged accuracy rates for each one of the evaluated dictionaries. Averaged accuracy rates (reference values) obtained by using the dictionary $\Phi 1$ for the detection of apnea-hypopnea events are shown on the left of this figure. It can be seen that sparse representations in terms of $\Phi 1$, using the smallest sub-dictionary size and the highest sparsity

degree, result in better performance than the ones obtained by using all other configurations of $\Phi 1$ and the over-complete dictionaries $\Phi 2$ and $\Phi 4$. In this way, two regions can be distinguished corresponding to a high performance region and a low performance one. The first one, which is or our interest, is yielded by simultaneously employing a small sub-dictionary size (10%) and a high sparsity degree (5%).

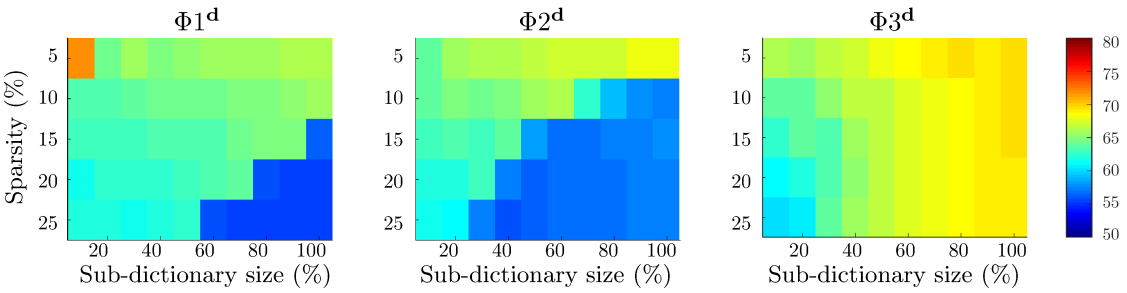


Figure 8: Averaged accuracy rates obtained by varying the percentages of the sub-dictionary size and the sparsity level according to a random ranking of the atoms.

Next, DCAF and four other discrepancy measures were used for appropriately constructing discriminative sub-dictionaries. Then, a grid search of hyper-parameters was performed by analyzing the performance that yields our system when using each one of the sub-dictionaries. Figure 9 shows five images corresponding to DCAF (upper-left) and the other four discrepancy measures. These images represent the differences between accuracy rates obtained by using discriminative measures and the reference one (random selection) for $\Phi 1$. Also, each pixel of these images correspond to particular percentages of sub-dictionary size and sparsity level. It can be observed that, independently of the discriminative measure, small percentages of sub-dictionary size yield good performances. It is appropriate to point out however that, the effect of the dimension (sub-dictionary size) in the performance of the system is more important than the one induced by using discriminative measures.

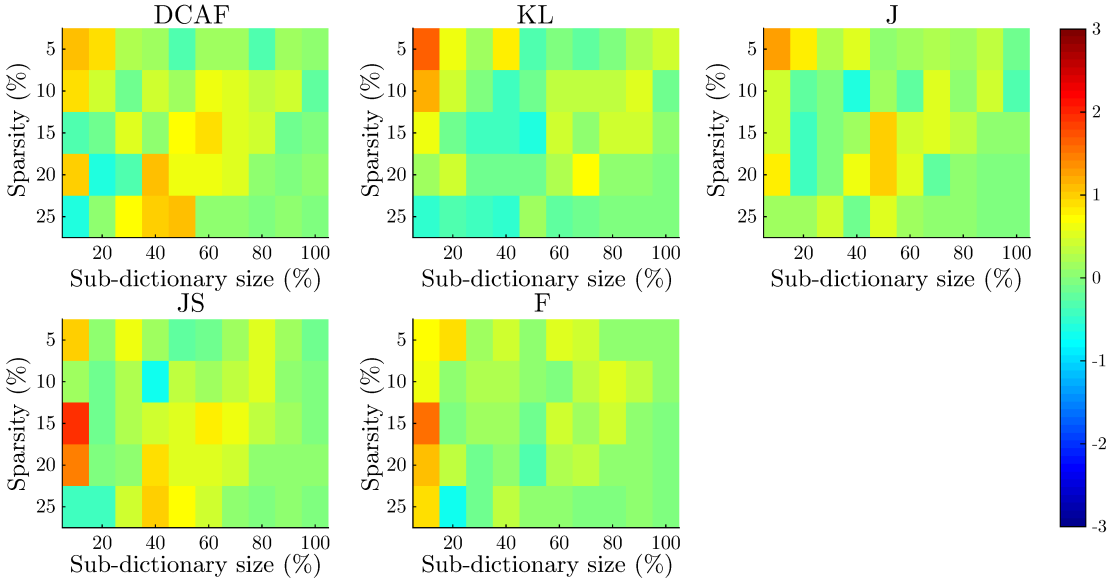


Figure 9: Five images representing differences between accuracy rates yielded by DCAF and all other discrepancy measures and random selection for $\Phi 1$.

Analogously, Figures 10 and 11 show five images which correspond to DCAF (upper-left) and all other discrepancy measures. The images depicted in Figures 10 and 11 represent the differences between accuracy rates obtained by using these measures and the reference one for dictionaries $\Phi 2$ and $\Phi 4$, respectively.

If we compare the results shown in Figures 9, 10, and 11, then it can be conclude that the proposed system presents the best performance, in terms of accuracy rate in the detection of apnea-hypopnea events, when using the full dictionary $\Phi 1$. Although similar results were obtained applying the proposed DCAF measure and those traditional ones (see Figure 9), it is important to point out that the use of

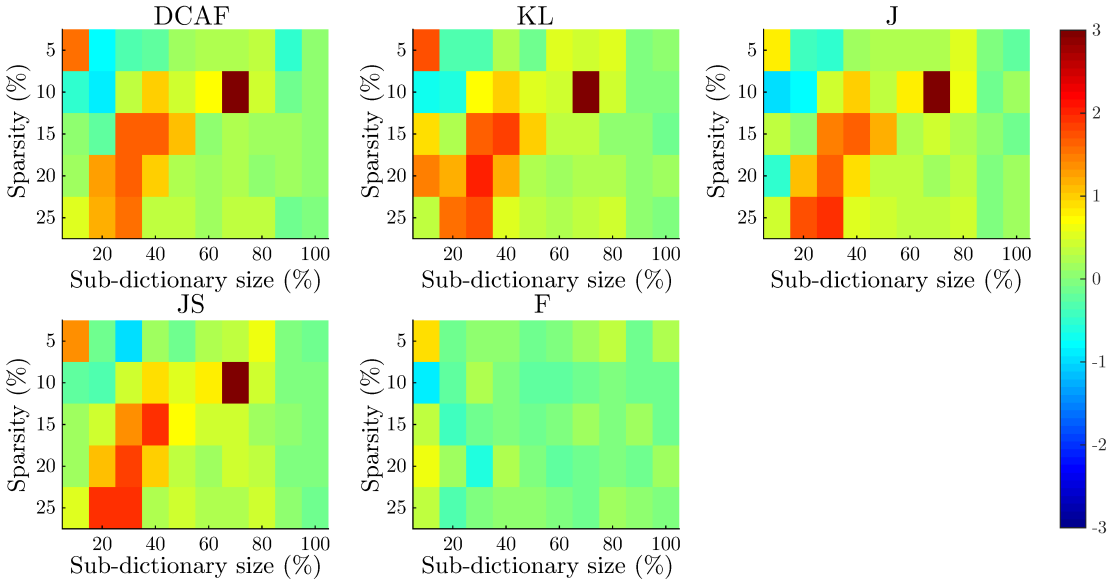


Figure 10: Five images representing differences between accuracy rates yielded by DCAF and all other discrepancy measures and random selection for Φ2.

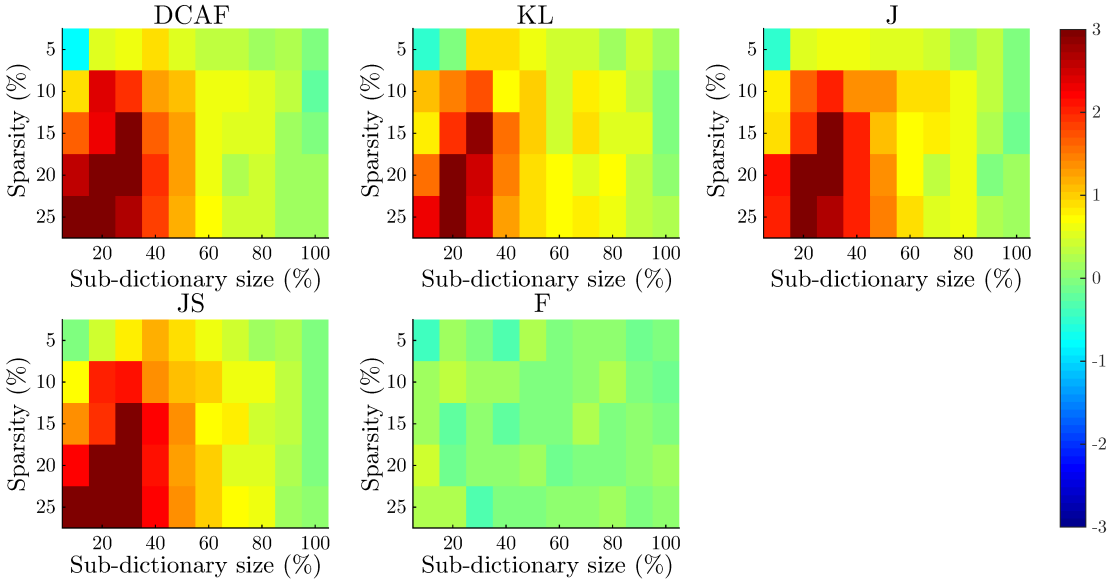


Figure 11: Five images representing differences between accuracy rates yielded by DCAF and all other discrepancy measures and random selection for Φ4.

discrepancy measures resulted in a significantly high improvement with respect to a "random" selection of the atoms. As discussed above, the dimension reduction in the sub-dictionary size as well as high sparse levels yielded high accuracy rates. This is the reason for which a small sub-dictionary size (10%) and high sparse level (5%) were chosen to perform the final test.

System performance changes were analyzed by performing a comparison between averaged accuracy rates obtained by using discriminative sub-dictionaries and the ones obtained by using full dictionaries. Table 1 shows averaged accuracy percentages obtained by taken into account fixed discriminative sub-dictionary sizes (10%) while allowing the sparsity level to change (rows from 3 to 7). The last row of this table presents averaged accuracy percentages yielded by using full dictionaries for different sparsity levels. It can be observed that, in all of cases, discriminative sub-dictionaries outperform full dictionaries in the detection of apnea-hypopnea events.

The impact of sparsity degree in the performance of our system is illustrated in Table 2. These results

Table 1: Averaged accuracy rates for sub-dictionary sizes of 10% regarding to each one of the evaluated full dictionaries.

	$\Phi 1^{\mathbf{d}}(128 \times 12)$		$\Phi 2^{\mathbf{d}}(128 \times 24)$		$\Phi 4^{\mathbf{d}}(128 \times 50)$	
Measure	Max	Avg	Max	Avg	Max	Avg
DCAF	72.62	64.68	65.20	63.15	65.19	64.21
KL	73.20	64.91	65.44	63.53	65.42	63.66
J	72.82	64.88	64.50	62.82	65.39	63.68
JS	72.55	64.10	65.02	63.18	65.87	64.01
F	72.23	65.21	64.57	63.04	65.64	62.71
Full dictionary	66.39	59.77	68.13	59.57	69.28	69.21

were yielded by averaging accuracy rates obtained for a sparsity level of 5% and considering all possible sub-dictionary sizes (from 10% to 90%). For example, the second row shows averaged accuracy rates obtained by means of discriminative sub-dictionaries whose atoms were taken from $\Phi 1$, $\Phi 2$ and $\Phi 4$ by using DCAF measure.

Table 2: Averaged accuracy rates by considering a sparsity level of 5% regarding to all possible sub-dictionary sizes.

Measure	$\Phi 1$	$\Phi 2$	$\Phi 4$
DCAF	66.41	66.51	67.95
KL	66.49	66.72	67.98
J	66.60	66.56	67.98
JS	66.41	66.57	68.15
F	66.53	66.54	67.58

7.2 Optimal system performance

Optimal system configurations were selected and fixed to perform the final test. In the previous section it was found that discriminative sub-dictionaries constructed by taken atoms from the dictionary $\Phi 1$ yields better performances than the ones constructed by selecting atoms from the dictionaries $\Phi 2$ and $\Phi 4$. Additionally, it was found that a discriminative sub-dictionary composed by only 12 atoms (10%) and a sparsity level of one (5%) yield in the best accuracy rate of our system.

In order to overcome the variance introduced by ELM predictors, 60 repetitions of the testing process were performed. Table 3 shows percentage values of minimum (Min), maximum (Max), average (μ) and standard deviation (σ) corresponding to obtained accuracy rates in the detection of apnea-hypopnea events. Although, DCAF perform similarly to the four other discrepancy measures, its performance is achieved with a relatively low computational cost. Additionally, results show that performances obtained by using discriminative measures for constructing sub-dictionaries always outperform the ones yielded by making use of randomly constructed sub-dictionaries.

Table 3: Averaged accuracy rates for a sub-dictionary percentage of 10 for the detection of apnea-hypopnea events.

Measure	Min	Max	μ	σ
DCAF	71.72	73.14	72.57	0.345
Kullback-Leibler	72.06	73.78	73.26	0.390
Jeffrey	71.77	73.31	72.66	0.319
Jensen-Shannon	71.79	73.11	72.55	0.295
Fisher	71.01	72.77	72.18	0.325
Random Selection	70.01	71.51	70.91	0.372

We have also evaluated the statistical significance of the results presented in Table 3 by computing the probability that using each one of the evaluated measures, including random selection (RS), yields in better classification performances than the others. In order to perform this test, we assumed the statistical independence of the classification errors for each study. Also it was possible to approximate the error's binomial probability distribution by a normal distribution due to a wide availability of signals (301,306). Table 4 summarizes the results of the performed statistical significance tests by considering a p-value

of 0.01. It can be seen that DCAF and three other discrepancy measures (KL, J and JS divergences) are significant different respect to random selection. Also, no significant difference was found between F score and random selection. Additionally is was found that DCAF does not perform significantly better that the KL, J and JS divergences.

Table 4: A summary of the performed statistical significance tests.

	RS	DCAF	KL	J	JS	F
RS	-	✓	✓	✓	✓	✗
DCAF	-	-	✗	✗	✗	✗
KL	-	-	-	✗	✗	✗
J	-	-	-	-	✗	✗
JS	-	-	-	-	-	✗
F	-	-	-	-	-	-

To determine the severity degree of the OSAH syndrome, a ROC curve analysis was successfully performed by considering a detection AHI of 15. This index was selected in order to be able to identify patients suspected of suffering from moderate-severe OSAH syndrome. Table Table 5 shows the minimum operating (cut-off) point of the ROC curves and maximum percentages of sensitivity, specificity and accuracy as well as maximum values of area under the ROC curve for AHI diagnostic threshold values of 15 (Figure 12 (left)). It can be seen that DCAF resulted in a maximum area under the ROC curve of 0.9250 and sensitivity and specificity percentages of 81.88 and 87.32, respectively. These are the maximum performance measures at which the minimum cut-off point of the ROC curve is attained. If we compare the performances attained between all of the evaluated measures, then the maximum SE and AUC value is yielded by J divergence. Also, JS divergence outperformed all the others in terms of ACC and DCAF resulted in the minimum cut-off point of the ROC curve.

Table 5: Maximum cut-off points for testing accuracy for a sub-dictionary percentage of 10 for the detection of apnea-hypopnea events.

Measure	d_{min}	SE	SP	ACC	AUC
DCAF	0.2211	81.88	87.32	84.60	0.9250
Kullback-Leibler	0.2242	81.46	87.39	84.43	0.9271
Jeffrey	0.2311	80.86	87.04	83.95	0.9283
Jensen-Shannon	0.2267	80.75	88.03	84.39	0.9244
Fisher	0.2280	80.66	87.91	84.29	0.9252

We additionally performed a ROC curve analysis of the averaged performances of DCAF and all the other discrepancy measures (Figure 12 (right)). A random selection was additionally included in our results in order to be able to compare performance changes. Table 6 the averaged minimum operating (cut-off) point of the ROC curves and averaged maximum percentages of sensitivity, specificity and accuracy as well as averaged maximum values of AUC values for the same OSAH syndrome diagnostic threshold. The result show that DCAF outperform all the other discrepancy measures in terms of minimum optimal operating cut-off point of the ROC curve as well as in terms of sensitivity and accuracy rate. Also KL divergence resulted in the best averaged area under the curve ROC and the maximum averaged specificity was yielded by JS divergence. A significantly high performance improvement was observed when using DCAF or any of the other discrepancy measures compared to random selection.

Table 6: Averaged cut-off points for testing accuracy for a sub-dictionary percentage of 10 for the detection of apnea-hypopnea events.

Measure	d_{min}	SE	SP	ACC	AUC
DCAF	0.2211	81.88	87.32	84.60	0.9250
Kullback-Leibler	0.2242	81.46	87.39	84.43	0.9271
Jeffrey	0.2311	80.86	87.04	83.95	0.9283
Jensen-Shannon	0.2267	80.75	88.03	84.39	0.9244
Fisher	0.2280	80.66	87.91	84.29	0.9252
Random Selection	0.2396	80.85	85.60	83.23	0.9222

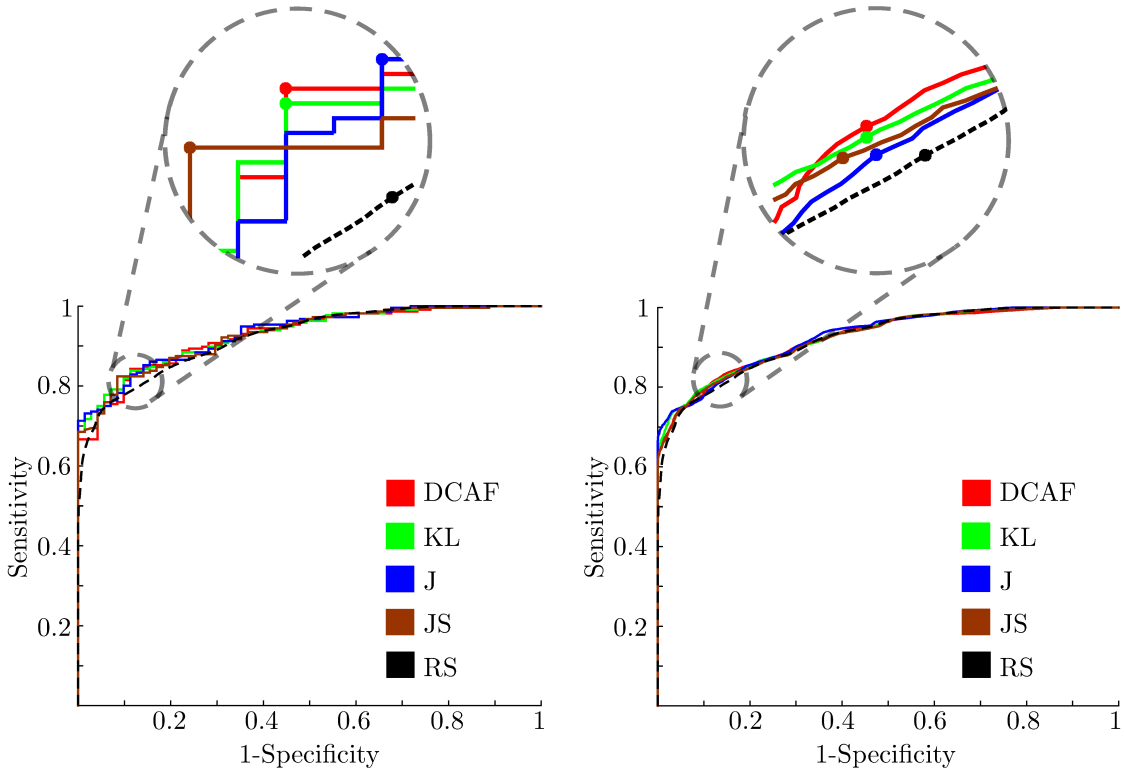


Figure 12: ROC curves corresponding to the performance measures described in Tables 5 and 6.

8 Conclusions

Sparse representations of signals constitute a powerful technique which yields high accuracy rates in the detection of apnea-hypopnea events. In this work the difference of conditional activation frequency (DCAF) measure was successfully used for accurately pointing out discriminative atoms in a full dictionary. Additionally, we compared the performance of the DCAF with four widely used discrepancy measures. It was found that DCAF and three other discrepancy measures (KL, J y JS divergences) outperform the random selection of atoms, unlike F score. Additionally, DCAF is cheaper to compute. Discriminative sub-dictionaries were successfully constructed by taking the best ranked atoms of full dictionaries according to their discriminative power. Results show that sparse representations of signals in terms of discriminative sub-dictionaries result in better performances than the ones obtained in terms of full dictionaries in the detection of apnea-hypopnea events by using only pulse oximetry signals. In this context, it was found that more sparse solutions almost always yielded in better performances. Additionally, it was observed that larger dictionary over-completeness worsens the performance of the system. Future research lines include more analysis of the DCAF measure, the study of its properties and an extension of such a measure to multi-class problems, among others.

9 Acknowledgments

This work was supported in part by Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, through PIP 2014-2016 Nro. 11220130100216-CO and PIP 2012-2014 Nro. 114 20110100284-KA4, by the Air Force Office of Scientific Research, AFOSR / SOARD, through Grant FA9550-14-1-0130 and by Universidad Nacional del Litoral through projects CAI+D PIC Nro. 504 201501 00098 LI (2016) and PIC Nro. 504 201501 00036 LI (2016).

References

[1] N. Saito and R. R. Coifman, "Local discriminant bases and their applications," *Journal of Mathematical Imaging and Vision*, vol. 5, no. 4, pp. 337–358, 1995.

[2] S. Tabibian, A. Akbari, and B. Nasersharif, "Speech enhancement using a wavelet thresholding method based on symmetric Kullback–Leibler divergence," *Signal Processing*, vol. 106, pp. 184–197, 2015.

[3] M. Sánchez-Gutiérrez, E. M. Albornoz, H. L. Rufiner, and J. G. Close, "Post-training discriminative pruning for rbms," *Soft Computing*, pp. 1–15, 2017.

[4] R. E. Rolón, L. D. Larrateguy, L. E. Di Persia, R. D. Spies, and H. L. Rufiner, "Discriminative methods based on sparse representations of pulse oximetry signals for sleep apnea–hypopnea detection," *Biomedical Signal Processing and Control*, vol. 33, pp. 358–367, 2017.

[5] V. Peterson, H. L. Rufiner, and R. D. Spies, "Generalized sparse discriminant analysis for event-related potential classification," *Biomedical Signal Processing and Control*, vol. 35, pp. 70–78, 2017.

[6] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[7] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[8] A. Gupta, S. Parameswaran, and C.-H. Lee, "Classification of electroencephalography (EEG) signals for different mental activities using Kullback Leibler (KL) divergence," pp. 1697–1700, 2009.

[9] H. Jeffreys, "An Invariant Form for the Prior Probability in Estimation Problems," *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 186, no. 1007, pp. 453–461, 1946.

[10] J. Lin, "Divergence Measures Based on the Shannon Entropy," *IEEE Trans. Inf. Theor.*, vol. 37, no. 1, pp. 145–151, 2006.

[11] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images," *SIAM Review*, vol. 51, no. 1, pp. 34–81, 2009.

[12] X. Zhang and Q. Ding, "Respiratory rate estimation from the photoplethysmogram via joint sparse signal reconstruction and spectra fusion," *Biomedical Signal Processing and Control*, vol. 35, pp. 1–7, 2017.

[13] Y. Zhou, X. Hu, Z. Tang, and A. C. Ahn, "Sparse representation-based ECG signal enhancement and QRS detection," *Physiological Measurement*, vol. 37, no. 12, pp. 2093–2110, 2016.

[14] M. Aharon, M. Elad, and A. Bruckstein, "KSVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," *IEEE Transactions on Signal Processing*, vol. 54, pp. 4311–4322, Nov. 2006.

[15] D. S. Pham and S. Venkatesh, "Joint learning and dictionary construction for pattern recognition," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[16] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2691–2698, June 2010.

[17] M. J. Sateia, "International classification of sleep disorders-third edition: Highlights and modifications," *Chest*, vol. 146, pp. 1387–1394, Nov. 2014.

[18] N. A. Dewan, F. J. Nieto, and V. K. Somers, "Intermittent hypoxemia and OSA: implications for comorbidities," *Chest*, vol. 147, no. 1, pp. 266–274, 2015.

[19] W. Kukwa, E. Migacz, K. Druc, E. Grzesiuk, and A. M. Czarnecka, "Obstructive sleep apnea and cancer: effects of intermittent hypoxia?," *Future Oncology (London, England)*, vol. 11, no. 24, pp. 3285–3298, 2015.

[20] M. Torres, R. Laguna-Barraza, M. Dalmases, A. Calle, E. Pericuesta, J. M. Montserrat, D. Navajas, A. Gutierrez-Adan, and R. Farré, "Male fertility is reduced by chronic intermittent hypoxia mimicking sleep apnea in mice," *Sleep*, vol. 37, no. 11, pp. 1757–1765, 2014.

[21] T. Young, L. Evans, L. Finn, and M. Palta, "Estimation of the clinically diagnosed proportion of sleep apnea syndrome in middle-aged men and women," *Sleep*, vol. 20, no. 9, pp. 705–706, 1997.

[22] J. Durán, S. Esnaola, R. Rubio, and A. Iztueta, "Obstructive sleep apnea-hypopnea and related clinical features in a population-based sample of subjects aged 30 to 70 yr," *American Journal of Respiratory and Critical Care Medicine*, vol. 163, pp. 685–689, 2001.

[23] R. Thurnheer, K. E. Bloch, I. Laube, M. Gugger, M. Heitz, and Swiss Respiratory Polygraphy Registry, "Respiratory polygraphy in sleep apnoea diagnosis. Report of the Swiss respiratory polygraphy registry and systematic review of the literature," *Swiss Medical Weekly*, vol. 137, no. 5-6, pp. 97–102, 2007.

[24] E. García-Díaz, E. Quintana-Gallego, A. Ruiz, C. Carmona-Bernal, A. Sánchez-Armengol, G. Botebol-Benhamou, and F. Capote, "Respiratory polygraphy with actigraphy in the diagnosis of sleep apnea-hypopnea syndrome," *Chest*, vol. 131, no. 3, pp. 725–732, 2007.

[25] A. Yadollahi, E. Giannouli, and Z. Moussavi, "Sleep apnea monitoring and diagnosis based on pulse oximetry and tracheal sound signals," *Medical & Biological Engineering & Computing*, vol. 48, no. 11, pp. 1087–1097, 2010.

[26] M. Elad, *Sparse and Redundant Representations*. Springer-Verlag New York, 2010.

[27] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[28] J. Tropp and A. Gilbert, "Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[29] R. R. Coifman, Y. Meyer, S. Quake, and M. V. Wickerhauser, "Signal processing and compression with wavelet packets," in *Wavelets and Their Applications*, pp. 363–379, Springer, Dordrecht, 1994.

[30] M. S. Lewicki and B. A. Olshausen, "Probabilistic framework for the adaptation and comparison of image codes," *Journal of the Optical Society of America A*, vol. 16, no. 7, p. 1587, 1999.

[31] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Computation*, vol. 12, no. 2, pp. 337–365, 2000.

[32] K. Engan, S. O. Aase, and J. H. Husoy, "Method of optimal directions for frame design," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 2443–2446, 1999.

[33] Z. Jiang, Z. Lin, and L. Davis, "Label Consistent K-SVD: Learning a Discriminative Dictionary for Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 2651–2664, Nov. 2013.

[34] M. Basseville, "Distance measures for signal processing and pattern recognition," *Signal Processing*, vol. 18, no. 4, pp. 349–369, 1989.

[35] W. Gersch, F. Martinelli, J. Yonemoto, M. D. Low, and J. A. Mc Ewan, "Automatic classification of electroencephalograms: Kullback-Leibler nearest neighbor rules," *Science (New York, N.Y.)*, vol. 205, no. 4402, pp. 193–195, 1979.

[36] P. J. Moreno, P. P. Ho, and N. Vasconcelos, "A Kullback-Leibler Divergence Based Kernel for SVM Classification in Multimedia Applications," in *Advances in Neural Information Processing Systems 16* (S. Thrun, L. K. Saul, and P. B. Schölkopf, eds.), pp. 1385–1392, MIT Press, 2004.

[37] C. C. Aggarwal, *Data classification: algorithms and applications*. CRC Press, 2014.

[38] S. F. Quan, B. V. Howard, C. Iber, J. P. Kiley, F. J. Nieto, G. T. O'Connor, D. M. Rapoport, S. Redline, J. Robbins, J. M. Samet, and P. W. Wahl, "The Sleep Heart Health Study: design, rationale, and methods," *Sleep*, vol. 20, no. 12, pp. 1077–1085, 1997.

[39] B. K. Lind, J. L. Goodwin, J. G. Hill, T. Ali, S. Redline, and S. F. Quan, "Recruitment of healthy adults into a study of overnight sleep monitoring in the home: experience of the Sleep Heart Health Study," *Sleep & Breathing = Schlaf & Atmung*, vol. 7, no. 1, pp. 13–24, 2003.

[40] F. Lestussi, L. Di Persia, and D. Milone, "Comparison of on-line wavelet analysis and reconstruction: with application to ECG," *5th International Conference on Bioinformatics and Biomedical Engineering (iCBBE 2011)*, 2011.

[41] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.

[42] J. Tang, C. Deng, and G. B. Huang, "Extreme Learning Machine for Multilayer Perceptron," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 4, 2016.

[43] J. A. Swets, "ROC analysis applied to the evaluation of medical imaging techniques," *Investigative Radiology*, vol. 14, no. 2, pp. 109–121, 1979.

[44] K. Hajian-Tilaki, "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation," *Caspian Journal of Internal Medicine*, vol. 4, no. 2, pp. 627–635, 2013.

Anexo B

Discriminative methods based on sparse representations of pulse oximetry signals for sleep apnea-hypopnea detection

Artículo publicado en la revista *Biomedical Signal Processing and Control*, Elsevier, Vol. 33, pp. 358-367, 2017, DOI: `http://dx.doi.org/10.1016/j.bspc.2016.12.013`.

Discriminative methods based on sparse representations of pulse oximetry signals for sleep apnea-hypopnea detection

R.E. Rolón[a], L.D. Larrateguy[b], L.E. Di Persia[a], R.D. Spies[c], H.L. Rufiner[a,d]

[a] *Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, sinc(i), FICH-UNL/CONICET, Santa Fe, Argentina*
[b] *Centro de Medicina Respiratoria de Paraná, Argentina*
[c] *Instituto de Matemática Aplicada del Litoral, IMAL, FIQ-UNL/CONICET, Santa Fe, Argentina*
[d] *Laboratorio de Cibernética, Fac. de Ing., Univ. Nacional de Entre Ríos, Argentina*

Abstract

The obstructive sleep apnea-hypopnea (OSAH) syndrome is a very common and generally undiagnosed sleep disorder. It is caused by repeated events of partial or total obstruction of the upper airway while sleeping. This work introduces two novel approaches called most dicriminative activation selection (MDAS) and most discriminative column selection (MDCS) for the detection of apnea-hypopnea events using only pulse oximetry signals. These approaches use discriminative information of sparse representations of the signals to detect apnea-hypopnea events. Complete (CD) and overcomplete (OD) dictionaries, and three different strategies (FULL sparse representation, MDAS, and MDCS), are considered. Thus, six methods (FULL-OD, MDAS-OD, MDCS-OD, FULL-CD, MDAS-CD, and MDCS-CD) emerge. It is shown that MDCS-OD outperforms all the others methods. A receiver operating characteristic (ROC) curve analysis of this method shows an area under the curve of 0.937 and diagnostic sensitivity and specificity percentages of 85.65 and 85.92, respectively. This shows that sparse representations of pulse oximetry signals is a very valuable tool for estimating apnea-hypopnea indices. The implementation of the MDCS-OD method could be embedded into the oximeter so as to be used by primary attention clinical physicians in the search and detection of patients suspected of suffering from OSAH.

Keywords: Sleep apnea-hypopnea syndrome, Sparse representations, Dictionary learning, Neural networks

1. Introduction

In the year 2014 the American academy of sleep medicine (AASM) released the third edition of the international classification of sleep disorders [1]. One of the most common sleep disorders is the obstructive sleep apnea-hypopnea (OSAH) syndrome, which is caused by repeated events of partial (hypopnea) or total (apnea) obstruction of the upper airway while sleeping. To establish the degree of severity of the syndrome, the apnea-hypopnea index (AHI) is created. The AHI represents the number of apnea-hypopnea events per hour of sleep. The OSAH is classified as normal, mild, moderate or severe if belongs to the interval $[0, 5)$, $[5, 15)$, $[15, 30)$, or $[30, \infty)$, respectively.

Nowadays, the gold standard test for diagnosing sleep disorders is a polysomnography (PSG) in a sleep medical center. However the accessibility to this type of study is usually very limited as well as costly in terms of both time and money. A complete PSG consists of simultaneous measurement of several physiological signals such as electrical activity of the brain along the scalp, electrical activity of the heart using electrodes placed on the body's surface, electrical activity produced by skeletal muscles, respiratory effort, airflow and blood oxygen saturation (SaO_2) signals, among others. Mainly due to its ease of acquisition, we are particularly interested in the latter. In a typical PSG study, after a normal period of sleep the recorded signals are provided to medical experts. Due to its complexity, different alternatives to PSG have been developed. One of the most popular alternatives to PSG is the so called home respiratory polygraphy [2]. Although some studies have shown that there is a very high correlation between AHIs generated by polygraphy and PSG studies and polygraphy requires no neurophysiological signals [3], it still needs several others physiological signals, whose acquisition affects the normal sleeping of the persons. It is therefore highly desirable to develop a reliable system which makes use of as few as possible physiological signals. Since pulse oximetry is a well know, quite cheap and non-invasive technique, it has become a very valuable alternative to detect persons suspected of suffering from OSAH [4]. A recent work has shown that statistical analysis and feature extraction methods applied to pulse oximetry signals provide satisfactory diagnostic performance in detecting severe OSAH patients [5]. Cessation of breathing associated with apnea-hypopnea events are always accompanied by a drop in the oxygen saturation level. It is appropri-

ate to mention however that this drop level can be very small and impossible to detect by a human observer, reason for which advanced signal processing techniques such as artificial intelligence methods could provide a very valuable alternative. A decrease in blood oxygen saturation usually produces changes in the pulse oximetry record corresponding to intermittent hypoxemia. The intermittent hypoxemia, with hypoxemia-reoxygenation cycles, very often indicates OSAH syndrome.

Pulse oximetry, besides providing information about blood oxygen saturation during sleeping, is used for computing some parameters which quantify desaturation levels in the SaO_2 signal. The seek of patients suspected of suffering from OSAH can be addressed by means of two different approaches. A *global* approach consists of obtaining general characteristics of the SaO_2 signal, such as its mean, variance and entropy values, among others with the only objective of classifying a person as healthy or sick without taking into consideration the degree of severity of the illness. In this work a *local* approach, which allows a more thorough analysis of the SaO_2 signal, is taken. This approach consists of detecting the apnea-hypopnea events from sparse representations of segments of SaO_2 signals using a neural network classifier. The local approach was previously used for estimating three parameters denoted by ODI4, ODI3, and ODI2, which are defined as the number of times per hour of sleep that the SaO_2 signal decreases below 4%, 3%, and 2% of a baseline level, respectively. It is timely to point out, however that although the concept of "baseline level" is very intuitive, it is not uniquely defined and different criteria and definitions have been adopted by different authors [6, 7].

In the last fifteen years, a wide variety of machine learning algorithms were used for detecting several health disorders [8]. Implementations of these algorithms were applied to detect particular sleep disorders and different signal processing techniques originating new methods based on non-linear systems, higher-order statistics, spectral analysis, including independent component analysis (ICA) [9, 10, 11]. Moreover pattern recognition algorithms based on artificial neural network (ANN) were successfully applied to assist OSAH diagnosis and classification [12]. Nowadays, a powerful method based on sparse representations of signals finds the solution corresponding to the most compact representation by means of a linear combination of atoms in a dictionary [13, 14]. It was found that this approach, when applied to biological sensory systems, results in internal representations having properties similar to the real ones, in particular similar to those found in the primary

auditive or visual cortex of the mammals [15, 16]. Some of the advantages of the sparse representations are: super resolution, robustness to noise and dimension reduction, among others. The sparse representations of signals provide new grounds for treating both the signal modeling and the representation problems. The dictionary is learned for the purpose of obtaining the best representation of a given set of signals, although the atoms involved in such representation are not necessarily the atoms which capture discriminative information. It is therefore clear that if the SaO_2 signal is to be used as the only input for detection of apnea-hypopnea events, advanced signal processing algorithms capable of extracting discriminative information from sparse representations of signals will be needed.

In this work we present two novel methods called "most discriminative activation selection" (MDAS) and "most discriminative column selection" (MDCS) based on sparse representations of SaO_2 signals. A preliminary related approach of this work has been reported in [17]. The methods MDAS and MDCS involve finding an optimal subset of most discriminative atoms and the corresponding configuration of a multilayer perceptron (MLP) neural network classifier for detecting apnea-hypopnea events from sparse representations of segments of SaO_2 signals. The apnea-hypopnea events were appropriately labeled by medical experts, who have been carefully analyzed the complete PSG. Our methods allow for a significant reduction in the dimension of the inputs to the MLP neural network, preserving the most important characteristics of the SaO_2 signal.

This article is organized as follows: in Section 2 the materials and methods used for obtaining sparse representations of SaO_2 signals are explained. In Section 3 the results are described and the discussion is finally included in Section 4.

2. Materials and Methods

A sparse representation problem can be divided into two separate subproblems: a *learning* problem and an *inference* problem. The first one, which is quite often more complex, consists of finding an "optimal" dictionary Φ to represent a given set of signals $\{\mathbf{x}_i\}$. A dictionary Φ is called complete (CD) or overcomplete (OD) depending on the number of basic waveforms be equal or greater, respectively than the signal's space dimension. The second problem consists of selecting a set of representation vectors $\{\mathbf{a}_i\}$ satisfying a given sparsity constraint. The MDAS and MDCS methods involve finding

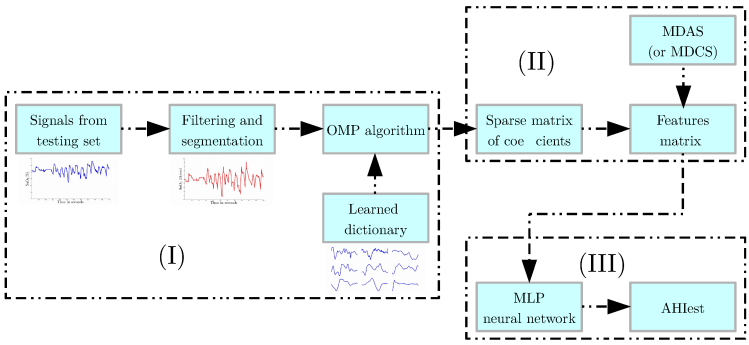


Figure 1: A simplified block diagram of the classification process.

a set of discriminative coefficients (feature vector) to be used as inputs of a MLP neural network [18]. In order to achieve this objective all possible number of inputs (F) and a large number of neurons in its hidden layer (NHL) are tested. Finally the optimal configuration is obtained by choosing the F and NHL values resulting in the best performance.

Figure 1 shows a simplified block diagram of the proposed system. In the first block (I) the signals are filtered and segmented by making use of wavelet filters [19] and segmentation techniques (as described in Subsection 2.1), respectively. The processes for obtaining sparse representations of the signals are presented by a previously learned dictionary and orthogonal matching pursuit (OMP) algorithm. The second block (II) shows the feature extraction stage by using the MDAS (or MDCS) method (see details in Subsection 2.4). In the last block (III), the estimated AHI ($\mathrm{AHI}_{\mathrm{est}}$) value is obtained by post-processing a previously trained MLP neural network output which produces the apnea-hypopnea event detection (see details in Sections 2.3 and 3).

We consider two types of dictionaries (complete and overcomplete) and three different methods (use of the FULL sparse representation, MDAS and MDCS). Thus, six methodologies emerged, which we call FULL-OD, MDAS-OD, MDCS-OD, FULL-CD, MDAS-CD, and MDCS-CD. Thus, for instance, the FULL-OD method makes use of an overcomplete dictionary Φ_{OD} and the whole representation vector $\mathbf{a}_i$ as input of the MLP neural network classifier, while the MDAS-OD method uses the dictionary Φ_{OD} and a selected set of features extracted from $\mathbf{a}_i$ by applying the MDAS method.

2.1. Filtering and segmentation

The set of biomedical signals used in this article was obtained from the sleep heart health study (SHHS) dataset [20, 21]. This dataset comprises valuable material about detailed PSGs which were properly obtained to explore correlations between sleep disorders and cardiovascular diseases. The complete dataset includes 995 studies, each of them containing several biomedical signals such as electrical activity of the brain, electrical activity of the heart, nasal airflow, SaO_2, among others. Annotations of sleep stages, arousals and apnea-hypopnea events are also added. For our work, only the SaO_2 signal and its corresponding apnea-hypopnea labels are considered.

The SaO_2 signals are usually highly degraded by patient movements, baseline wander, disconnections and the limited resolution of the pulse oximeter, among others factors. When a disconnection occurs, the values during the time interval where the sensor signal is invalid are linearly interpolated. A wavelet processing technique proposed in [19] is chosen for denoising the signals. The signals are also sampled at 1Hz and the denoising process is carried out by discarding the approximation coefficients, at level 8, as well as the first three detail coefficients of the discrete dyadic wavelet transform with mother wavelet Daubechies 2. The application of this process has the effect of a band-pass filter where the baseline wander and both the low frequency noise and the high frequency noise as well as the quantization noise are eliminated. Figure 2 shows a portion of the airflow signal (top) as well as the original raw pulse oximetry signal (middle) and the wavelet-filtered pulse oximetry signal (bottom). The corresponding labels of apnea-hypopnea events (dash lines) are also included. By observing both the airflow and the raw pulse oximetry signals, it can be seen that there is generally a causal relation between an apnea-hypopnea event and the oxygen desaturation in the pulse oximetry signal. However, the time interval between the blockage of nasal airflow and the start of the oxygen desaturation is highly variable. Although, as previously mentioned an apnea-hypopnea event is not always accompanied with "noticeable" oxygen desaturations (which are used by medical experts to detect and label the apnea-hypopnea events), artificial intelligent algorithms can detect slight changes in the pulse oximetry signal. Note that the time duration of each desaturation, which is associated to an apnea-hypopnea event, is also variable. Figure 2 also shows the effect of the wavelet-filter in avoiding "disconnections" in the pulse oximetry signal. In what follows, by the "SaO_2 signal", we will always mean the denoised one.

In order to apply the sparse representation technique, an appropriate

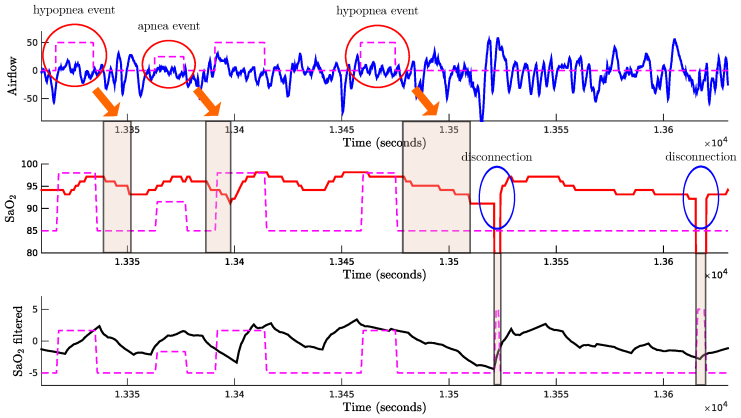


Figure 2: A portion of airflow and pulse oximetry signals. Original raw airflow and pulse oximetry signals (top and middle) and its wavelet-filtered version (bottom). Dashed lines represent labels of apnea-hypopnea events introduced by the medical expert.

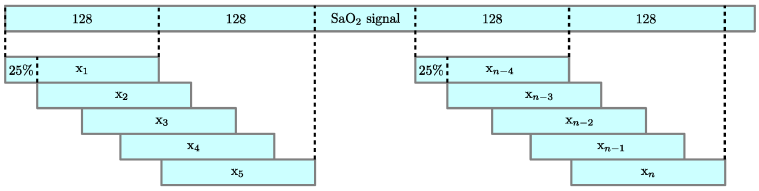


Figure 3: Schematic representation of SaO_2 signal segmentation.

segmentation of the signals is required. For this reason, segments of length $N = 128$ (corresponding to 128 seconds) with a 75% overlapping between two consecutive segments are taken. In this process, the time intervals where a disconnection occurs are not taken into account. The segmentation process is depicted in Figure 3. The segments of pulse oximetry signals are simultaneously arranged as column vectors $\mathbf{x}_i \in \mathbb{R}^N$ and labeled with ones and minus ones, where a one is associated to an apnea-hypopnea event, and a minus one to the lack of it, respectively. Finally a signal matrix X is built by stacking side-by-side the column vectors $\mathbf{x}_i$, i.e. the signal matrix is defined as $\mathrm{X} \doteq [\mathbf{x}_1\ \mathbf{x}_2\ \mathbf{x}_3\ \cdots\ \mathbf{x}_n]$, where n represents the total number of segments.

2.2. Sparse representations

The problem of obtaining the sparse representation of a signal $\mathbf{x}_i$ in terms of a given overcomplete dictionary Φ can be described as follows: Given both a matrix $\Phi \in \mathbb{R}^{N \times M}$ (with $M \geq N$) formed by M columns ϕ_j (called atoms

of the dictionary) and a signal $\mathbf{x}_i \in \mathbb{R}^N$, the sparse representation problem can be written as $\mathbf{x}_i = \Phi \mathbf{a}_{\mathrm{SR}(i)}$; where

$$\mathbf{a}_{\mathrm{SR}(i)} = \underset{\mathbf{a}_i}{\operatorname{argmin}} ||\mathbf{a}_i||_0 \text{ subject to } \Phi \mathbf{a}_i = \mathbf{x}_i, \tag{1}$$

where the operator $||\cdot||_0$ denotes the zero-norm.

The term "basis" is often replaced by "dictionary" because the atom-by-atom linear independence is not necessary needed, and many times the number of atoms is greater than the dimension of the signals. In that case, i.e. $M > N$, or more generally when the atoms do not form a basis, then the representation of a given signal may not be unique and therefore a good enough constraint is required to choice only one of them. In our case, sparsity (a criterion for selecting a representation using the least number of atoms) is used, although many other available criteria can be taken into account.

By considering the representation given by $\mathbf{x}_i = \Phi \mathbf{a}_i$. It is important to point out that although the synthesis of the signals is linear, the opposite operation (obtain $\mathbf{a}_i$ in terms of $\mathbf{x}_i$ and Φ) can be non-linear.

In practical applications not just one but a given set of signals is normally obtained. In this case the problem of sparse representation of such signals becomes very difficult because the build up of the dictionary is part of the problem. Naturally the dictionary could be constructed by staking side-by-side the whole signals. Although the sparse representation problem will be optimal, this kind of solution is highly undesired because of its huge size and long redundancy. Thus it is very appropriate to use a method which learn an optimal dictionary, in certain sense, from de signals in the given dataset. To achieve this objective a statistical approach called noise overcomplete ICA (NOCICA) [13, 22, 23] was taken. Equations (2) and (3) describe iterative rules for updating both the dictionary Φ and the representation vector $\mathbf{a}$ by means of this method:

$$\Delta\Phi = \eta \Lambda_\epsilon ((\mathbf{x} - \Phi \mathbf{a}_{\mathrm{MAP}}) \mathbf{a}_{\mathrm{MAP}}^T - \Phi H^{-1}), \tag{2}$$

where $\eta \in (0, 1)$ is the so called "learning coefficient", Λ_ϵ is the noise covariance matrix, $\mathbf{a}_{\mathrm{MAP}}$ is the maximum-a-posteriori (MAP) estimator of $\mathbf{a}$ and H is minus the Hessian of the log-posterior evaluated at $\mathbf{a}_{\mathrm{MAP}}$, and

$$\Delta\mathbf{a} = \Phi^T \Lambda_\epsilon (\mathbf{x} - \Phi \mathbf{a}) - \boldsymbol{\rho}^T |\mathbf{a}|, \tag{3}$$

where $\boldsymbol{\rho} = (\rho_1 \ \rho_2 \ \cdots \ \rho_n)^T$ corresponds to a proposed a Laplacian a-priori distribution $\pi(a_j) \propto \exp(\rho_j |a_j|)$ and $\rho_j < 0$.

2.3. MLP neural network

The MLP is a special type of neural networks which consist of input units (input layer), at least one hidden layer and an output layer [18]. Both the hidden an the output layers are composed of computation units (neurons). The inputs, sometimes called feature vector, are processed layer-by-layer moving forward through the network. The output of a neuron is given by the application of an activation function (linear or non-linear) to the weighted sum of the inputs plus a bias term. In general the output of a neuron y_j is given by Equation (4).

$$y_j = f(\sum_{i=1}^{d} \omega_{ji} x_i + \omega_{j0}) = f(\sum_{i=0}^{d} \omega_{ji} x_i), \tag{4}$$

where the activation function (sometimes called transfer function) is denoted by $f(\cdot)$, and the weights connecting the i-th input to the j-th neuron for a given layer is represented by ω_{ji}.

2.4. Detection of discriminative atoms

As already explained, the problem of sparse representations of a signal consist essentially in approximating such a signal by a linear combination of only a few atoms in a given dictionary. In applications whose final objective is signal classification we are not much interested in the accuracy of such a representation but rather in its discriminative power, that is in its ability to distinguish between the different classes. With this in mind, in this work we introduce an atom selection process by means of discriminative information. Roughly speaking, when an atom has a high activation frequency for one of the classes (but not for the others), then this atom is classified as containing significant "discriminative" information. The MDAS and MDCS methods are explained below.

The MDAS method: let Φ be a given dictionary, $\mathrm{X_{train}}$ and $\mathrm{X_{val}}$ training and validation signal matrices, respectively (built as explained in Subsection 2.1), $\mathrm{T_{train}}$ and $\mathrm{T_{val}}$ training and validation target vectors, respectively, and p_0 the sparsity level. We describe now the building steps of the MDAS method together with the corresponding lines in its implementation algorithm (Algorithm 1). First, each representation vector $\mathbf{a}_{\mathrm{SR(i)}}$ is obtained by applying a greedy pursuit algorithm called OMP [24] (line 2). Then a coefficient matrix A is assembled by stacking side-by-side the vectors $\mathbf{a}_{\mathrm{SR(i)}}$ (line 3). After that, the atom activation frequencies η_κ^j are obtained for each one of the atoms $\boldsymbol{\phi}_j$ and each one of the classes $\kappa = 1$ and $\kappa = 2$ (line

5). Here, η_κ^j represents the number of times that the atom $\boldsymbol{\phi}_j$ was used to represent segments belonging to the class κ and τ_κ represents the column indices corresponding to class κ. The proposed discriminative approach begins by computing the absolute difference between the activation frequencies, i.e. $\mathrm{D}(j) = |\eta_1^j - \eta_2^j|$ (line 6). Clearly $\mathrm{D}(j)$ will be large if the j^{th}-atom is much more active in one class than in the other. Otherwise, if the j^{th}-atom has similar activation frequencies in both classes then $\mathrm{D}(j)$ is close to zero. After that the vector D is redefined by rearranging its elements in decreasing order and saving the corresponding vector of indices Ind (lines 8 and 9). Next the MLP neural network is trained by varying the feature vector size and the number of neurons located in the hidden layer (lines 10 to 19). The features taken as input of the MLP neural network are those corresponding to the most discriminative atoms of Φ according to D ($\mathrm{F}_{\mathrm{MDAS}}$ for training and $\mathrm{F}_{\mathrm{val}}$ for validation). Once the MLP neural network training stage is finished, an optimal configuration of the MLP neural network is obtained (line 20). An schematic representation of the coefficient selection process is depicted in Figure 4. Figure 5 shows, in decreasing order, the absolute difference of activation frequencies of the atoms corresponding to a dictionary which was learned using segments of signals belonging to class 1. By observing this figure it is reasonable to conclude that a large percentage of the discriminative information can be captured by the first 40 or 50 atoms. Figure 6 shows the waveforms of some atoms in three different regions of the curve shown in Figure 5. In particular the first row in Figure 6 shows the waveforms of the first three most discriminative atoms while rows 2 and 3 present the waveforms corresponding to atoms in the middle and low discrimination ranges, respectively. It is very interesting to see that the three first most discriminative atoms present waveforms which are clearly associated with desaturations in the SaO_2 signals.

The MDCS method: this method (whose implementation is described by Algorithm 2) is similar to the previous one except for the stage 2 that we describe next. Once the vector D is rearranged, a new sub-dictionary Φ_{new} is built (line 4) and consequently the feature vector $\mathbf{f}_i$ is obtained by applying the OMP algorithm (line 5). Finally each feature vector $\mathbf{f}_i$ is assigned to be the input of the MLP neural network (line 7).

At the training stage most of the computational cost (about 80%) is due to dictionary learning. The remaining cost corresponds to the inference of the coefficients and the MLP neural network training. At the testing stage the computational cost is significantly reduced (at about 30% of the training

Algorithm 1 MDAS algorithm

procedure MDAS($\Phi, \mathrm{X}_{\mathrm{train}}, \mathrm{X}_{\mathrm{val}}, \mathrm{T}_{\mathrm{train}}, \mathrm{T}_{\mathrm{val}}, p_0$)
stage 1:
  $\mathbf{a}_{\mathrm{SR(i)}} \leftarrow \underset{\|\mathbf{a}_i\|}{\operatorname{argmin}} \|\mathbf{x}_i - \Phi\mathbf{a}_i\|_2^2$
    subject to $\|\mathbf{a}_i\|_0 \leq p_0,\ \forall \mathbf{x}_i \in \mathrm{X}_{\mathrm{train}}$
  $\mathrm{A} \leftarrow [\mathbf{a}_{\mathrm{SR}(1)}\ \mathbf{a}_{\mathrm{SR}(2)}\ \mathbf{a}_{\mathrm{SR}(3)}\ \cdots\ \mathbf{a}_{\mathrm{SR}(n)}]$
  for $j \leftarrow 1, M$ **do**
    $\eta_\kappa^j \leftarrow \|\mathrm{A}(j, \tau_\kappa)\|_0$
    $\mathrm{D}(j) \leftarrow |\eta_1^j - \eta_2^j|$
  end for
  $\mathrm{D} \leftarrow [\mathrm{d}_{\gamma(1)}\ \mathrm{d}_{\gamma(2)}\ \mathrm{d}_{\gamma(3)}\ \cdots\ \mathrm{d}_{\gamma(M)}]$
  $\mathrm{Ind} \leftarrow [\gamma(1)\ \gamma(2)\ \gamma(3)\ \cdots\ \gamma(M)]$
stage 2:
  for $m \leftarrow 1, M$ **do**
    $\mathrm{Ind}_{\mathrm{new}} \leftarrow [\mathrm{Ind}(1)\ \cdots\ \mathrm{Ind}(m)]$
    for $h \leftarrow 1, N$ **do**
      $\mathbf{f}_i \leftarrow \mathbf{a}_{\mathrm{SR(i)}}(\mathrm{Ind}_{\mathrm{new}})$
      $\mathrm{F}_{\mathrm{MDAS}} \leftarrow [\mathbf{f}_1\ \mathbf{f}_2\ \mathbf{f}_3\ \cdots\ \mathbf{f}_n]$
      $\mathrm{NHL} \leftarrow h$
      net $\leftarrow$TRAIN($\mathrm{F}_{\mathrm{MDAS}}, \mathrm{T}_{\mathrm{train}}, \mathrm{NHL}$)
      $\mathrm{PM}(n, m) \leftarrow$ VALID(net, $\mathrm{F}_{\mathrm{val}}, \mathrm{T}_{\mathrm{val}}$)
    end for
  end for
stage 3:
  $[\mathrm{F}_{\mathrm{op}}, \mathrm{NHL}_{\mathrm{op}}] \leftarrow \underset{\mathrm{F,NHL}}{\operatorname{argmax}}\ \mathrm{PM}$
  return $\mathrm{F}_{\mathrm{op}}, \mathrm{NHL}_{\mathrm{op}}$
end procedure

Algorithm 2 MDCS algorithm

procedure MDCS($\Phi, \mathrm{X}_{\mathrm{train}}, \mathrm{X}_{\mathrm{val}}, \mathrm{T}_{\mathrm{train}}, \mathrm{T}_{\mathrm{val}}, p_0$)
stage 1: same as MDAS algorithm
stage 2:
  for $m \leftarrow 1, M$ **do**
    $\mathrm{Ind}_{\mathrm{new}} \leftarrow [\mathrm{Ind}(1) \ \cdots \ \mathrm{Ind}(m)]$
    $\Phi_{\mathrm{new}} \leftarrow \Phi(:, \mathrm{Ind}_{\mathrm{new}})$
    $\mathbf{f}_i \leftarrow \underset{\|\mathbf{a}_i\|}{\operatorname{argmin}} \|\mathbf{x}_i - \Phi_{\mathrm{new}}\mathbf{a}_i\|_2^2$
      subject to $\|\mathbf{a}_i\|_0 \leq p_0$
    for $h \leftarrow 1, N$ **do**
      $\mathrm{F}_{\mathrm{MDCS}} \leftarrow [\mathbf{f}_1 \ \mathbf{f}_2 \ \mathbf{f}_3 \ \cdots \ \mathbf{f}_n]$
      $\mathrm{NHL} \leftarrow h$
      $\mathrm{net} \leftarrow \textsc{Train}(\mathrm{F}_{\mathrm{MDCS}}, \mathrm{T}_{\mathrm{train}}, \mathrm{NHL})$
      $\mathrm{PM}(n, m) \leftarrow \textsc{Test}(\mathrm{net}, \mathrm{F}_{\mathrm{val}}, \mathrm{T}_{\mathrm{val}})$
    end for
  end for
stage 3: same as MDAS algorithm
end procedure

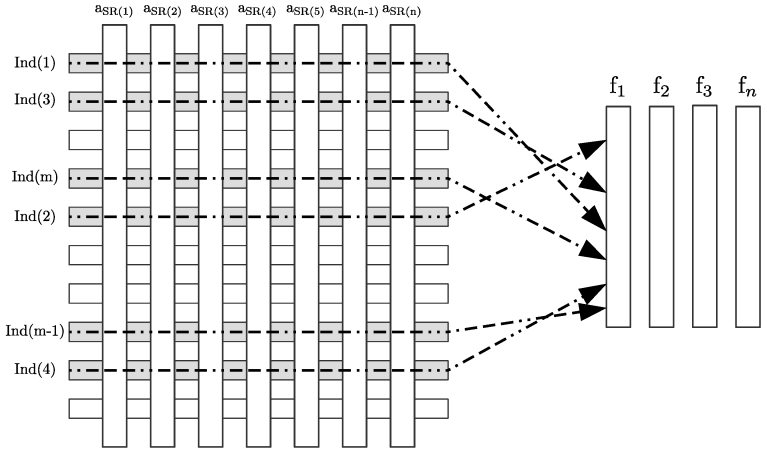


Figure 4: Schematic representation of the coefficient selection process. Here $\mathbf{f}_i$ is a vector whose components are the features extracted from $\mathbf{a}_{\mathrm{SR(i)}}$.

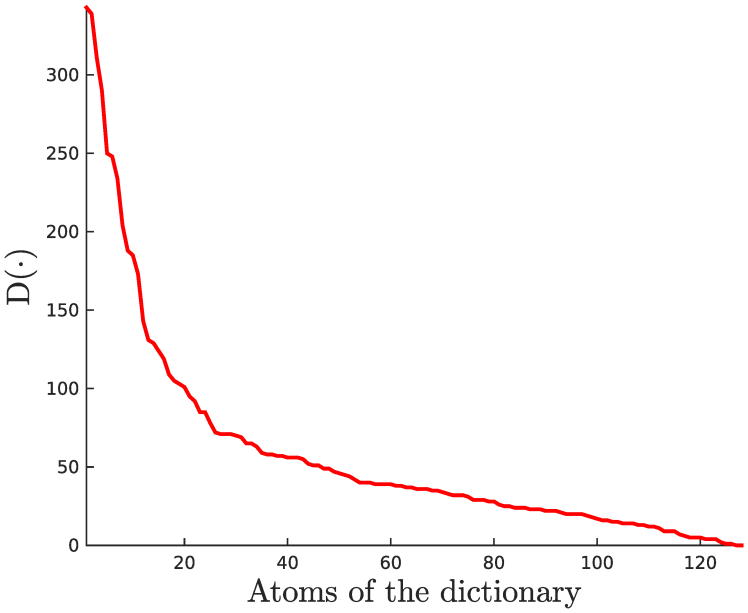


Figure 5: Absolute difference of activation frequency D(·) of the atoms of a dictionary learned with segments of signals belonging to class 1, in decreasing order of magnitude.

cost). The experiments were run on a PC with a 3.5 GHz, 6 cores AMD FX-6300 processor and 8 GB of RAM.

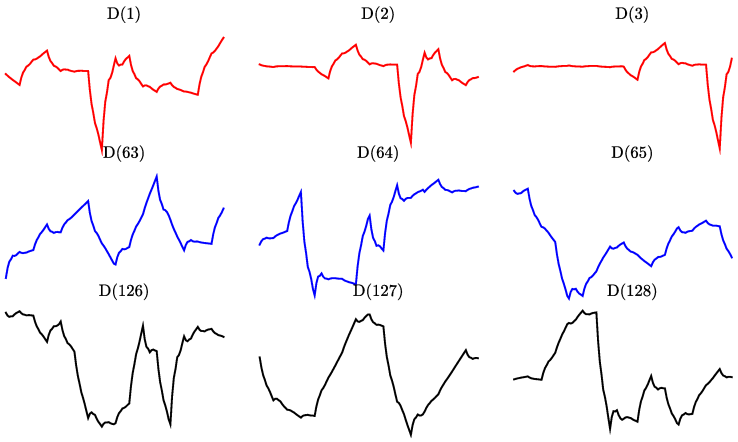


Figure 6: Examples of some atoms of a dictionary learned with segments of signals belonging to class 1 from three different regions of the curve of absolute difference of activation frequency (Figure 5): most discriminative atoms (top), medium discriminative atoms (middle row) and lowest discriminative atoms (bottom row).

3. Results

As mentioned in Subsection 2.1, the complete dataset contains 995 studies, 41 of which were discarded due to incomplete information. Among the remaining 954 studies, a subset of 667 (70%) studies were randomly selected and fixed in order to learn the dictionary and train the MLP neural network. The final test was made using the remaining 287 (30%) studies of the database. The SaO_2 signals were filtered and segmented (see details in Subsection 2.1) into vectors of length 128 (this window size corresponds to 128 seconds of the recording). A matrix $\mathrm{X}_{\mathrm{train}}$ of size 128×455515 was built as $\mathrm{X}_{\mathrm{train}} \doteq [\mathrm{X}^{\mathrm{c1}}_{\mathrm{train}}\ \mathrm{X}^{\mathrm{c2}}_{\mathrm{train}}]$, where the matrices $\mathrm{X}^{\mathrm{c1}}_{\mathrm{train}}$ of size 128×183163 and $\mathrm{X}^{\mathrm{c2}}_{\mathrm{train}}$ of size 128×272352 were constructed considering segments belonging to class 1 and class 2, respectively. Another matrix $\mathrm{X}_{\mathrm{test}}$ was constructed stacking side-by-side all vectors $\mathbf{x}_i$ corresponding to each signal from the testing set.

At the dictionary learning stage, two types of dictionaries were learned using both the $\mathrm{X}^{\mathrm{c1}}_{\mathrm{train}}$ and the $\mathrm{X}^{\mathrm{c2}}_{\mathrm{train}}$ signal matrices. First a complete dictionary Φ_{CD} of size 128×128 was learned using the matrix $\mathrm{X}_{\mathrm{train}}$, without taking into consideration any information about the classes. Second, an overcomplete dictionary Φ_{OD} of size 128×256 was assembled by stacking side-by-side the atoms of two previously learned 128×128 dictionaries Φ^{c1} and Φ^{c2}, which

were learned by using the matrices X^{c1}_{train} and X^{c2}_{train}, respectively. At the dictionary learning stage the atoms were initially taken by random selection from the corresponding signal matrix. The NOCICA method [23] was used for the dictionary learning stage.

The representation coefficients $\mathbf{a}_{SR(i)}$ were obtained by applying the OMP algorithm [25]. The reason for having chosen this greedy algorithm is because it guarantees convergence to the projection of $\mathbf{x}_i$ into the span of the dictionary atoms, in no more than p_0 iterations.

Since our problem involved a big and redundant dataset (big data problem), a variation of the back-propagation algorithm, called mini-batch training procedure, was used to train the MLP neural network. In order to avoid overfitting and estimate the neural network hyper-parameters, a large number of trials with different hyper-parameter values were performed. In what follows, the final choice of the neural network hyper-parameters are described. Batches of 1000 balanced segments were randomly selected from the 455515 available training segments. To avoid overtraining, the number of steps in the scaled conjugate gradient algorithm was set to 4. In addition, to minimize classification bias, the above training scheme was repeated 455 times with re-sampling.

In the proposed algorithms, two parameters need to be empirically determined: the sparsity level p_0 and the threshold of the outputs of the MLP neural network. To determine an adequate sparsity level, several trials were performed. It was found that a percentage value of 12.5 of the signal's space dimension presented the best trade-off between representativity and discriminability of the segments. Hence, sparsity level $p_0 = 16$ was chosen. On the other hand, to establish an optimal threshold of the MLP neural network outputs, different values in the interval $[-0.2, 0.2]$ were tested. A value of zero of the MLP neural network outputs was chosen. Hence an output value greater than 0 was considered as containing an apnea-hypopnea event, and considered to be normal otherwise. Finally the AHI_{est} value was determined as the number of detected events divided by the record length of each study (in seconds).

In Table 1, the columns labeled "F" and "NHL" show the number of inputs (feature vector size) and the number of neurons in the hidden layer of the MLP neural network, respectively. Clearly the application of the MDAS (or MDCS) method produces a significant dimension reduction and therefore, the computing time required for classification is also significant reduced. Thus, for instance, the MDAS-OD method used only 32 features

Table 1: MLP neural network's hyper-parameters. Feature vector size and number of neurons in the hidden layer.

Dictionary	Method	F	NHL
	FULL	256	32
OD	MDAS	32	16
	MDCS	64	32
	FULL	128	32
CD	MDAS	64	32
	MDCS	64	32

(12.5% of the total) compared with the FULL-OD method, which used 256 features.

For analyzing the capability of the proposed classifier in the detection of patients suspected of suffering from OSAH, two measures were introduced. The sensitivity (SE), defined as the ratio of persons with OSAH for whom the trial process is positive, and the specificity (SP), defined as the ratio of patients without OSAH for whom the trial process is negative. Also a receiver operating characteristics (ROC) [26] analysis allows to obtain the following values: true positive (TP), true negative (TN), false positive (FP), false negative (FN), cut-off point (cut-off), and area under the curve (AUC).

The objective of our experiment was to compare the performances of our methods with those of other local approaches used for OSAH detection. In particular, we compared our methods with those introduced by Chiner *et al.* [6] and Vázquez *et al.* [7], and with that presented by Schlotthauer *et al.* [10]. Tables 2, 3, and 4 show the AUC values as well as SE, SP, and accuracy (ACC) measures for AHI diagnostic threshold values of 10 and 15 for the reference.

Table 2 shows the results obtained with the use of sparse representations by means of overcomplete dictionaries. We observed a significant increment in the AUC and SE values obtained with the use of the MDCS (MDCS-OD) method. It can also be seen that the application of the MDAS (MDAS-OD) method does not produce significant changes in the AUC, SE, and SP values. Hence, the best performance of the classifier for the case of overcomplete

Table 2: Performance measures for OSAH detection using an overcomplete dictionary.

Method	AHI_{thr}	AUC	SE(%)	SP(%)	ACC(%)
FULL-OD	10	0.896	**88.37**	75.86	**82.12**
	15	0.923	83.33	**87.32**	85.33
MDAS-OD	10	0.847	86.05	72.41	79.23
	15	0.891	81.02	83.10	82.06
MDCS-OD	10	**0.906**	81.40	**79.31**	80.35
	15	**0.937**	**85.65**	85.92	**85.78**

dictionaries is obtained with the MDCS (MDCS-OD) method.

Table 3 shows the results obtained with the use of sparse representations by means of complete dictionaries. Although the MDAS method produces slight improvements in the AUC, SE, SP, and ACC values as compared with the MDAS-OD method, the results are not the best. In fact, it can be seen that the application of the MDCS method results in the best AUC, SP, and ACC values. A comparison of Tables 2 and 3 allows us to conclude that the application of the MDCS method to sparse representations results in the best option for OSAH detection.

Finally Table 4 presents a comparative summary of the best results (MDCS-OD method) and of those obtained with the other three previously mentioned methods. As observed, our method outperforms all the others. For AHI threshold values of both 10 and 15, our method reaches the maximum AUC values of 0.906 and 0.937, respectively. Also for an AHI threshold value of 15, our method achieves sensitivity and specificity percentages of 85.65% and 85.92%, respectively. The optimal operating point was chosen in order to maximize both the sensitivity and specificity percentages. Figure 7 shows the ROC plots for the four methods presented in Table 4 corresponding to AHI threshold values of 10 (Figure 7a) and 15 (Figure 7b). We also tested the use of a support vector machine (SVM) classifier with a Gaussian kernel function instead of the MLP neural network classifier. No improvements in the results were observed.

Finally, a detailed account of the computational costs for the four methods

Table 3: Performance measures for OSAH detection using a complete dictionary.

Method	AHI_{thr}	AUC	SE(%)	SP(%)	ACC(%)
FULL-CD	10	**0.903**	78.68	**82.76**	80.72
	15	0.930	**85.65**	85.92	85.78
MDAS-CD	10	0.870	73.64	**82.76**	78.20
	15	0.906	**85.65**	85.92	85.78
MDCS-CD	10	0.901	**86.82**	75.86	**81.34**
	15	**0.934**	85.19	**87.32**	**86.25**

at the testing stage is presented in Table 5. It can be observed that although our method needs more than twice of the CPU time required for the other three methods, 2.85 seconds for analyzing the data corresponding to study of ten hours of duration is insignificant, even more so taking into account the improvements in OSAH's detection reached by our method, as it can be observed in Table 4.

4. Discussion

OSAH is a highly prevalent syndrome in the general human population. From a sample of 602 workers, with ages between 30 and 60, Young *et al.* [27] found that 24% of men and 9% of women had an AHI value above 5. Durán *et al.* [28] also found that aging, being male, snoring and obesity are all factors increasing the risk of suffering from OSAH. Given this high prevalence of OSAH, primary attention medicine is determinant in the identification of patients suffering from it and therefore simple and cheap diagnostic tools are highly important. An additional valuable aspect of our work is the fact that we were able to establish a relationship between the final feature vectors and the apnea-hypopnea events. This relationship can be seen in Figure 8. On the upper right of this figure a portion of the wavelet-filtered SaO_2 signal with the marks of apnea-hypopnea events labeled by the medical expert is shown. Immediately below a curve (in green) representing the cumulative absolute activation of the sixteen most discriminative coefficients and the la-

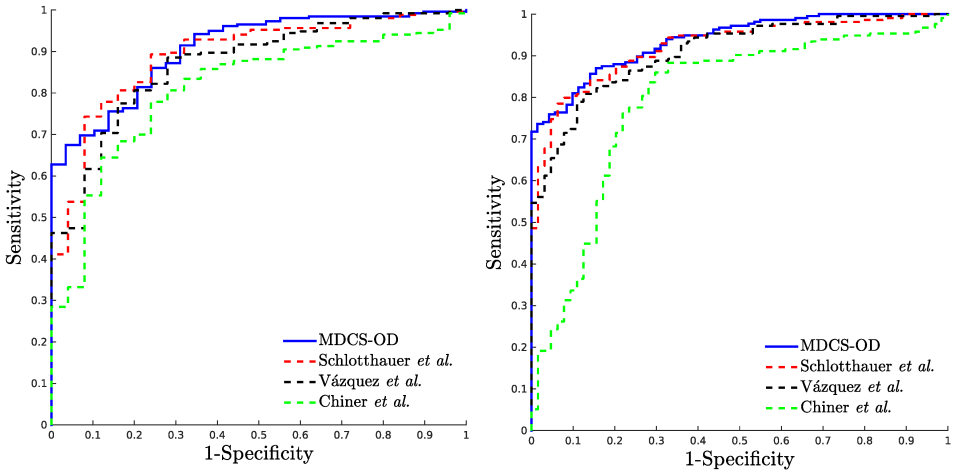
Table 4: Performance measures for OSAH detection using different methods.

Method	AHI_{thr}	AUC	SE(%)	SP(%)	ACC(%)
MDCS-OD	10	**0.906**	**81.40**	79.31	80.35
	15	**0.937**	**85.65**	85.92	**85.78**
Chiner *et al.* [6]	10	0.810	77.87	76.00	76.93
	15	0.795	76.17	78.12	77.15
Vázquez *et al.* [7]	10	0.870	77.47	**84.00**	80.74
	15	0.909	80.84	**87.50**	84.17
Schlotthauer *et al.* [10]	10	0.890	80.63	**84.00**	**82.32**
	15	0.922	84.11	85.94	85.02

Table 5: Computational cost: average CPU time for each study.

Method	Time (seconds)
MDCS-OD	2.85
Chiner *et al.* [6]	0.81
Vázquez *et al.* [7]	1.21
Schlotthauer *et al.* [10]	1.35

bels of apnea-hypopnea events (in red) are presented. The image appearing on the lower right part of Figure 8 shows the absolute value of the sixteen most discriminative coefficients of our method. A high correlation between the tags labeled by the medical experts and the most discriminative coefficients can be clearly observed. On the other hand, on the upper left corner of Figure 8 a segment of 128 seconds of the wavelet-filtered SaO_2 signal with the corresponding marks of apnea-hypopnea events is shown, while immediately below the three most discriminative atoms (ϕ_1, ϕ_8, and ϕ_{13}, respectively) involved in its representation are shown. It can be clearly seen how these three most discriminative atoms assemble together to capture the main features of



(a) $\mathrm{AHI}_{\mathrm{thr}} = 10$. (b) $\mathrm{AHI}_{\mathrm{thr}} = 15$.

Figure 7: ROC plots for the methods described in Table 4 for two different AHI threshold values.

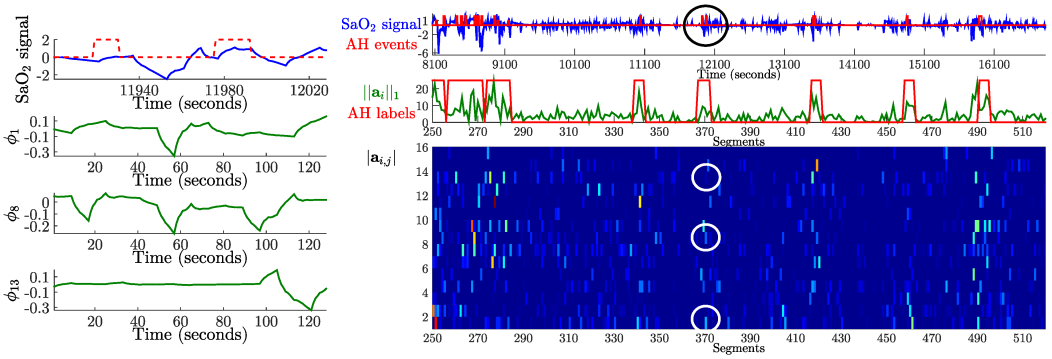


Figure 8: Final feature vectors to apnea-hypopnea events correlation.

the waveform of the filtered SaO_2 signal.

An adequate use of simplified and correctly validated systems would allow, once the cases have been selected, to decentralize the diagnosis of the reference units which are usually saturated. This decentralization would favor the creation of new smaller diagnostic units equipped with oximeters. This decentralization of the diagnostic process will have to be accompanied by appropriate training of the personnel as well as of good coordination with the reference sleep units for a deeper study of the difficult or doubtful cases

[29]. Networks of increasing complexity will have to be created in order to allow immediate consultation with a sleep medicine expert and the possibility of performing, whenever necessary, a polisomnography for the diagnostic and treatment of this real public health problem which is OSAH [29]. The design of diagnostic tools and equipment which could be handled by non-expert personnel for detecting patients with severe OSAH is a priority, since an early identification will allow immediate access to a correct treatment.

Apnea-hypopnea events during sleeping occur as a consequence of a functional-anatomic disturbance of the upper airway producing its collapse. At the end of each apnea-hypopnea event, a desaturation of the hemoglobin usually occurs. This desaturation originates a characteristic pattern in the pulse oximetry record corresponding to intermittent hypoxemia. The intermittent hypoxemia, with hypoxemia-reoxygenation cycles, promotes oxidative stress and angiogenesis, increases the sympathetic activation with increment of blood pressure and systemic and vascular inflammation with endothelial dysfunction which contributes to multi-organic chronic morbility, metabolic dysfunction, cognitive impairment and cancer progression [30].

Due to the intermittent hypoxemia in the cells (hypoxemia-reoxygenation cycles) which induce angiogenesis and tumor growth, a strong correlation between neoplastic diseases and OSAH has been described [31]. On the other hand, a recent study among male mice suggests that the intermittent hypoxia associated with OSAH could induce fertility reduction [32].

In this work we presented two novel methods which allow for the detection of apnea-hypopnea events using only the SaO_2 signals. These methods were successfully applied to signals coming from the polysomnography records in the study database [20, 21]. As it can be observed in Section 3, the application of the FULL, MDAS, and MDCS strategies, both to complete and overcomplete dictionaries, resulted in six different methods. Tables 2 and 3 show the results of each one of the six methods for two different AHI threshold values ($AHI_{thr} = 10$ and $AHI_{thr} = 15$). These threshold values were strategically chosen in order to be able to analyze the performance of each method, independently of the severity of the OSAH (or the AHI value) that one wishes to detect. Although usually an AHI threshold value of 5 is used as the lower limit for detecting mild cases of OSAH, in our case, a reliable ROC analysis for that threshold value could not be made. The main reason for that is the fact our database is highly unbalanced, containing only 16 studies with AHI values below 5. Our random selection of studies resulted in only three of them being considered for testing purposes. A statistically significant

correlation between OSAH's severity and comorbilities, such as hypertension, diabetes, dyslipidemia and metabolic syndrome, has been found in previous works. Although this correlation is found in mild OSAH, it increases consideulrably with the OSAH's degree, reaching its highest value with severe OSAH. Hence if the objective is OSAH treatment and the prevention of the associated comorbilities, an AHI threshold value of 15 is clearly pathological [33]. There is evidence that close to 93% of women and 82% of men with moderate to severe OSAH remain undiagnosed [34]. Since sleep fragmentation, intermittent hypoxemia, increased sympathetic tone and hypertension are main causes of mortality and morbidity, it is highly desirable to have everyone with moderate to severe OSAH appropriately diagnosed. Although the gold standard for diagnosing sleep disorders is the complete PSG, this diagnosing procedure presents many limitations, such as: limited resources, limited number of recording beds, high costs, long waiting lists, and high labor requirements, among others. It is for those reasons that there is a lot of interest in exploring the possibility of using screening devices together with automated algorithms as alternative methods for diagnosing OSAH. Mild cases can be analyzed by standard methods. The ideal screening device should be cheap and easy to be used with minimal risks to the patient.

By considering an $\mathrm{AHI}_{\mathrm{thr}} = 15$, a detailed analysis of Tables 2 and 3 show that, although most methods have good performances, MDCS-OD outperforms all the others. The application of this method results in an area under the ROC curve of 0.937 and sensitivity and specificity percentages of 85.65 and 85.92, respectively. Taking into account that out of the 287 records in the testing database, 216 had an AHI value above 15, and the remaining 71 were below that threshold value, a 85.65% sensitivity indicates that of the 216 cases with AHI values above 15, 185 were appropriately identified while 31 were erroneously detected. On the other hand, an 85.92 specificity indicates that of the 71 cases with AHI values below 15, 61 were appropriately identified while only 10 were erroneously detected. It is timely to point out here that for the 10 cases that the MDCS-OD method yielded an AHI value higher than 15, the registry database indicated an AHI average value of 10.62 with a standard deviation of 3.88. By analyzing each one of these studies in detail, it was observed that most of the respiratory events informed by the medical expert were hypopneas and not all of them were related to SaO_2 desaturations. This fact indicates that the medical experts have not taken into account the AASM criteria.

The MDCS-OD method was compared with those proposed by Chiner *et*

al. [6], Vázquez *et al.* [7], and Schlotthauer *et al.* [10]. These four methods were successfully applied to pulse oximetry signals included in the study database [20, 21]. Table 4 shows a detailed comparison of the performances of such methods. The results clearly show that the MDCS-OD method is a very attractive tool to assist physicians in the detection of patients whose AHI values are above the objective threshold $\mathrm{AHI}_{\mathrm{thr}} = 15$. Thus, the sparse representation of pulse oximetry signals is undoubtedly a promising technique for the design of new methods for OSAH detection.

Since there exist applications where a particular value of sensitivity or specificity is highly desirable, other operation points in the ROC curves (Figure 7) can be chosen. If the primary purpose of the test is "screening", i.e. detection of early disease in large numbers of apparently healthy individuals, then a high sensitivity is generally chosen. With this in mind, if a sensitivity of 98% is chosen in the ROC curves in Figure 7a, our method achieves a specificity of 44.83%, followed by Schlotthauer's *et al.* which reaches 28.00%. For an operating point of 98% sensitivity in the ROC curves in Figure 7b, our method achieves a specificity of 46.48%, followed by Schlotthauer's *et al.* which reaches 34.37%. On the other hand, if the objective test is "diagnostic", i.e. to establish the presence (or absence) of disease, then a high specificity is usually selected. Thus, if a specificity of 100% is chosen in the ROC curves in Figure 7a, our method achieves a sensitivity of 62.79%, followed by Vázquez's *et al.* which reaches 46.25%. For an operating point of 100% sensitivity in the ROC curves in Figure 7b, our method achieves a sensitivity of 71.76%, followed by Vázquez's *et al.* which reaches 54.67%.

There are several technical and physiological limitations associated with pulse oximetry which hinder the acquisition of a "good" signal in some cases. This is so, for instance in the following cases: weak contact between the probe and the finger due to body motions, anemia, use of nail polish, use of artificial nails, skin pigmentation, onychomycosis, cold fingers and low perfusion of vascular bed [35, 36]. Even so, pulse oximetry has shown its effectiveness in clinical practice and therefore an alert and well informed clinical physician must be aware of both its proper use and limitations.

5. Conclusions

It has been shown that the sparse representations of pulse oximetry signals is a tool which allows a very good performance for estimating AHI values above 15. The previous results have been shown that there is a high corre-

lation between the AHI observed by the medical physician via PSG and the AHI_{est} obtained by using sparse representations of pulse oximetry signals. This fact constitutes a strong evidence that such a procedure could be helpful in the detection of individuals suspected of suffering from OSAH, which require a complete PSG study for their correct diagnosis. The MDCS-OD algorithm could be embedded into the oximeter so as to be used by primary attention clinical physicians in the search and detection of patients with moderate OSAH.

Acknowledgment

The authors would like to acknowledge the financial support of Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, through projects PIP 2014-2016 Nro. 11220130100 216-CO and PIP 2012-2014 Nro. 11420110100284-KA4, of the Air Force Office of Scientific Research, AFOSR /SOARD, through Grant FA9550-14-1-0130 and of the Universidad Nacional del Litoral through projects CAI+D 50120110100519 and CAI+D 5012011010 0525.

6. References

[1] M. J. Sateia, International classification of sleep disorders-third edition: Highlights and modifications, Chest 146 (2014) 1387–1394.

[2] R. Thurnheer, K. E. Bloch, I. Laube, M. Gugger, M. Heitz, Swiss Respiratory Polygraphy Registry, Respiratory polygraphy in sleep apnoea diagnosis. Report of the Swiss respiratory polygraphy registry and systematic review of the literature, Swiss Medical Weekly 137 (2007) 97–102.

[3] E. García-Díaz, E. Quintana-Gallego, A. Ruiz, C. Carmona-Bernal, A. Sánchez-Armengol, G. Botebol-Benhamou, F. Capote, Respiratory polygraphy with actigraphy in the diagnosis of sleep apnea-hypopnea syndrome, Chest 131 (2007) 725–732.

[4] A. Yadollahi, E. Giannouli, Z. Moussavi, Sleep apnea monitoring and diagnosis based on pulse oximetry and tracheal sound signals, Medical & Biological Engineering & Computing 48 (2010) 1087–1097.

[5] L.-W. Hang, H.-L. Wang, J.-H. Chen, J.-C. Hsu, H.-H. Lin, W.-S. Chung, Y.-F. Chen, Validation of overnight oximetry to diagnose patients with moderate to severe obstructive sleep apnea, BMC Pulmonary Medicine 15 (2015) 24.

[6] E. Chiner, J. Signes-Costa, J. M. Arriero, J. Marco, I. Fuentes, A. Sergado, Nocturnal oximetry for the diagnosis of the sleep apnoea hypopnoea syndrome: a method to reduce the number of polysomnographies?, Thorax 54 (1999) 968–971.

[7] J.-C. Vázquez, W. H. Tsai, W. W. Flemons, A. Masuda, R. Brant, E. Hajduk, W. A. Whitelaw, J. E. Remmers, Automated analysis of digital oximetry in the diagnosis of obstructive sleep apnoea, Thorax 55 (2000) 302–307.

[8] D. Alvarez-Estevez, V. Moret-Bonillo, Computer-Assisted Diagnosis of the Sleep Apnea-Hypopnea Syndrome: A Review, Sleep Disorders 2015 (2015).

[9] L. M. Sepulveda-Cano, E. Gil, P. Laguna, G. Castellanos-Dominguez, Selection of nonstationary dynamic features for obstructive sleep apnoea detection in children, EURASIP Journal on Advances in Signal Processing 11 (2011) 1–10.

[10] G. Schlotthauer, L. E. Di Persia, L. D. Larrateguy, D. H. Milone, Screening of obstructive sleep apnea with empirical mode decomposition of pulse oximetry, Medical Engineering & Physics 36 (2014) 1074–1080.

[11] A. R. Hassan, Computer-aided obstructive sleep apnea detection using normal inverse gaussian parameters and adaptive boosting, Biomedical Signal Processing and Control 29 (2016) 22–30.

[12] H. Karamanli, T. Yalcinoz, M. A. Yalcinoz, T. Yalcinoz, A prediction model based on artificial neural networks for the diagnosis of obstructive sleep apnea, Sleep and Breathing 20 (2015) 509–514.

[13] M. S. Lewicki, B. A. Olshausen, Probabilistic framework for the adaptation and comparison of image codes, Journal of the Optical Society of America A 16 (1999) 1587.

[14] M. Aharon, M. Elad, A. Bruckstein, KSVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation, IEEE Transactions on Signal Processing 54 (2006) 4311–4322.

[15] P. König, K. P. Körding, D. J. Klein, Sparse spectrotemporal coding of sounds, EURASIP Journal on Advances in Signal Processing (2003) 659–667.

[16] C. E. Martínez, J. Goddard, D. H. Milone, H. L. Rufiner, Bioinspired sparse spectro-temporal representation of speech for robust classification, Computer Speech and Language 26 (2012) 336–348.

[17] R. Rolón, L. Di Persia, H. L. Rufiner, R. Spies, Most discriminative atom selection for apnea-hypopnea events detection, in: Anales del VI Congreso Latinoamericano de Ingeniería Biomédica (CLAIB 2014), pp. 709–712.

[18] S. Haykin, Neural Networks: A Comprehensive Foundation, Prentice Hall PTR, Upper Saddle River, NJ, USA, 2nd edition, 1998.

[19] F. Lestussi, L. Di Persia, D. Milone, Comparison of on-line wavelet analysis and reconstruction: with application to ECG, 5th International Conference on Bioinformatics and Biomedical Engineering (iCBBE 2011) (2011).

[20] S. F. Quan, B. V. Howard, C. Iber, J. P. Kiley, F. J. Nieto, G. T. O'Connor, D. M. Rapoport, S. Redline, J. Robbins, J. M. Samet, P. W. Wahl, The Sleep Heart Health Study: design, rationale, and methods, Sleep 20 (1997) 1077–1085.

[21] B. K. Lind, J. L. Goodwin, J. G. Hill, T. Ali, S. Redline, S. F. Quan, Recruitment of healthy adults into a study of overnight sleep monitoring in the home: experience of the Sleep Heart Health Study, Sleep & Breathing = Schlaf & Atmung 7 (2003) 13–24.

[22] S. Abdallah, Towards music perception by redundancy reduction and unsupervised learning in probabilistic models, Ph.D. Thesis, Department of Electronic Engineering, King's College London., 2002.

[23] M. S. Lewicki, T. J. Sejnowski, Learning overcomplete representations, Neural Computation 12 (2000) 337–365.

[24] J. Tropp, A. Gilbert, Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit, IEEE Transactions on Information Theory 53 (2007) 4655–4666.

[25] Y. Pati, R. Rezaiifar, P. Krishnaprasad, Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition, in: Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers, pp. 40–44.

[26] R. Kumar, A. Indrayan, Receiver operating characteristic (ROC) curve for medical researchers, Indian Pediatrics 48 (2011) 277–287.

[27] T. Young, M. Palta, J. Dempsey, J. Skatrud, S. Weber, S. Badr, The occurrence of sleep-disordered breathing among middle-aged adults, The New England Journal of Medicine 328 (1993) 1230–1235.

[28] J. Durán, S. Esnaola, R. Rubio, A. Iztueta, Obstructive sleep apnea-hypopnea and related clinical features in a population-based sample of subjects aged 30 to 70 yr, American Journal of Respiratory and Critical Care Medicine 163 (2001) 685–689.

[29] Tratamiento médico del SAHS, Archivos de Bronconeumología 41 (2005) 43–50.

[30] N. A. Dewan, F. J. Nieto, V. K. Somers, Intermittent hypoxemia and OSA: implications for comorbidities, Chest 147 (2015) 266–274.

[31] W. Kukwa, E. Migacz, K. Druc, E. Grzesiuk, A. M. Czarnecka, Obstructive sleep apnea and cancer: effects of intermittent hypoxia?, Future Oncology (London, England) 11 (2015) 3285–3298.

[32] M. Torres, R. Laguna-Barraza, M. Dalmases, A. Calle, E. Pericuesta, J. M. Montserrat, D. Navajas, A. Gutierrez-Adan, R. Farré, Male fertility is reduced by chronic intermittent hypoxia mimicking sleep apnea in mice, Sleep 37 (2014) 1757–1765.

[33] M. Fusetti, A. B. Fioretti, M. Valenti, F. Masedu, M. Lauriello, M. Pagliarella, Cardiovascular and metabolic comorbidities in patients with obstructive sleep apnoea syndrome, Acta Otorhinolaryngologica Italica 32 (2012) 320–325.

[34] T. Young, L. Evans, L. Finn, M. Palta, Estimation of the clinically diagnosed proportion of sleep apnea syndrome in middle-aged men and women, Sleep 20 (1997) 705–706.

[35] R.-P. Eduardo Martín, Factores que afectan la oximetría de pulso, Revista Mexicana de Anestesiología 29 (2006) S193–S198.

[36] M. George, S. Ronald E., Limitations of Pulse Oximetry, Anesthesia Progress 39 (1992) 194–196.

AppendixA. Dictionary updating rule.

Proof.

$$\begin{aligned}\Delta\Phi &= \eta\Lambda_\epsilon\mathbb{E}[(\mathbf{x}-\Phi\mathbf{a})\mathbf{a}^T] \\ &= \eta\Lambda_\epsilon\mathbb{E}[(\mathbf{x}\mathbf{a}^T-\Phi\mathbf{a}\mathbf{a}^T)] \\ &= \eta\Lambda_\epsilon(\mathbf{x}\mathbb{E}[\mathbf{a}^T]-\Phi\mathbb{E}[\mathbf{a}\mathbf{a}^T]).\end{aligned}$$

But

$$\mathbb{E}[\mathbf{a}^T] = \mathbf{a}_{\text{MAP}}^T,$$

and

$$\begin{aligned}\text{cov}(\mathbf{a}) &= \mathbb{E}[(\mathbf{a}-\mathbf{a}_{\text{MAP}})(\mathbf{a}^T-\mathbf{a}_{\text{MAP}}^T)] \\ &= \mathbb{E}[\mathbf{a}\mathbf{a}^T-\mathbf{a}\mathbf{a}_{\text{MAP}}^T-\mathbf{a}_{\text{MAP}}\mathbf{a}^T+\mathbf{a}_{\text{MAP}}\mathbf{a}_{\text{MAP}}^T] \\ &= \mathbb{E}[\mathbf{a}\mathbf{a}^T]-\mathbb{E}[\mathbf{a}\mathbf{a}_{\text{MAP}}^T]-\mathbb{E}[\mathbf{a}_{\text{MAP}}\mathbf{a}^T]+\mathbb{E}[\mathbf{a}_{\text{MAP}}\mathbf{a}_{\text{MAP}}^T] \\ &= \mathbb{E}[\mathbf{a}\mathbf{a}^T]-\mathbb{E}[\mathbf{a}]\mathbf{a}_{\text{MAP}}^T-\mathbf{a}_{\text{MAP}}\mathbb{E}[\mathbf{a}^T]+\mathbf{a}_{\text{MAP}}\mathbf{a}_{\text{MAP}}^T \\ &= \mathbb{E}[\mathbf{a}\mathbf{a}^T]-\mathbf{a}_{\text{MAP}}\mathbf{a}_{\text{MAP}}^T-\mathbf{a}_{\text{MAP}}\mathbf{a}_{\text{MAP}}^T+\mathbf{a}_{\text{MAP}}\mathbf{a}_{\text{MAP}}^T \\ &= \mathbb{E}[\mathbf{a}\mathbf{a}^T]-\mathbf{a}_{\text{MAP}}\mathbf{a}_{\text{MAP}}^T \\ \mathbb{E}[\mathbf{a}\mathbf{a}^T] &= \text{cov}(\mathbf{a})+\mathbf{a}_{\text{MAP}}\mathbf{a}_{\text{MAP}}^T.\end{aligned}$$

Hence,

$$\begin{aligned}\Delta\Phi &= \eta\Lambda_\epsilon(\mathbf{x}\mathbf{a}_{\text{MAP}}^T-\Phi(\text{cov}(\mathbf{a})+\mathbf{a}_{\text{MAP}}\mathbf{a}_{\text{MAP}}^T) \\ &= \eta\Lambda_\epsilon(\mathbf{x}\mathbf{a}_{\text{MAP}}^T-\Phi(H^{-1}+\mathbf{a}_{\text{MAP}}\mathbf{a}_{\text{MAP}}^T) \\ &= \eta\Lambda_\epsilon(\mathbf{x}\mathbf{a}_{\text{MAP}}^T-\Phi H^{-1}-\Phi\mathbf{a}_{\text{MAP}}\mathbf{a}_{\text{MAP}}^T) \\ &= \eta\Lambda_\epsilon((\mathbf{x}-\Phi\mathbf{a}_{\text{MAP}})\mathbf{a}_{\text{MAP}}^T-\Phi H^{-1}).\end{aligned}$$

□

Anexo C

A method for discriminative dictionary learning with application to pattern recognition

Trabajo presentado en el VI congreso de Matemática Aplicada, Computacional e Industrial, páginas 1-4, 2017.

A METHOD FOR DISCRIMINATIVE DICTIONARY LEARNING WITH APPLICATION TO PATTERN RECOGNITION

Román E. Rolón[♭], Leandro E. Di Persia[♭], Hugo L. Rufiner[♭,†] and Rubén D. Spies[‡]

[♭]*Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional (sinc(i), UNL-CONICET), Facultad de Ingeniería y Ciencias Hídricas, Univ. Nacional del Litoral, Santa Fe, Argentina, rrolon@sinc.unl.edu.ar, ldipersia@sinc.unl.edu.ar, lrufiner@sinc.unl.edu.ar*

[†]*Laboratorio de Cibernética, Facultad de Ingeniería, Univ. Nacional de Entre Ríos, Argentina.*

[‡]*Instituto de Matemática Aplicada del Litoral (IMAL), UNL-CONICET, Santa Fe, Argentina, rspies@santafe-conicet.gov.ar*

Abstract: Pattern recognition is a scientific discipline whose purpose is the classification of objects into different categories or classes. Object categorization deals with the detection or recognition of "generic" categories, reason for which it known as "generic object recognition". In this article, sparse representation of signals in terms of a discriminative multi-class dictionary for image recognition is presented. A sparse representation approximates an input signal over a linear combination of a few atoms of the given dictionary. A balanced set of input signals selected from the Caltech 101 database is used for learning the discriminative dictionary. The sparse vectors are then used as input of a multi-class classifier. The proposed method shows improvements over the standard KSVD method.

Keywords: *Dictionary learning, Inverse problems, Discriminative information*
2000 AMS Subject Classification: 21A54 - 55P54

1 INTRODUCTION

The problem of sparse representation of signals consists of obtaining representations of such signals by means of a linear combination of only a few atoms of an appropriately constructed dictionary [1]. Some of the advantages of sparse representations are super resolution, robustness to noise and dimension reduction, among others. The sparse representation problem can be divided into two separate sub-problems: an inference and a learning problem. The first one, which is usually called "sparse coding", consists of selecting a set of representation vectors $\{\mathbf{a}_i\}$ satisfying a given sparsity constraint. The second problem, which is quite often more complex, consists of finding an "optimal" dictionary Φ for representing a given set of signals $\{\mathbf{x}_i\}$. The formulation of the learning problem focuses only on minimizing the reconstruction error without taking into account the discriminative classification power of the dictionary [1]. For that reason, some authors have proposed different supervised approaches in order to take advantage of the discriminative power of the dictionary [2], [3]. In such supervised approaches the dictionary and a linear classifier are simultaneously optimized.

In a previous work [4], two different methods for identifying the most discriminative atoms in an overcomplete dictionary were presented. A method for learning discriminative dictionaries to be used for classification tasks is presented in this work.

The article is organized as follows: the method used for learning discriminative dictionaries is described in Section 2. Experiments and results are presented and discussed in Section 3. Finally, conclusions are presented in Section 4.

2 METHODS

2.1 SPARSE CODING

A sparse representation of a given data matrix of N input signals $X \doteq [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]$ where $\mathbf{x}_i \in \mathbb{R}^n$ in terms of a given dictionary $\Phi \in \mathbb{R}^{N \times M}$ (with $M \geq N$), whose columns ϕ_j are sometimes called atoms, can be obtained by minimizing the following problem,

$$A^* = \underset{A}{\operatorname{argmin}} \sum_{i=1}^{N} \left(||\mathbf{x}_i - \Phi \mathbf{a}_i||_2^2 + \lambda ||\mathbf{a}_i||_1 \right), \tag{1}$$

where λ is a sparsity constraint factor and the terms $||\mathbf{x}_i - \Phi \mathbf{a}_i||_2$ are the reconstruction errors. Finally, each input signal $\mathbf{x}_i$ is approximated by a linear combination of only a few atoms of the dictionary in an appropriate way.

2.2 DICTIONARY LEARNING

The aim of dictionary learning is to obtain an efficient dictionary that provides a good representation for most of the signals under study. The dictionary learning problem can be stated as follows:

$$< \Phi^*, A^* >= \underset{\Phi, A}{\operatorname{argmin}} \sum_{i=1}^{N} \left(||\mathbf{x}_i - \Phi \mathbf{a}_i||_2^2 + \lambda ||\mathbf{a}_i||_1 \right). \tag{2}$$

Problem (2) is convex in each one of the variables Φ and A, individually, but not simultaneously convex [5]. Therefore the dictionary learning problem is usually iteratively and sequentially solved, by first optimizing in Φ (while holding A fixed) and then optimizing in A (while holding Φ fixed), so proceeding until a prescribed stop criteria is met. Clearly, formulation (2) minimizes the reconstruction error and this formulation does not take into account the discriminative information of the dictionary, which is completely neglected for the classification task.

2.3 DISCRIMINATIVE DICTIONARY LEARNING

In a previous work [4] we have introduced a novel approach for identifying the most discriminative atoms in a binary classification problem. In this work a method for learning a discriminative dictionary to be used for solving a multi-class problem is proposed. The activation frequency of the atoms is denoted by η_κ^j where j represents the activation of atom ϕ_j for input signals labeled as belonging to class κ. With η_κ^j we shall denote the number of times that the atom ϕ_j is used to represent data belonging to class κ. Suppose that the atom ϕ_j has a very high activation for the class "κ" but very low activation for the remaining classes. In such a case the atom ϕ_j is considered to be highly discriminative for classifying elements belonging to the class "κ". Otherwise, if the atom ϕ_j has similar activation for the class "κ" and for the remaining classes, then the atom is considered as not carrying any significant discriminative information.

The discriminative approach begins by defining a matrix $D \in \mathbb{R}^{M \times \kappa}$, which represents a measure of the discriminative power of the atoms:

$$D(j,k) = \frac{1}{N} \left(\eta_k^j N_{\kappa \neq k} - \sum_{\kappa \neq k} \eta_k^j N_k \right), \tag{3}$$

where N_k denotes the number of input signals belonging to the class "k" while $N_{\kappa \neq k} (= n - N_k)$ denotes the number of signals not belonging to that class. The elements in the rows of D represent the difference of activations of the atoms in the dictionary and the elements of its columns represent the difference of activation frequencies (see Eq. (3)). Clearly $D(j,k)$ will be positive and large if the j^{th}-atom is much more active in one class than in the others. Otherwise, if the j^{th}-atom has similar activation frequencies for all the classes, then $D(j,k)$ will be small or even negative.

We describe now the building steps of the discriminative dictionary learning method together with the corresponding lines of its implementation algorithm (Algorithm 1). Let X be the training data, p_0 the sparsity level, I the final number of discriminative atoms for each class and κ the number of classes. The algorithm starts by learning a dictionary Φ by using the unsupervised KSVD algorithm [1] (line 4). Then each representation matrix A_κ is obtained by applying a greedy pursuit algorithm called OMP [6] (line 5). Then, matrix D is obtained according to Ec. (3) (line 6). An atom selection process follows by selecting the most discriminative ones for each one of the classes. Those discriminative atoms constitute the columns of the matrix Φ_d. It could happen that no discriminative atoms are available for certain classes. In such a case, the corresponding columns of Φ_d are represented by vectors of zeros and the indices of those classes are saved into the vector Rem (line 7). Finally a new data matrix $\hat{X}$ is constructed by removing all the input signals of X belonging to the classes corresponding to the discriminative atoms (line 12). This process is repeated until all of the discriminative atoms are acquired.

The idea behind this approach is to learn a discriminative dictionary where the activations of the atoms contain significant information to be used for classification. Thus, a sparse version of a multi-class linear discriminant analysis (LDA) has been chosen for classification [7].

Algorithm 1 DKSVD

procedure DKSVD(X, p_0, I, κ)
  $inc = 1$
  for $i \leftarrow 1, I$ **do**
    $\Phi \leftarrow$ KSVD(X, p_0)
    Get sparse matrix $A = [A_1\ A_2\ A_3\ \cdots\ A_\kappa]$ that accomplish $X = \Phi A$
    Get D according to Ec. (3)
    Get Φ_d and Rem
    if $Rem = \emptyset$ **then**
      $inc = 0$
    end if
    while $inc = 1$ **do**
      Get $\hat{X}$ by removing the input signals corresponding to the discriminative atoms
      $\Phi \leftarrow$ KSVD($\hat{X}, p_0$)
      Get sparse matrix A that accomplish $\hat{X} = \Phi A$
      Get D according to Ec. (3)
      Get Φ_d and Rem
      if $Rem = \emptyset$ **then**
        $inc = 0$
      end if
    end while
    $\Phi_D \leftarrow [\Phi_D\ \Phi_d]$
    Remove randomly a signal per class
  end for
  $\Phi_D \leftarrow [\Phi_{c1}\ \Phi_{c2}\ \cdots\ \Phi_{c\kappa}]$ where $\Phi_{c\kappa} = [\phi_{c1}^1\ \phi_{c2}^2\ \cdots\ \phi_{c\kappa}^I]$
  return Φ_D
end procedure

3 Experiments and results

The Caltech 101 database is used for this work [8]. This database is widely used and it contains 9144 images corresponding to 101 different objects (classes) with an extra background class. The images have an average size of 60,000 pixels (300×200). The number of images belonging to each class varies from 31 (inline_skate) to 800 (airplanes). The images corresponding to the background and faces classes were not taken into account for this work. Thus, a total of 100 classes were considered.

To balance the database, a total of 30 images from each class were randomly selected, of which 15 were used for training and the remaining ones for testing purposes. In the pre-processing stage, the region of interest of each image was automatically cropped, converted into gray scale, normalized using histogram normalization, resized to 32×32 and finally converted into vectors of length 144 using Principal Components Analysis (PCA) [9].

The matrix of input signals $X \in \mathbb{R}^{144\times3000}$ was built by staking side-by-side the input matrices corresponding to each class, i.e. $X = [X_{c1}, X_{c2}, \cdots, X_{c100}]$. Next, the signal matrices $X_{train} \in \mathbb{R}^{144\times1500}$ and $X_{test} \in \mathbb{R}^{144\times1500}$ were constructed by randomly selecting signals from each class in X.

The first step of our method is to obtain the initial dictionary. In this step no information about classes is required. To accomplish this task the standard KSVD algorithm is used [1]. In the following step, the representation coefficients of X_{train} are obtained by applying the OMP algorithm [6]. By considering the representation coefficients, the matrix D which contains significant information about the discriminative power of each atom in the initial dictionary, is constructed. The image appearing on the left of Figure 1 shows a representation of the matrix D. It can be seen that, for most of the classes, the atoms have a high variability of activation frequency. On the right of Figure 1, a 3-D representation of matrix D, where the elements of each column are now shown in decreasing order of magnitude, is presented. It can be clearly seen that, for instance, the 258^{th}-atom contains high discriminative information for class 3

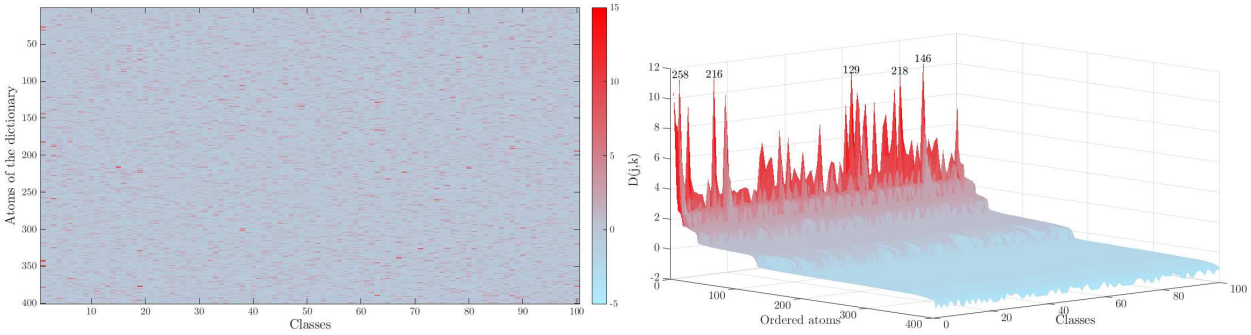


Figure 1: Difference of activation frequency. Matrix D (left); Matrix D with decreasingly reordered columns (right).

($D(258, 3) = 11.48$). On the studied database, by considering 15 input signals for training and testing, a classification accuracy of 24,6% was obtained. Compared with the use of a dictionary learned by means of the standard KSVD, whose classification accuracy resulted in 21,3%, our result is encouraging.

4 Conclusions

A new approach to pattern recognition using sparse representations was presented. The results show that the design of a discriminative dictionary is a suitable technique to be used for multi-class classification tasks. Although far more research is needed in order to improve the classification performances, the approach constitutes a promising method for learning a discriminative dictionary. For future work we propose to use over-sampling techniques [10] for increasing the size of the database. The effects of using different types of classifiers will also be studied.

Acknowledgments

This work was supported in part by Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, through PIP 2014-2016 Nro. 11220130100216-CO and PIP 2012-2014 Nro. 114 20110100284-KA4, by the Air Force Office of Scientific Research, AFOSR / SOARD, through Grant FA9550-14-1-0130 and by Universidad Nacional del Litoral through project CAI+D 501 201101 00519 LI.

References

[1] Aharon M., Elad M. and Bruckstein A. *K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation*. IEEE Transactions on Signal Processing, Vol. 54, (2006), pp. 4311-4322.

[2] Pham D. and Venkatesh, S. *Joint learning and dictionary construction for pattern recognition*. In proceedings CVPR Workshop on Generative-Model Based Vision, (2008).

[3] Jiang Z., Lin Z. and Davis L. *Learning a discriminative dictionary for sparse coding via label consistent k-svd*. In proceedings CVPR Workshop on Generative-Model Based Vision, (2011).

[4] Rolón R., Larrateguy L., Di Persia L., Spies R. and Rufiner L. *Discriminative methods based on sparse representations of pulse oximetry signals for sleep apnea-hypopnea detection*. Biomedical Signal Processing and Control, Vol. 33, (2017), pp. 358-367.

[5] Guo H., Jiang Z. and Davis L. *Discriminative Dictionary Learning with Pairwise Constraints*. In proceedings ACCV Asian Conference on Computer Vision, Vol. 7724, (2012), pp. 328-342.

[6] Pati Y.C., Rezaiifar R. and Krishnaprasad P.S. *Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition*. IEEE Asilomar Conference on Signals, Systems and Computers, Vol. 1, (1993), pp. 40-44.

[7] Clemmensen L., Hastie T., Witten D. and Ersbøll B. *Sparse Discriminant Analysis*. Technometrics, 53, (2011), pp. 406-413.

[8] Fei-Fei L., Fergus R. and Perona P. *Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories*. In proceedings CVPR Workshop on Generative-Model Based Vision, (2004).

[9] Jolliffe I.T. *Principal Component Analysis*. Springer, 2nd ed, (2002).

[10] Chawla N., Bowyer K., Hall L. and Kegelmeyer P. *SMOTE: Synthetic Minority Over-sampling Technique*. Journal of Artificial Intelligence Research, Vol. 16, (2002), pp. 321-357.

[11] Crammer K. and Singer Y. *On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines*. Journal of Machine Learning Research 2, (2001), pp. 265-292.

Anexo D

A multi-class structured dictionary learning method using discriminant atom selection

Artículo enviado para su publicación en septiembre de 2018.

A multi-class structured dictionary learning method using discriminant atom selection

R.E. Rolón[a,d], L.E . Di Persia[a], R.D. Spies[b] and H.L. Rufiner[a,c]

September 20, 2018

[a] Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, sinc(i), FICH, UNL, CONICET, Santa Fe, Argentina
[b] Instituto de Matemática Aplicada del Litoral, IMAL, FIQ, UNL, CONICET, Santa Fe, Argentina
[c] Laboratorio de Cibernética, Fac. de Ing., Univ. Nacional de Entre Ríos, Argentina
[d] Facultad Regional Paraná, Universidad Tecnológica Nacional, Entre Ríos, Argentina

Abstract

In the last decade, traditional dictionary learning methods have been successfully applied to various pattern classification tasks. Although these methods produce sparse representations of signals which are robust against distortions and missing data, such representations quite often turn out to be unsuitable if the final objective is signal classification. In order to overcome or at least to attenuate such a weakness, several new methods which incorporate discriminative information into sparse-inducing models have emerged in recent years. In particular, methods for discriminative dictionary learning have shown to be more accurate (in terms of signal classification) than the traditional ones, which are only focused on minimizing the total representation error. In this work, we present both a novel multi-class discriminative measure and an innovative dictionary learning method. For a given dictionary, this new measure, which takes into account not only when a particular atom is used for representing signals coming from a certain class and the magnitude of its corresponding representation coefficient, but also the effect that such an atom has in the total representation error, is capable of efficiently quantifying the degree of discriminability of each one of the atoms. On the other hand, the new dictionary construction method yields dictionaries which are highly suitable for multi-class classification tasks. Our method was tested with a widely used database for handwritten digit recognition and compared with three state-of-the-art classification methods. The results show that our method significantly outperforms the other three achieving good recognition rates and additionally, reducing the computational cost of the classifier.

Keywords: Multi-class discriminative measure, structured dictionary learning, sparse coding, handwritten digit recognition.

1 Introduction

Sparse representation of signals is considered a very powerful signal processing technique which has drawn massive interest in recent years mainly due to its success in solving a wide variety of problems in different fields such as biomedical signal processing [1, 2], computer vision [3] and image analysis [4], including image denoising [5], color image restoration [6] and image classification [7]. Roughly speaking, the problem of sparse representation consists of obtaining approximations of the involved signals in terms of linear combinations of only a few prescribed very simple characteristic signals taken from a large set [8, 9]. Besides providing a robust framework against distortions, missing data and noise, sparse representation of signals has many other advantages such as super resolution and dimensionality reduction [10].

A sparse representation problem (SRP) is usually divided into two sub-problems: an inference problem and a learning problem. The first one, which is often called "sparse coding", consists of computing a representation vector satisfying a particular sparsity constraint given a predefined dictionary. The second one, which involves solving a more complex problem, consists of finding an "optimal", in certain sense, dictionary for representing a given set of training signals. It is important to point out however, that most formulations of SRPs only focus on minimizing a prescribed total representation error and they do not take into account any a-priori discriminative information which could significantly improve the performance in the case of multi-object classification problems.

The first data-driven dictionary learning algorithms were originally developed almost two decades ago [8, 11, 12]. Some of them have their roots in probabilistic frameworks by considering the observed data as

realizations of certain random variables [8, 11]. In [11] for example, the authors developed an algorithm for finding a redundant dictionary maximizing the likelihood function of the probability distribution of the data. In that work, an analytic expression for the likelihood function was derived by approximating the posterior distribution by Gaussian functions. On the other hand, an iterative approach for dictionary learning, known as the "Method for Optimal Directions" (MOD), was presented in [12]. The sparse coding stage of this method makes use of a greedy algorithm called "Orthogonal Matching Pursuit" (OMP) [13] followed by a simple dictionary updating rule.

A new iterative algorithm was proposed by Aharon *et.al.* in [9]. This new approach, called "K Singular Value Decomposition" (KSVD), consists mainly of two stages: a sparse coding stage and a dictionary learning stage. The OMP algorithm is used in the sparse coding stage, which is followed by a dictionary updating step where the atoms are updated one at a time and the representation coefficients are allowed to change in order to minimize the total representation error.

In the last decade, the interest in developing algorithms based on sparse representation of signals for pattern recognition purposes has notably increased [7, 14, 15]. This is so because a large number of authors have proposed new supervised approaches for pattern recognition using sparse representations of signals. For instance, a discriminative version of the standard KSVD method applied to face recognition was presented by Zhang Q. *et.al.* [7]. In that work, the authors included a discriminative term into the objective function of the standard KSVD algorithm. Results have shown that this modification constitutes an appropriate way to learn dictionaries which satisfy both criteria: low reconstruction error and high recognition rates. Also, Pham D. *et.al.* [14] proposed an iterative method that simultaneously optimizes a dictionary and a linear classifier. The authors successfully used the method in an image categorization problem. More recently, a novel approach called "Label Consistent KSVD" (LC-KSVD) for dictionary learning was proposed in [15]. In that work a discriminative sparse representation and a single predictive linear classifier were efficiently integrated into the objective function.

However, besides supervised dictionary learning methods, many other new alternative options were presented [4, 16, 17]. These new alternatives are mainly based on the pursuit of discriminability of sparse representations through the development of "structured" or, more precisely, category-specific dictionary methods. In [4], a method for learning multiple dictionaries that uses the reconstruction errors yielded by these dictionaries on image patches to derive a pixel-wise classification. This algorithm has proved to be robust specially for local image classification tasks. A method for learning multiple non-redundant dictionaries for complex object categorization was proposed in [16]. This method was assessed on both visual object categorization and document classification image-related problems yielding competitive performances. In [17], a method that simultaneously optimizes both a structured dictionary (category-specific visual words for each feature) and a classifier was introduced. This method yielded good recognition rates showing a significant improvement over state-of-the-art object classification methods. A new method for structured dictionary learning was recently proposed by Sun *et.al.* [18]. In that work, the learned dictionary was decomposed into class-specific sub-dictionaries for the classification that is conducted measuring the minimum reconstruction error among all the classes. The method was tested using both the synthetic data and the real-world data showing good performances.

In this work we propose a novel multi-class discriminative measure and a new dictionary learning method which yields structured dictionaries which are composed by category-specific sub-dictionaries specially constructed for multi-class classification purposes. Thus, the novelty of our approach is twofold. First, we introduce an innovative and effective multi-class discriminative measure whose main property is precisely its capability for quantifying the discriminative degree of each one of the atoms in a given dictionary. This measure takes into account not only when a particular atom is used for representing a signal coming from a certain class and the magnitude of its corresponding representation coefficient, but also the effect that such an atom has in the total representation error. Secondly, this work presents a novel method for discriminative structured dictionary learning which yields a dictionary increasing the classifier recognition rate.

The organization of this article is as follows. A recall of sparse representation of signals is presented in Section 2. In Section 3, we make a brief description of the database used in the experiments as well as we present our new proposed discriminative measure and structured dictionary learning method. Section 4 contains details on all the experiments, while results and discussion are presented in Section 5. Finally, concluding comments and future works are given in Section 6.

2 Sparse representation of signals

Sparse representation is a signal processing technique that seeks the sparsest representation of all the signals in a given set in terms of linear combinations of certain basic waveforms. The sparse representation

problem can be separated into two sub-problems. Namely the so-called sparse coding problem and the dictionary learning problem. We shall now proceed to describe in detail each one of these sub-problems. For that, let $\mathbf{x} \in \mathbb{R}^N$ be a discrete signal and let $\Phi \in \mathbb{R}^{N \times M}$ (generally with $M \geq N$) be a dictionary whose columns $\boldsymbol{\phi}_j \in \mathbb{R}^N$ are atoms that we want to use for obtaining representations of $\mathbf{x}$ of the form $\mathbf{x} = \Phi\mathbf{a}$. Here, and in the sequel, we shall refer to the vector $\mathbf{a} = [a_1 \ a_2 \ \cdots \ a_M]^T \in \mathbb{R}^M$ as a "representation" of $\mathbf{x}$. Sparsity consists essentially of obtaining a representation with as few non-zero elements as possible. A way of obtaining such representations consists of solving the following problem:

$$(P_0): \quad \min_{\mathbf{a} \in \mathbb{R}^M} ||\mathbf{a}||_0 \text{ subject to } \mathbf{x} = \Phi\mathbf{a},$$

where $||\mathbf{a}||_0$ denotes the l_0 pseudo-norm, defined as the number of non-zero elements of $\mathbf{a}$. It turns out that imposing an exact representation of $\mathbf{x}$ is a too restrictive constraint, which makes (P_0) an NP hard problem [19, §1.8], yielding the approach highly unsuitable for most practical applications.

Hence, the exact representation requirement $\mathbf{x} = \Phi\mathbf{a}$ is often relaxed by allowing small representation errors and imposing an upper bound on the l_0 pseudo-norm of the representations. Thus, a small error representation tolerant version of (P_0) is defined as follows:

$$(P_0^q): \quad \min_{\mathbf{a} \in \mathbb{R}^M} ||\mathbf{x} - \Phi\mathbf{a}||_2^2 \text{ subject to } ||\mathbf{a}||_0 \leq q,$$

where q is a prescribed integer parameter. This formulation considers the presence of possible additive noise terms. In other words it assumes that a particular signal $\mathbf{x} = \Phi\mathbf{a} + \mathbf{e}$ where $\mathbf{e} \in \mathbb{R}^N$ is a small energy noise term. Thus, this approach is more appropriate in a wide variety of real applications (such as biomedical signal or image processing) where the captured raw signals are always contaminated by noise. Several greedy strategies have been proposed for solving problem (P_0^q) [20, 13]. Among them, the OMP algorithm is perhaps the most commonly used strategy. This greedy algorithm ensures convergence to the projection of $\mathbf{x}$ into the span of atoms in a given dictionary, in no more than q iterations. It is important to note that the representation vector $\mathbf{a}$ has only $q \ll M$ non-zero entries, while the remaining ones are strictly equal to zero. Figure 1 shows an example of the representation vectors obtained with this (P_0^q) approach for two images of different classes coming from a widely used database which we shall describe in detail in Section 3. Note that, in this case, most coefficients are strictly equal to zero.

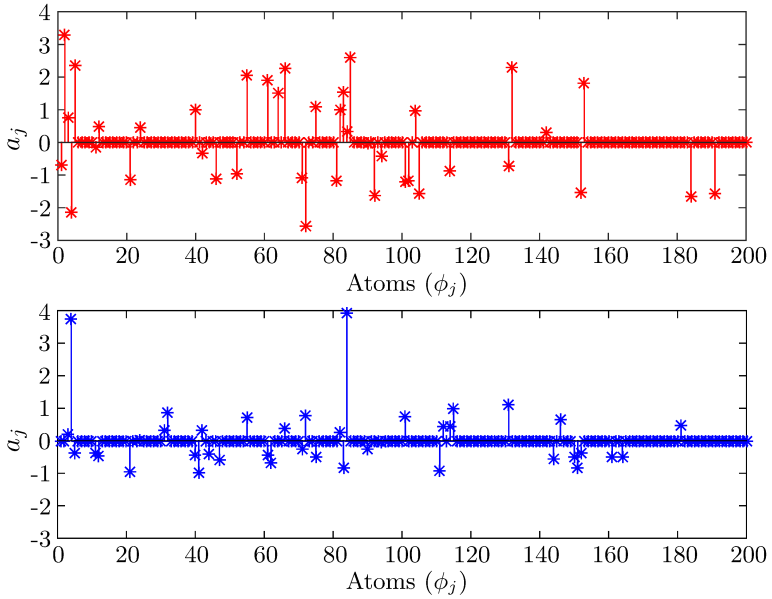


Figure 1: Example of two representation vectors of handwritten digits of two different classes obtained with OMP algorithm.

Although pre-constructed dictionaries such as the well known wavelet packets [21] typically lead to fast sparse coding, they are almost always highly restricted to certain classes of signals. Hence, due to their lack of generalization, new approaches introducing data-driven dictionary learning techniques have emerged. A dictionary learning problem associated to the data: k, M, $N \in \mathbb{N}$, $M \geq N$ and a collection

of n signals in $\mathbb{R}^N$, $\mathbf{x}_1, \cdots, \mathbf{x}_n$, can be formally written as:

$$(DL): \quad \min_{\substack{\Phi \in \mathbb{R}^{N \times M} \\ \mathbf{a}_i \in \mathbb{R}^M, ||\mathbf{a}_i||_0 \leq q, 1 \leq i \leq n.}} \sum_{i=1}^{n} ||\mathbf{x}_i - \Phi \mathbf{a}_i||_2^2$$

The solution of this problem yields on one hand a dictionary Φ and, on the other hand, representations for all the signals in terms of that dictionary complying with the sparsity constraint for each one of the "n" involved signals $\mathbf{x}_1, \cdots, \mathbf{x}_n$. It is important to point out that in such a process, the total representation error is minimized.

Although data-driven dictionary learning algorithms produce sparse representations of signals which are robust against distortions and missing data, such representations quite often turn out to be unsatisfactory if the final objective is signal classification. This is mainly because those algorithms do not take into account any prior available information concerning class membership. To overcome this flaw, several alternative approaches producing sparse representations in terms of a unique (and shallow) dictionary for signal classification were presented [7, 14, 15].

A different approach is the construction of structured dictionaries composed by sub-dictionaries whose atoms are discriminative, in certain sense, for each one of the classes, i.e. each sub-dictionary has a group of atoms that are discriminative only for a particular class. The use of structured dictionaries could be useful for reducing the features dimension, avoiding over-fitting and optimizing the performance of a classifier, among others. In recent years, there has been increasing interest in developing algorithms whose main purpose is to obtain "optimal" sub-dictionaries to be used for signal classification [1, 22, 23]. In [22], a method called clustering based online learning of dictionaries (COLD) was presented. This algorithm makes use of the mean shift clustering procedure [24] to identify modes in the distribution of the atoms and, hence obtain a dictionary of minimal size. Recently, Chen *et.al.* [23] introduced a dictionary learning method for image and video editing tasks. In that work, the problem of seeking an optimal dictionary is solved by using a symmetric version of the Kullback-Leibler (KL) divergence [25]. This divergence has been successfully used for detecting redundant atoms in a given dictionary. Our proposal consists of defining and using a new discriminative measure for selecting the most discriminative atoms for each one of the classes and therefore use them for building a new structured dictionary.

3 Materials and methods

In this section we make a brief description of the database used in the experiments. Additionally, we describe in detail both the new proposed multi-class discriminative measure and the novel structured dictionary learning method.

3.1 Database

One of the most popular databases used to assess Computer Vision and Pattern Recognition methods is the Modified NIST (MNIST) database [26]. This database has been widely used for assessing new methods including Deep Learning techniques [27], Extreme Learning Machines [28] and a many types of neural networks [29], among others. The MNIST database contains a total of 70,000 normalized and centered (in a gray-scale) images of handwritten digits ranging from 0 (zero) to 9 (nine) each one of size 28×28 (or a feature vector of length 784). Additionally, this database provides information about standard partitions used for training (60,000) and testing (10,000).

Although each one of all original (raw) images coming from the MNIST database can be represented as a single column vector consisting of 784 elements (features), it becomes necessary to reduce its dimensionality for practical reasons. In this work, the image dimension reduction process is carried out by using the well known bicubic interpolation method [30] which is not only accurate, but smooth and computationally efficient. The bicubic interpolation method has been used for obtaining new (reduced) images each one of size 16×16 which correspond to a feature vector of length 256.

3.2 A new discriminative measure

Discriminative dictionaries can be thought of as a collection of atoms specially learned for signal classification. These dictionaries not only produce accurate representations of the training signals (in terms of their waveforms) coming from different classes, but they also render their representations easy to distinguish by a suitable classifier. However, the problem of finding a discriminative dictionary is computationally very challenging. A way to overcome the computational complexities entailed by such a problem consists

of defining an appropriate discriminative value functional that independently evaluates each one of the atoms in a given dictionary. This simplification is based on the assumption that each atom in the dictionary is used to model specific characteristics that are not modeled by any of the other atoms. Thus, the discriminative information provided by a particular atom is different from the information contributed by all other atoms.

In a previous work [1], we presented a simple approach for quantifying the discriminative degree of the atoms of a given dictionary Φ in the context of a binary classification problem. The approach essentially consists of counting the number of times a particular atom is used, i.e. the number of atoms that become active for representing signals belonging to each one of both classes $\ell = 1$ and $\ell = 2$. As a result of this counting process, an activation frequency (η) for each atom given the class, is considered. To quantify the discriminative degree of the j^{th}-atom (ϕ_j, the j^{th}-column of Φ), the absolute difference of activation frequencies of that atom for classes $\ell = 1$ and $\ell = 2$ ($|\eta_1^j - \eta_2^j|$) is computed. This value will be large if (an only if) the atom ϕ_j is much more frequently used for representing signals in one of the two classes and, in that case, it can be thought of as a quantifier of the capability of ϕ_j to supply important discriminative information regarding class membership. The use of this discriminative quantifier gave rise to a method called Most Discriminative Column Selection (MDCS) for discriminative sub-dictionary construction [1]. The MDCS method has shown to be robust for efficiently extracting meaningful features from segments of pulse oximetry signals for detecting apnea-hypopnea events.

In this work we propose an extension of the measure described above to multi-class classification problems. This extension consists of defining and using a new multi-objective function $m_{\alpha,\beta}$ aimed at quantifying the discriminative degree of each one of the atoms in a given dictionary. This function $m_{\alpha,\beta}$ will be defined as a convex combination of three discriminative terms, all based on the affine sparse representations of the data. In what follows, a detailed description of each one of such terms as well as a formal definition of $m_{\alpha,\beta}$ are presented.

3.2.1 Activation frequency measure

Conditional activation frequencies provide a reasonable starting point for determining the discriminative degree of individual atoms in a given dictionary. For this reason, our approach begins by computing the activation frequency η_ℓ^j of ϕ_j given the class ℓ, for $\ell = 1, 2, \cdots, k$. Moreover, the conditional activation probability of ϕ_j given (that a signal $\mathbf{x}$ belongs to) class ℓ is defined as $p_\ell^j \doteq P(a_j \neq 0 | \mathbf{x} \in \ell)$. Given a set of n_ℓ signals belonging to class ℓ, this conditional probability can be approximated by the quotient η_ℓ^j / n_ℓ. Note that if the problem is balanced, i.e. if the number of available signals belonging to each one of the k classes is the same, say $\hat{n}$, then $\eta_\ell^j \propto p_\ell^j$, more precisely $\eta_\ell^j = k\hat{n}p_\ell^j$, for all ℓ and j. In this work, the problem of quantifying the discriminability of each atom is tackled by analyzing their individual contributions to the signal classification process. More specifically, a particular atom ϕ_j is considered as having important discriminative information for class ℓ signals if $p_\ell^j > p_m^j$, for all $m \neq \ell$. Hence, if ϕ_j is discriminative for class ℓ, the activation of the representation coefficient a_j will be strongly associated to class ℓ membership. Since the performance of a classifier highly depends on the discriminability of their inputs, it is reasonable to think that using the representation coefficients a_1, a_2, $\cdots$, a_M as inputs of a classifier, for atoms selected using that criterion, could result in good recognition rates.

For a given j, $1 \leq j \leq M$, we shall denote by ℓ_j^+ the class that maximizes all conditional activation probabilities p_ℓ^j, for all $\ell = 1, 2, \cdots, k$, i.e. such that

$$p_{\ell_j^+}^j = \max_{1 \leq \ell \leq k} p_\ell^j. \tag{1}$$

In the (unlikely) case that there is more than one value of ℓ maximizing p_ℓ^j, ℓ_j^+ is defined by randomly chosen one of them, for instance the smallest one (note that the order of the classes is completely irrelevant). Similarly, for a fixed j, $1 \leq j \leq M$, ℓ_j^* is defined as the class leading to the second largest conditional activation probability, i.e. such that

$$p_{\ell_j^*}^j = \max_{\substack{1 \leq \ell \leq k \\ \ell \neq \ell_j^+}} p_\ell^j. \tag{2}$$

Here again if there is more than one value of ℓ_j^* satisfying (2), then ℓ_j^* is chosen randomly as any one of them.

Next we define the function $m_{af} : \{1, 2, \cdots, M\} \to \mathbb{R}_0^+$ by

$$m_{af}(j) \doteq \frac{p_{\ell_j^+}^j - p_{\ell_j^*}^j}{p_{\ell_j^+}^j}, \tag{3}$$

we shall refer to $m_{af}(\cdot)$ as the "activation frequency measure".

Note that $0 \leq m_{af}(\cdot) \leq 1$. The atom ϕ_j is said to be discriminative (for class ℓ_j^+) if and only if $m_{af}(j) > 0$. Clearly, within this setting, if an atom ϕ_j is discriminative, it will be so only for the class ℓ_j^+, otherwise it will be discriminative for none of them. Moreover, the value of $m_{af}(j)$ can be thought of as a "measure" of the degree of discriminability of the atom ϕ_j (for the corresponding class ℓ_j^+), based solely on the activation frequency information.

Figure 2 shows examples of two graphic representations of conditional activation probabilities p_ℓ^1 and p_ℓ^2, for $\ell = 1, 2, \cdots, 10$, associated to atoms ϕ_1 (top) and ϕ_2 (bottom), respectively. The vertical bars represent the value of each conditional activation probability p_ℓ^1 (top) and p_ℓ^2 (bottom), for $\ell = 1, 2, \cdots, 10$. Clearly, for the top case (atom ϕ_1) $\ell_1^+ = 4$ and $\ell_1^* = 5$, $m_{af}(1) > 0$ and therefore the atom ϕ_1 is considered to be discriminative (for class 4). For the bottom case (atom ϕ_2) $\ell_2^+ = 2$ and $\ell_2^* = 7$ (although these values could be interchanged), but since $p_2^2 = p_7^2$, one has $m_{af}(2) = 0$ implying that ϕ_2 is not discriminative for class ℓ_2^+, and therefore is not discriminative for any one of the classes.

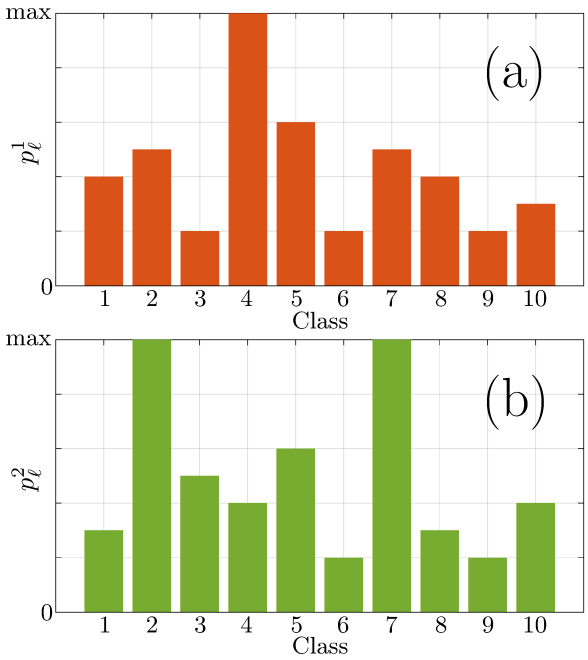


Figure 2: Vertical bars representing conditional activation probabilities for a discriminative atom ϕ_1 (top) and for a non-discriminative atom ϕ_2 (bottom) in the sense proposed here.

3.2.2 Coefficient magnitude measure

On one hand, the sparse representation of signals provides valuable information regarding the activation of atoms and, on the other hand, it can highlight important characteristics or features contained in particular event related waveforms of signals or images such as brightness variations in images and slight changes in biomedical signals, to name but a few. With the above observation in mind, we proceed now to define a second measure that takes into account the magnitude of the representation coefficients. For that, given an atom ϕ_j, let ℓ_j^+ and ℓ_j^* be the classes as defined in (1) and (2), respectively, and let $\mathbf{A}_{\ell_j^+}$ and $\mathbf{A}_{\ell_j^*}$ the matrices which provide the sparse representations of $\mathbf{X}_{\ell_j^+}$ and $\mathbf{X}_{\ell_j^*}$, respectively, in terms of the dictionary Φ, i.e. $\mathbf{X}_{\ell_j^+} = \Phi\mathbf{A}_{\ell_j^+}$ and $\mathbf{X}_{\ell_j^*} = \Phi\mathbf{A}_{\ell_j^*}$. Additionally, let q_ℓ^j denote the quotient $||\,[\mathbf{A}_\ell]_{j,:}\,||_1/n_\ell$, where $[\mathbf{A}_\ell]_{j,:}$ represents the j^{th}-row of the matrix $\mathbf{A}_\ell$. The coefficient magnitude measure is the function $m_{cm} : \{1, 2, \cdots, M\} \to \mathbb{R}_0^+$ defined by

$$m_{cm}(j) \doteq \frac{q_{\ell_j^+}^j - q_{\ell_j^*}^j}{q_{\ell_j^+}^j}. \tag{4}$$

Here again $0 \leq m_{cm}(\cdot) \leq 1$. Based on this measure, an atom ϕ_j is said to be discriminative (for the class ℓ_j^+) if and only if $m_{cm}(j) > 0$ and, in that case, the value of $m_{cm}(j)$ quantifies the corresponding degree of discriminability of ϕ_j for the class ℓ_j^+.

3.2.3 Representation error measure

We now proceed to describe the third measure for quantifying on an indirect way the discriminative degree of each atom in a dictionary. This measure takes into account the contribution of each atom ϕ_j to the total representation error. Let $\mathbf{A}_\ell \doteq [\mathbf{a}_1\ \mathbf{a}_2\ \cdots \mathbf{a}_{n_\ell}]$ be the matrix providing the sparse representation of $\mathbf{X}_\ell$, $\mathbf{X}_\ell \doteq [\mathbf{x}_1\ \mathbf{x}_2\ \cdots \mathbf{x}_{n_\ell}]$, as in the previous measure. Clearly, the contribution of the class ℓ to the total representation error can be written as [9]

$$\begin{aligned}
\sum_{i=1}^{n_\ell} ||\mathbf{x}_i - \Phi \mathbf{a}_i||_2^2 &= \|\mathbf{X}_\ell - \Phi \mathbf{A}_\ell\|_F^2 \\
&= \left\| \mathbf{X}_\ell - \sum_{j=1}^{M} \phi_j \left[\mathbf{A}_\ell\right]_{j,:} \right\|_F^2 \\
&= \left\| \left(\mathbf{X}_\ell - \sum_{i \neq j} \phi_i \left[\mathbf{A}_\ell\right]_{i,:} \right) - \phi_j \left[\mathbf{A}_\ell\right]_{j,:} \right\|_F^2 \\
&\doteq \left\| \mathbf{E}_\ell^j - \phi_j \left[\mathbf{A}_\ell\right]_{j,:} \right\|_F^2, \qquad (5)
\end{aligned}$$

where $\mathbf{E}_\ell^j$ denotes the total representation error for all class ℓ signals when ϕ_j is removed. Hence, a large value of $\mathbf{E}_\ell^j$ indicates that the contribution of ϕ_j to the representation of the class ℓ signals is large. We then define a "representation error measure" $m_{re} : \{1, 2, \cdots, M\} \to \mathbb{R}_0^+$ by

$$m_{re}(j) \doteq \frac{r_{\ell_j^+}^j - r_{\ell_j^*}^j}{r_{\ell_j^+}^j}, \qquad (6)$$

where $r_\ell^j \doteq \mathbf{E}_\ell^j / n_\ell$, for $\ell = 1, 2, \cdots, k$, $j = 1, 2, \cdots, M$.

Here again $0 \leq m_{re}(\cdot) \leq 1$, and an atom ϕ_j is said to be discriminative (for class ℓ_j^+) with respect to this measure if and only if $m_{re}(j) > 0$. In such a case, the value of $m_{re}(j)$ represents the corresponding degree of discriminability.

3.2.4 A combined discriminative measure

Each one of the three measures previously defined takes into account different properties related with the discriminability of each one of the atoms (in a given dictionary). It is then reasonable to think of a measure that appropriately combines all three of them. With that in mind, given two positive parameters α and β, with $\alpha + \beta \leq 1$, we define the function $m_{\alpha,\beta} : \{1, 2, \cdots, M\} \to \mathbb{R}_0^+$ as

$$m_{\alpha,\beta}(j) \doteq \alpha m_{af}(j) + \beta m_{cm}(j) + (1 - \alpha - \beta) m_{re}(j). \qquad (7)$$

We shall refer to $m_{\alpha,\beta}$ as a "combined discriminative measure". Clearly, (7) exhausts all possible convex combinations of the three single measures m_{af}, m_{cm} and m_{re}. A challenging problem, on which we shall shed some light in Section 4.3, consists precisely of finding the "optimal" pair of parameters (α^*, β^*) leading to the best recognition rate, for a given problem.

3.3 Dictionary learning algorithm

Supervised dictionary learning methods have observed great interest in recent years. Implementations of these methods were originally focused on efficiently learning simple dictionaries (unstructured) that incorporate information of "discriminability" (in terms of signal classification) in their optimization process. This information can be introduced to the learning model by considering different discriminative criteria [31, 32, 33]. The most commonly used criteria are the so called "softmax" cost function [16], Fisher criterion [34] and linear predictive classification error [7, 14], to name but a few.

Although there exist several ways to simultaneously optimize both a dictionary, i.e. to solve a representation learning problem, and a classifier, i.e. to find a solution to a classification problem, a very often used strategy consists simply of dividing that problem into two sub-problems [4, 16]. Hence, it is possible to use all existing traditional dictionary learning techniques such as MOD and KSVD (to name but a few) and therefore train a single classifier at a later stage. Our proposal is based precisely on this strategy. For that, we propose a new method for multi-class structured dictionary learning called

"Discriminant Atom Selection KSVD" (DAS-KSVD) in which we use the proposed discriminative measure $m_{\alpha,\beta}$ to efficiently select discriminant atoms in a given dictionary. The DAS-KSVD method aims at building a dictionary $\Phi_D^{(I)}$ by stacking side-by-side k sub-dictionaries Φ_ℓ each one of size $N \times I$, for all $\ell = 1, 2, \cdots, k$, $\Phi_D^{(I)} \doteq [\Phi_1\ \Phi_2\ \cdots\ \Phi_k]$. It is important to point out that each sub-dictionary Φ_ℓ contains atoms that are discriminative, in terms of $m_{\alpha,\beta}$, for class ℓ signals.

Here, and in the sequel, we shall consider the vectors $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ as realizations of a particular random vector $\mathcal{X}$. Its time to start describing the building steps of the proposed DAS-KSVD method (Algorithm 1). For that, given an $N \times n$ signal matrix $\mathbf{X}_{trn}$ composed by $n = \sum_{\ell=1}^{k} n_\ell$ samples such that $\mathbf{X}_{trn} \doteq [\mathbf{X}_1\ \mathbf{X}_2\ \cdots\ \mathbf{X}_k]$, the sparsity level q, the redundancy factor r_f, the number of class ℓ training signals $t \ll n_\ell$, the number of iterations I and the class label vector $\mathbf{c}$, the proposed algorithm begins by defining a data uniform probability distribution p_1 over $\mathbf{X}_{trn}$ so $p_1(i) = 1/n$, for all i (Alg. 1, line 2). Here, the value of $p_1(i)$ refers to the probability that the random vector $\mathcal{X}$ is equal to $\mathbf{x}_i$ ($P(\mathcal{X} = \mathbf{x}_i)$) or, more precisely, $P(\mathbf{x}_i \in \mathbf{X}_{lrn})$. The matrix $\mathbf{X}_{lrn}$ is composed by all samples selected from $\mathbf{X}_{trn}$ and it is used for learning a dictionary Φ.

The iterative process of this algorithm (Alg. 1, lines from 3 to 10) begins by statistically sampling t samples (note that $t \ll n_\ell$, for instance 10 times smaller) from each class ℓ signal matrix $\mathbf{X}_\ell$. As a result of such a sampling process, a matrix $\mathbf{X}_{lrn}$ of size $N \times (t * k)$ is built (Alg. 1, line 4). Also, to compute the distribution p_{l+1} from both p_l and $\mathbf{X}_{lrn}$, we multiply the value of $p_l(i)$ by a non-negative number $\tau_1 < 1$ if (and only if) $\mathbf{x}_i$ has been selected, i.e. $p_{l+1}(i) = p_l(i)\tau_1$ (in that case $p_{l+1}(i) < p_l(i)$). Otherwise $p_l(i)$ is left unchanged. It is important to point out that an appropriate normalization of these weights forcing them to sum one is needed. Figure 3 shows five graphic representations of probability distributions p_l, for $l = \{1, 5, 10, 15, 20\}$. It can be observed that, at the first iteration, all samples have the same probability to be selected. In addition, see that the probability value of most samples decreases as the iteration order increases.

In order to increase robustness, all training signals used to learn the dictionary Φ (Alg. 1, line 5) are also degraded by incorporating an additive zero-mean Gaussian noise $\epsilon_{l,i}$ whose magnitude increases proportionally according to the iteration level. The magnitude of the noise is updated by $\epsilon_{l,i} = l\sigma_i\tau_2$, where σ_i is the variance of $\mathbf{x}_i$ and τ_2 is a (prescribed) non-negative number, $\tau_2 < 1$. For instance, the magnitude of the noise associated to the signal $\mathbf{x}_1$ at iteration 5 will be $\epsilon_{5,1} = 5\sigma_1\tau_2$. It is important to point out however that, the first iteration ($l = 1$) of the proposed learning algorithm leaves the original image undegraded. On the other hand, the dictionary Φ is learned by means of the traditional unsupervised KSVD algorithm [9]. Then the sparse matrix $\mathbf{A}_{lrn}$ is obtained by applying the previously mentioned OMP algorithm (Alg. 1, line 6). The reason for having chosen this pursuit algorithm is because it guarantees convergence to the projection of each one of the signals into the span of the dictionary atoms, in no more than q iterations leaving the rest of the coefficients equal to zero.

As previously mentioned, at the beginning of each iteration, the standard unsupervised KSVD algorithm was used to learn a dictionary Φ of size 256×256. Note that this stage of the learning process requires no information regarding training signals class membership. The selected subset of $t * k$ signals ($\mathbf{X}_{lrn}$) used to learn the dictionary was degraded by incorporating additive Gaussian noise with different magnitudes. Figure 4 shows some atoms of the dictionary Φ learned at iteration 1 (left) and at iteration 20 (right). It can be seen that, at the first iteration, the dictionary is learned by considering noise-free signals. Although the algorithm at iteration 20 makes use of highly degraded signals for learning dictionary Φ, it still preserves the structure of the handwritten digits on a blurred background.

The proposed discriminative approach consists of optimizing and using the new combined discriminative measure $m_{\alpha,\beta}$ for selecting the most discriminative atoms of Φ for each one of the k classes (Alg. 1, lines from 7 to 9). As explained in Section 3.2, the value of $m_{\alpha,\beta}(j)$ corresponds to the degree of discriminability of the atom ϕ_j for one (and only one) class, which is denoted by ℓ_j^+. Note that the process of selecting the most discriminative atoms carries a serious trouble since the problem of finding the optimal pair of parameters (α^*, β^*) is very challenging. For more details about the tuning of that pair of parameters, we refer the reader to Section 4.3 and Appendix A. Also, the construction of the sub-dictionary Φ_d (Alg 1, line 8) basically consists of taking one-by-one the most discriminative atoms of Φ for each one of the k classes and stacking them side-by-side. In the case that there is more than one ℓ_j^+ class-related candidate complying with the proposed discriminative criterion, ϕ_j is defined as the atom that maximizes all possible values of m_{α^*,β^*}. Otherwise, in case that Φ lacks of discriminative atoms, the signal selection process (Alg. 1, line 4) is restarted.

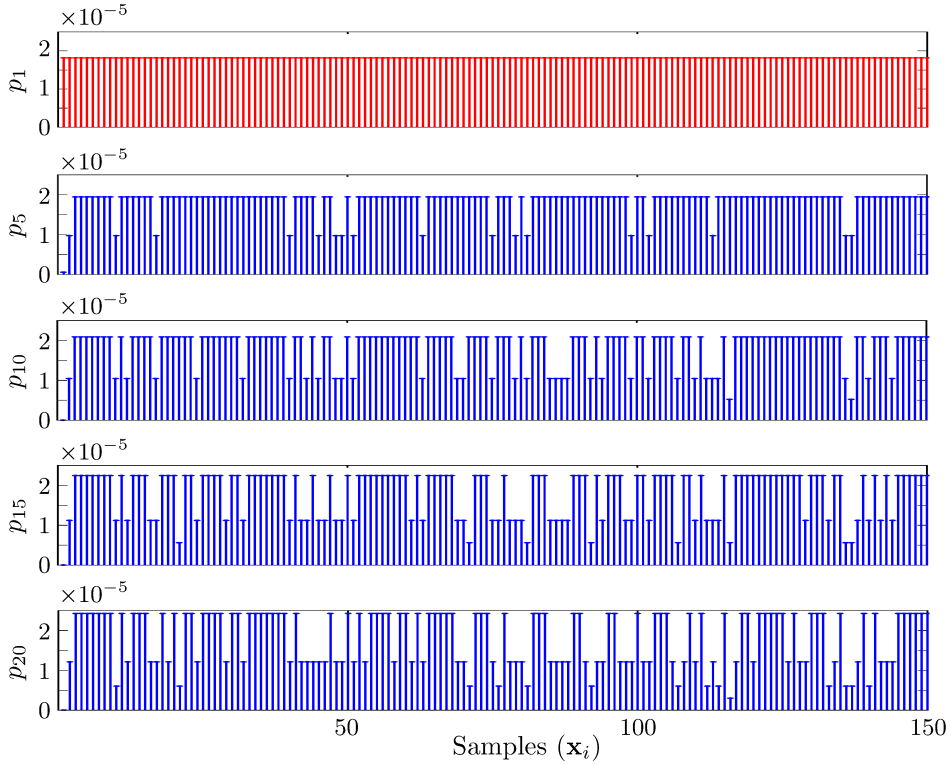


Figure 3: Data probability distributions for five different iterations of the proposed algorithm.

Iteration 1 ($l = 1$) Iteration 20 ($l = 20$)

Figure 4: Some atoms of the dictionary Φ for two different iterations of the DAS-KSVD algorithm. Iteration 1 (left) and iteration 20 (right).

Algorithm 1 Pseudocode of the new DAS-KSVD method

procedure DAS-KSVD($\mathbf{X}_{trn}, q, r_f, t, I, \mathbf{c}$)
  $p_1(i) = 1/n$, for all i
  for $l \leftarrow 1, I$ **do**
    $[\mathbf{X}_{lrn}, p_{l+1}] \leftarrow$ SAMPLEDATA($\mathbf{X}_{trn}, t, p_l, l$)
    $\Phi \leftarrow$ KSVD($\mathbf{X}_{lrn}, r_f, q$)
    $\mathbf{A}_{lrn} \leftarrow$ OMP($\mathbf{X}_{lrn}, \Phi, q$)
    $m_{\alpha^*,\beta^*} \leftarrow$ DISCMEASURE($\mathbf{A}_{lrn}, \mathbf{c}, q$)
    $\Phi_d \leftarrow$ GETATOMS($\Phi, m_{\alpha^*,\beta^*}$)
    $\Phi_D^{(i)} \leftarrow$ SAVEATOMS(Φ_d)
  end for
  return $\Phi_D^{(I)}$
end procedure

3.4 Classifier

In this work a Multilayer Perceptron (MLP) neural network is used in order to assess the proposed method. The MLP neural network is one of the most popular classes of neural networks whose architecture consists

of a fully connected assembly of single artificial neurons. The MLP neural network is typically comprised by an input layer, one (or more) hidden layers and an output layer [35]. The inputs (features) are processed layer-by-layer moving forward through the network. Each artificial neuron receives one (or more) inputs from its preceding nodes, processes the information and produces an output that is transmitted to the next node. The output of each neuron is reached by applying an activation (transfer) function (linear or not) to the weighted sum of the inputs plus a bias term. More precisely, the output of a neuron y_j is defined as

$$y_j = f\left(\sum_{i=1}^{d} \omega_{ji} x_i + \omega_{j0}\right) = f\left(\sum_{i=0}^{d} \omega_{ji} x_i\right), \tag{8}$$

where the transfer function is denoted by $f(\cdot)$, and the weights that connect the i^{th}-input to the j^{th}-neuron for a given layer is represented by ω_{ji}.

Since the MLP neural network training process is supervised, the desired outputs (labels) are required. The most popularly used method for training MLP neural networks is the back-propagation algorithm [36]. This algorithm iteratively adjusts the synaptic weights in the network by minimizing a given measure which quantifies the difference between the current output vector and the desired one.

4 Experiments

In this section we present a brief description regarding the experimental setup. Additionally, we make a brief recall of the evaluation metric used for assessing the proposed dictionary learning method. Finally, we comment on appropriate ways for tuning the parameters.

4.1 Experimental setup

As mentioned above, the performance of the new DAS-KSVD method is evaluated using standard partitions for training and testing of the MNIST database. This database contains 70,000 images of handwritten digits corresponding to 10 ($k = 10$) different classes which are numbers from 0 to 9. Also, the number of images per class varies from 5,421 to 6,742 which correspond to classes 5 and 1, respectively.

Although it is not a requirement, our experiments were performed by using a balanced set of training and validation samples. For that, subsets consisting of 4,000 and 1,000 images for each one of the classes coming from the standard partition of the training dataset were randomly chosen. Hence, new training and validation matrices ($\mathbf{X}_{trn}$ and $\mathbf{X}_{val}$) comprised by 40,000 and 10,000 samples, respectively, were built. It is important to point out however that, the standard partition of the testing dataset $\mathbf{X}_{tst}$ of size $256 \times 10,000$ was left unchanged.

It becomes appropriate to mention that the matrix $\mathbf{X}_{trn}$ was used both for dictionary learning and training the MLP neural network while the matrix $\mathbf{X}_{val}$ was used for testing the MLP neural network as well as for parameters tuning. Furthermore, the matrix $\mathbf{X}_{tst}$ was only taken into account for performing the final test.

We shall now proceed to describe the parameter settings for the DAS-KSVD method that were used in the experiments. We evaluated the effect that produces the size of $\Phi_D^{(I)}$ in the final recognition rate. For that, we have considered four structured dictionaries denoted by $\Phi_D^{(5)}$, $\Phi_D^{(10)}$, $\Phi_D^{(15)}$ and $\Phi_D^{(20)}$ which are composed by 50, 100, 150 and 200 atoms, respectively. Hence, the DAS-KSVD algorithm was run 20 iterations, i.e. $I = 20$.

4.2 Evaluation metric

Overall accuracy rate constitutes one of the most popular performance measures used to assess Pattern Recognition-related methods. The accuracy measure (Acc) is defined as the proportion of correctly predicted testing samples. Let n the number of testing samples, λ_i and $\hat{\lambda}_i$ the label and prediction, respectively, regarding $\mathbf{x}_i$ and $\delta(x, y)$ the well known delta function whose output is true (one) if $x = y$ and false (zero) otherwise. The Acc measure is defined as:

$$\text{Acc} = \frac{1}{n} \sum_{i=1}^{n} \delta(\lambda_i, \hat{\lambda}_i). \tag{9}$$

4.3 Parameters tuning

Although the pursuit for discriminative atoms is perhaps one of the most challenging issues to be addressed in this work, finding optimal pair of parameters (α^*, β^*) leading to the best recognition rate is also a very difficult task. However, the problem of finding that optimal pair of parameters strongly depends on the application under study. For that reason, we propose applying the well known and widely used "grid search" method for parameter optimization. For more details regarding grid search method, we refer the reader to Appendix A. In what follows, the final choice of the remaining parameters of the proposed algorithm are described.

At each iteration of the proposed DAS-KSVD method, one (and only one) discriminant atom for each one of the k classes is selected. Hence, each iteration of this method generates k discriminative atoms and therefore, if the algorithm is configured to perform I iterations, then the final structured dictionary will be composed by $I * k$ discriminant atoms. In order to explore the effect of the final structured dictionary size, the experiments were performed by considering a total of 20 iterations, i.e. $I = 20$. Thus, the final discriminative dictionary $\Phi_D^{(I)}$ is composed by 200 atoms (assuming $k = 10$). On the other hand, the number of samples for each class used to learn the full dictionary was set to $t = 500$.

As described in Section 3.3, τ_1 and τ_2 are two parameters $(0 \leq \tau_1, \tau_2 < 1)$ that need to be adjusted and fixed. Several trials were performed in order to obtain appropriate values for those parameters. A value of $\tau_1 = 0.5$ was finally selected and used in our experiments. Additionally, it was found that a value of $\tau_2 = 0.1$ presented the best trade-off between image degradation and iteration order.

The standard KSVD algorithm starts by performing a random selection of 256 samples coming from the learning signal matrix $\mathbf{X}_{lrn}$. Note that the redundancy factor (r_f) used for constructing the dictionary is equal to one, i.e. $M = N = 256$. Also, the maximum number of KSVD iterations was fixed to 50 in the code. It is also well known that the KSVD algorithm internally computes sparse codes representing each one of all involved signals. These codes were obtained by means of OMP algorithm. To establish an appropriate sparsity level, a great variety of sparse solutions were tested. It was found that a sparsity degree of 20% presents the best trade-off between discriminability and representativity of all signals.

The MLP neural network training process was performed using back-propagation method. This algorithm was optimized minimizing the Mean Squared Error (MSE) function through Scaled Conjugate Gradient (SCG) method. Also, the output of each neuron was determined by applying a saturating linear transfer function. Additionally, the structure of the MLP neural network was configured such that the sizes of its hidden and input layers are equal.

5 Results and discussion

As already explained above, the matrices denoted by $\mathbf{A}_{trn}$ and $\mathbf{A}_{val}$ provide the sparse representations of $\mathbf{X}_{trn}$ and $\mathbf{X}_{val}$, respectively, in terms of a dictionary Φ through $\mathbf{X}_{trn} = \Phi\mathbf{A}_{trn}$ and $\mathbf{X}_{val} = \Phi\mathbf{A}_{val}$. Also, the feature vectors $\mathbf{a}_i$ comprising the matrices $\mathbf{A}_{trn}$ and $\mathbf{A}_{val}$ were used as inputs for training and testing, respectively, the MLP neural network. The final test was performed by taken into account the standard partition of the testing dataset $\mathbf{X}_{tst}$ and each one of the previously learned structured dictionaries $\Phi_D^{(I)}$. The matrix $\mathbf{A}_{tst}$ was obtained by means of OMP algorithm. Also, the inputs of the already trained MLP neural networks are the feature vectors $\mathbf{a}_i$ coming from $\mathbf{A}_{tst}$ and, moreover, the outputs of these networks are evaluated to compute the final accuracy. In addition, structured dictionaries composed by 50, 100, 150 and 200 discriminant atoms were evaluated. Table 1 presents a comparative summary of the averaged accuracy rates yielded by MLP neural networks trained using as input the matrix $\mathbf{A}_{trn}$ obtained by taken into account each one of the evaluated structured dictionaries. The value of the accuracy rates represent the mean value over 10 rounds while maintaining fixed each one of the dictionaries. Also, details regarding the computational cost for each one of such dictionaries are included. It is important to point out that these results were obtained by considering a fixed hidden layer size coinciding with the input feature vector size. It can be observed that averaged recognition rates greater than 90% were obtained in all cases. Accuracy rates of 96.2, 95.2,95.0 and 92.2 were obtained for feature vector sizes of 200, 150, 100 and 50, respectively. Hence, results show that "discriminative" feature vectors of length 200 are the best option for handwritten digits recognition. On the other hand, the last column of Table 1 shows the total number of parameters required to train each one of the MLP neural networks.

Lecun *et.al.* [29] tested several configurations of one-hidden layer fully connected MLP neural networks trained for handwritten digit recognition. One of them consists of directly using the original (raw) data, i.e. without tacking into account any signal pre-processing or feature selection, as input of the classifier. Thus, vectors containing 784 features corresponding to images of size 28×28 were used as inputs of the classifier. The first two rows of Table 2 shows averaged accuracy rates (Acc) yielded by MLP neural

Table 1: Best recognition rates on the test set yielded by MLP neural networks using DAS-KSVD feature vectors as well as the number of parameters required to train each one of the classifiers.

Dictionary	Classifier	Acc (%)	Number of parameters
$\Phi_D^{(5)}$	MLP-50-50-10	92.23	3,060
$\Phi_D^{(10)}$	MLP-100-100-10	95.03	11,110
$\Phi_D^{(15)}$	MLP-150-150-10	95.90	24,160
$\Phi_D^{(20)}$	MLP-200-200-10	**96.20**	42,210

networks with 300 (MLP-784-300-10) and 1000 (MLP-784-1000-10) neurons in their hidden layer. The number of parameters needed to train the networks are also included in the last column. Accuracy rates on the standard test partition of 95.3% and 95.5% were yielded by MLP neural networks with 300 and 1000 hidden neurons, respectively. It can be observed that, as a result of increasing the number of hidden neurons (from 300 to 1000), a slight improvement in the result was achieved. Also, the number of parameters of the network has increased from 238,510 to 795,010, which represent an increment of 333%.

Table 2: Average over 10 rounds of recognition rates for raw data and features derived from dictionaries learned with different methods.

Method	Classifier	Acc (%)	Number of parameters
Raw data [29]	MLP-784-300-10	**95.3**	**238,510**
	MLP-784-1000-10	95.5	795,010
DAS-KSVD	MLP-200-50-10	**95.3**	**10,560**
	MLP-200-100-10	96.1	21,110
	MLP-200-200-10	96.2	42,210
	MLP-200-300-10	96.4	63,310
	MLP-200-1000-10	**96.7**	**211,010**
KSVD [9]	MLP-200-50-10	93.5	10,560
	MLP-200-100-10	92.8	21,110
	MLP-200-200-10	92.3	42,210
	MLP-200-300-10	92.8	63,310
	MLP-200-1000-10	92.7	211,010
LC-KSVD2 [15]	MLP-200-50-10	91.8	10,560
	MLP-200-100-10	91.9	21,110
	MLP-200-200-10	92.0	42,210
	MLP-200-300-10	92.1	63,310
	MLP-200-1000-10	92.3	211,010

Table 2 also shows a comparative summary of the results yielded by MLP neural networks with a reduction in the dimension of the feature vectors. For that, the proposed DAS-KSVD method was used for obtaining feature vectors of length 200. As shown in Table 1, structured dictionaries composed by 200 discriminative atoms ($\Phi_D^{(20)}$) are the best option for handwritten digit recognition. Clearly, the use of small dimensional feature vectors produce a significant dimension reduction and therefore, the computing time required for classification is reduced. Thus, the number of input units of the MLP neural network was reduced (from 784 to 200) in 74.49% compared with those required by the original raw data. The table shows the average over 10 rounds of accuracy rates yielded by MLP neural networks with 200 input units while varying the number of hidden neurons from 50 to 1000. The last column of this table shows the required number of training parameters. Accuracy rates on the standard testing dataset of 96.4% and 96.7% were achieved by MLP neural networks with 300 (MLP-200-300-10) and 1000 (MLP-200-1000-10) hidden neurons, respectively. Additionally, the performance of MLP neural networks with 50, 100 and 200 hidden neurons were tested without showing significant improvements in the results.

It is also important to point out that the classifier MLP-200-50-10 (DAS-KSVD method) has achieved the same recognition rate (95.3%) as MLP-784-300-10 (Raw data) using a MLP neural network composed by only a 4.42% of the required parameters. It was also found that taking into account the best option that uses the original raw data as inputs of the classifier (MLP-784-1000-10), it has 795,010 training parameters while DAS-KSVD method (MLP-200-1000-10) has not only 211,010 parameters, but also increases a 1.2% in the performance of the classifier. As a result of that analysis, one might think that

the proposed DAS-KSVD method produces a significant dimension reduction while enhancing the overall recognition rate. Summing up, it was demonstrated that using the proposed DAS-KSVD method for dimension reduction undoubtedly enhances the recognition rate of MLP neural networks.

We have compared the performance of the new DAS-KSVD method with the standard KSVD method as well as with the discriminative-based LC-KSVD2 method. It can be observed from Table 2 that the proposed DAS-KSVD method outperform all the others showing robustness and effectiveness in the recognition of handwritten digits images coming from MNIST database. The maximum recognition rate yielded by the DAS-KSVD method was 96.7% which clearly outperforms those yielded by both KSVD (93.5%) and LC-KSVD2 (92.3%) methods.

We have also evaluated the statistical significance of the results presented in Table 2 by computing the probability that the DAS-KSVD method yields better recognition rates than all the other evaluated methods ($P(\epsilon_{ref} < \epsilon)$). In order to perform this test we assumed the statistical independence of the classification errors for each image and we approached the error's Binomial distribution by means of a Gaussian distribution. This is possible because we have a sufficiently high number of testing samples (10,000). In this way, for 95.5% (Raw data) and 96.7% (DAS-KSVD) we have that $P(\epsilon_{ref} < \epsilon) > 0.9999$.

6 Conclusions

In this work, both a new discriminative measure and a novel method for learning structured dictionaries for multi-class classification problems were introduced. This new measure is capable of efficiently quantifying the degree of discriminability of each one of the atoms in a particular dictionary. The use of such a measure gave rise to what we called the Discriminant Atom Selection KSVD (DAS-KSVD) method for dictionary learning. The method was tested with a widely used database for handwritten digit recognition and compared with three state-of-the-art classification methods. Experimental results showed that DAS-KSVD significantly outperforms the other three methods achieving good recognition rates and additionally, reducing the computational cost of the classifier.

Clearly, there is much further room for improvements. In particular, future research lines include the evaluation of our learning method with other well known databases, more analysis of the combined discriminative measure as well as the study of its properties and the exploration of new deep structures. Also, we plan to study the properties of the obtained learned dictionaries and also explore different ways to incorporate discriminative information into the dictionary learning process. It is also of our interest to apply these techniques to other breathing-related sleep disorder problem as well as to other problems coming from different areas such as computer vision, speech processing, etc.

7 Acknowledgments

This work was supported in part by Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, through PIP 2014-2016 Nro. 11220130100216-CO and PIP 2012-2014 Nro. 114 20110100284-KA4, by the Air Force Office of Scientific Research, AFOSR/SOARD, through Grant FA9550-14-1-0130 and by Universidad Nacional del Litoral, through project CAI+D 501 201101 00519 LI.

References

[1] R. Rolón, L. Larrateguy, L. D. Persia, R. Spies, and H. Rufiner, "Discriminative methods based on sparse representations of pulse oximetry signals for sleep apnea–hypopnea detection," *Biomedical Signal Processing and Control*, vol. 33, pp. 358–367, 2017.

[2] V. Peterson, H. L. Rufiner, and R. D. Spies, "Generalized sparse discriminant analysis for event-related potential classification," *Biomedical Signal Processing and Control*, vol. 35, pp. 70–78, 2017.

[3] L. Li, S. Li, and Y. Fu, "Learning low-rank and discriminative dictionary for image classification," *Image and Vision Computing*, vol. 32, no. 10, pp. 814–823, 2014.

[4] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, June 2008.

[5] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[6] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Transactions on Image Processing*, vol. 17, no. 1, pp. 53–69, 2008.

[7] Q. Zhang and B. Li, "Discriminative K-SVD for dictionary learning in face recognition," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2691–2698, June 2010.

[8] M. S. Lewicki and B. A. Olshausen, "Probabilistic framework for the adaptation and comparison of image codes," *Journal of the Optical Society of America A*, vol. 16, no. 7, p. 1587, 1999.

[9] M. Aharon, M. Elad, and A. Bruckstein, "KSVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, pp. 4311–4322, Nov. 2006.

[10] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[11] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Computation*, vol. 12, no. 2, pp. 337–365, 2000.

[12] K. Engan, S. O. Aase, and J. H. Husoy, "Method of optimal directions for frame design," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 2443–2446, 1999.

[13] J. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, pp. 4655–4666, Dec. 2007.

[14] D. S. Pham and S. Venkatesh, "Joint learning and dictionary construction for pattern recognition," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, June 2008.

[15] Z. Jiang, Z. Lin, and L. Davis, "Label Consistent K-SVD: Learning a discriminative dictionary for recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 2651–2664, Nov. 2013.

[16] W. Zhang, A. Surve, X. Fern, and T. Dietterich, "Learning non-redundant codebooks for classifying complex objects," pp. 1–8, ACM Press, 2009.

[17] L. Yang, R. Jin, R. Sukthankar, and F. Jurie, "Unifying discriminative visual codebook generation with classifier training for object category recognition," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, June 2008.

[18] Y. Sun, Y. Quan, and J. Fu, "Sparse coding and dictionary learning with class-specific group sparsity," *Neural Computing and Applications*, vol. 30, pp. 1265–1275, Aug. 2018.

[19] M. Elad, *Sparse and redundant representations*. Springer-Verlag New York, 2010.

[20] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[21] R. R. Coifman, Y. Meyer, S. Quake, and M. V. Wickerhauser, "Signal processing and compression with wavelet packets," in *Wavelets and Their Applications*, pp. 363–379, Springer, Dordrecht, 1994.

[22] N. Rao and F. Porikli, "A clustering approach to optimize online dictionary learning," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1293–1296, 2012.

[23] X. Chen, J. Li, D. Zou, and Q. Zhao, "Learn sparse dictionaries for edit propagation," *IEEE Transactions on Image Processing*, vol. 25, no. 4, pp. 1688–1698, 2016.

[24] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[25] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 186, no. 1007, pp. 453–461, 1946.

[26] Y. Lecun, L. D. Jackel, L. Bottou, A. Brunot, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, U. A. Muller, E. Sackinger, P. Simard, and V. Vapnik, "Comparison of learning algorithms for handwritten digit recognition," *International Conference on Artificial Neural Networks, Paris*, 1995.

[27] S. Kim, Z. Yu, R. M. Kil, and M. Lee, "Deep learning of support vector machines with class probability output networks," *Neural Networks*, vol. 64, pp. 19–28, 2015.

[28] P. d. Chazal, J. Tapson, and A. v. Schaik, "A comparison of extreme learning machines and back-propagation trained feed-forward networks processing the mnist database," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2165–2168, 2015.

[29] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[30] W. S. Russell, "Polynomial interpolation schemes for internal derivative distributions on structured grids," *Applied Numerical Mathematics*, vol. 17, no. 2, pp. 129–171, 1995.

[31] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Advances in Neural Information Processing Systems 19*, pp. 801–808, MIT Press, 2007.

[32] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *J. Mach. Learn. Res.*, vol. 11, pp. 19–60, Mar. 2010.

[33] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3360–3367, June 2010.

[34] K. Huang and S. Aviyente, "Sparse representation for signal classification," in *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, (Cambridge, MA, USA), pp. 609–616, MIT Press, 2006.

[35] S. Haykin, *Neural networks: A comprehensive foundation*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2nd ed., 1998.

[36] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

Appendices

A Grid search

The grid search method starts by dividing the interval $[0;1]$ into segments of length Δ and generating different combinations of the parameters α and β such that $\alpha + \beta \leq 1$. This constraint suggests that the boundary of the work space coincides with a right triangle whose vertexes are the pair of parameters corresponding to $(0,0)$, $(1,0)$ and $(0,1)$. Figure 5 shows an example of the grid search method for three different values of Δ. It can be observed that small values of Δ entail evaluating a large number of combinations.

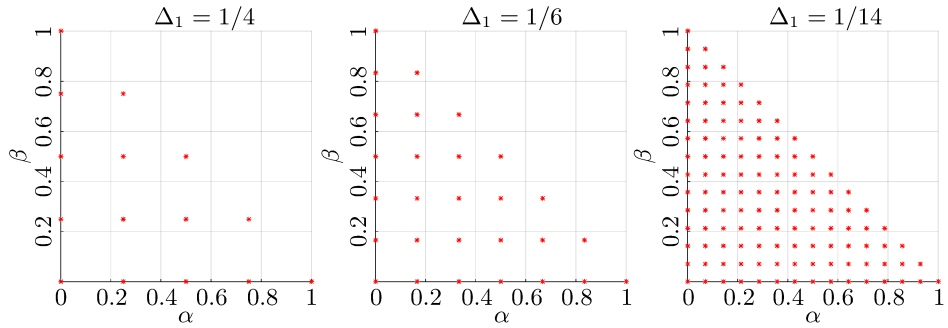


Figure 5: Different possible combinations of weights.

In order to reduce the computational cost, we have performed a grid search of the optimal pair of parameters into two stages. The first one consists of defining and using $\Delta_1 = 1/6$ in order to locate potential "regions" in the search space where recognition rates are maximized. Also, the second stage takes into account these regions and, moreover, performs a more refined search using $\Delta_2 = 1/100$. In that way, each new refined region of search is established by considering all possible pair of parameters complying with $(\alpha - \alpha^*)^2 + (\beta - \beta^*)^2 \leq (2\Delta_2)^2$ (see Figure 6). This definition coincides with all (α, β) that are inside to a close disc of radius $2\Delta_2$ centered at (α^*, β^*).

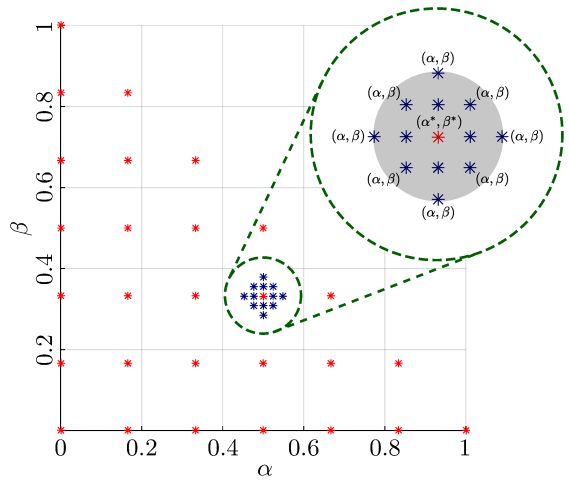


Figure 6: An example illustrating a second stage grid search.

The most discriminative atoms of Φ according to the combined measure $m_{\alpha,\beta}$ were selected and taken in for building structured dictionaries. As mentioned above, the problem of finding the optimal pair of parameters (α^*, β^*) was solved by applying the grid search method. This search was initially carried out by taking into account an interval length of $\Delta = 1/6$ which leads to 28 different pair of parameters. Figure 7 shows a summary of the results obtained by applying the grid search method for each one of the four evaluated dictionaries. In particular, we have found that using structured dictionaries comprised by more than 5 class-related discriminative atoms, the MLP neural networks achieved good recognition rates. This figure also shows, for each one of the evaluated dictionaries, two highlighted regions denoted by R_1 and R_2 where recognition rate are maximal. Among all highlighted regions, one might think that simultaneous values of α and β close to zero allow selecting the most discriminative atoms of Φ. In case

of using a structured dictionary comprised by 5 discriminative atoms for each one of the classes, we found that search regions R_1 and R_2 are centered at $(0.33, 0.17)$ and $(0.83, 0)$, respectively, and centered at $(0, 0)$ and $(0.33, 0.17)$, otherwise.

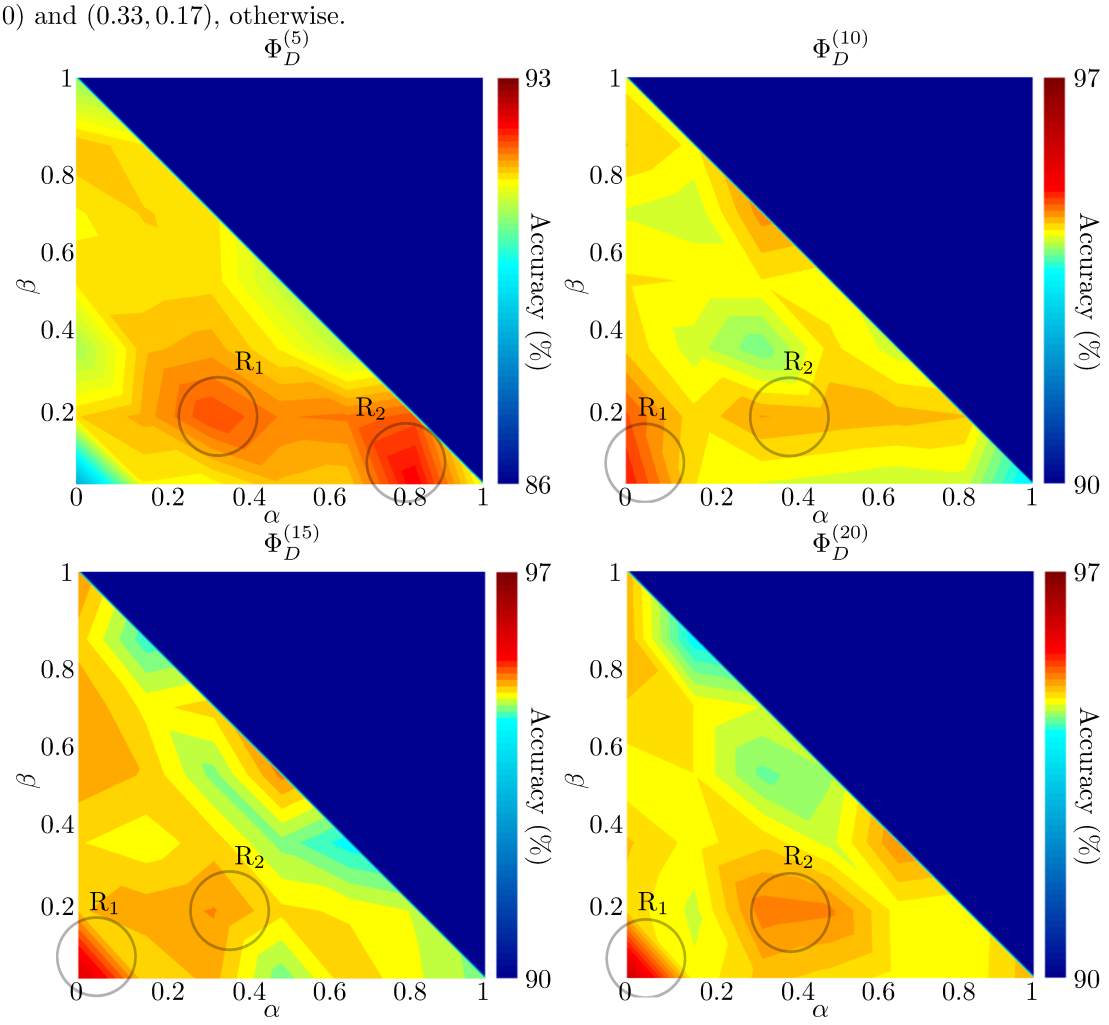


Figure 7: Fist step grid search results for each one of the evaluated dictionaries of sizes 50 (upper-left), 100 (upper-right), 150 (bottom-left) and 200 (bottom-right).

We also analyzed the overall performance (taken over 10 rounds) of the classifier for each one of the evaluated dictionaries. As it can be seen in Table 3, $\Phi_D^{(20)}$ outperforms all the others yielding the maximum (Max) recognition rate. Also, it can be seen that small structured dictionary sizes entail low classification rates. This may be due to the fact that low dimensional sparse vectors are not capable of capturing relevant information for signal classification. Otherwise, if the dimension of such vectors increases (from 100 to 200) then significant improvements are observed.

Table 3: Mean, standard deviation and maximum recognition rates obtained for each one of the evaluated dictionaries.

Dictionary	Classifier	Acc (%)	Max (%)
$\Phi_D^{(5)}$	MLP-50-50-10	91.25 ($\pm$0.67)	92.23
$\Phi_D^{(10)}$	MLP-100-100-10	94.36 ($\pm$0.27)	95.03
$\Phi_D^{(15)}$	MLP-150-150-10	94.74 ($\pm$0.27)	95.90
$\Phi_D^{(20)}$	MLP-200-200-10	94.87 ($\pm$0.33)	**96.20**

The second stage grid search method was successfully applied to each one of the tested structured dictionaries. Results have shown that, in this case, no improvements in the recognition rates were found. Thus, the optimal pair of parameters α^* and β^* are the ones found in the first stage. Figure 8 shows the results obtained by applying the refined grid search to regions R_1 (left) and R_2 (right) corresponding to the structured dictionary $\Phi_D^{(20)}$. It can be clearly seen that the values of $\alpha = 0$ and $\beta = 0$ suggest that the

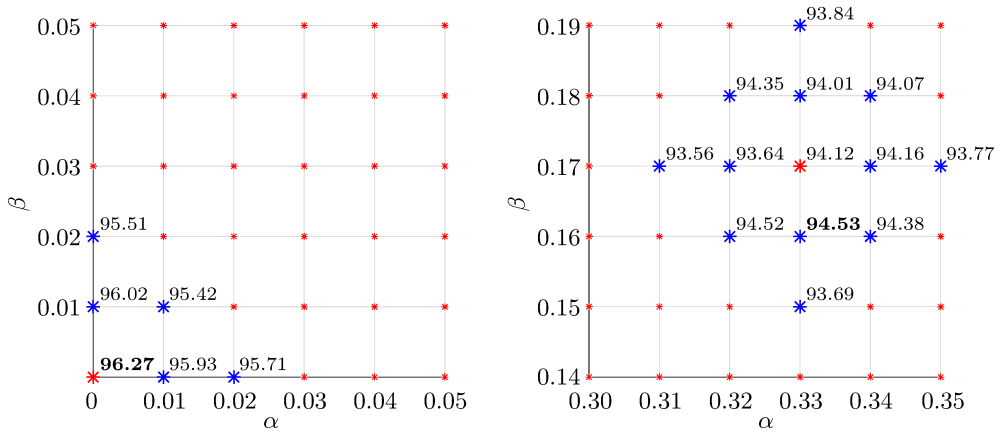


Figure 8: A second stage grid search taken into account regions R_1 (left) and R_2 (right) of $\Phi_D^{(20)}$.

most discriminative atoms of a particular dictionary Φ are not only those more frequently used for signal representation, but also the ones that minimize the total signal representation error. This imply that using only the third term of the proposed combined measure, we ensure finding the most discriminative atoms of a given dictionary.

Bibliografía

[1] Michael J. Thorpy. Classification of Sleep Disorders. *Neurotherapeutics*, 9(4):687–701, October 2012.

[2] J. Durán, S. Esnaola, R. Rubio, and A. Iztueta. Obstructive sleep apnea-hypopnea and related clinical features in a population-based sample of subjects aged 30 to 70 yr. *American Journal of Respiratory and Critical Care Medicine*, 163:685–689, March 2001.

[3] Consenso Nacional sobre el síndrome de apneas-Hipopneas del sueño (SAHS) - Resumen. *Archivos de Bronconeumología*, 41:7–9, 2005.

[4] Wojciech Kukwa, Ewa Migacz, Karolina Druc, Elzbieta Grzesiuk, and Anna M. Czarnecka. Obstructive sleep apnea and cancer: effects of intermittent hypoxia? *Future Oncology (London, England)*, 11(24):3285–3298, December 2015.

[5] Wei-Pin Chang, Mu-En Liu, Wei-Chiao Chang, Albert C. Yang, Yan-Chiou Ku, Jei-Tsung Pai, Yea-Wen Lin, and Shih-Jen Tsai. Sleep apnea and the subsequent risk of breast cancer in women:

a nationwide population-based cohort study. *Sleep Medicine*, 15(9):1016–1020, 2014.

[6] Síndrome de apnea/hipopnea obstructiva del sueño y su asociación con las enfermedades cardiovasculares. *Revista Colombiana de Cardiología*, 22.

[7] S. D. Ross, I. E. Allen, K. J. Harrison, M. Kvasz, J. Connelly, and I. A. Sheinhait. *Systematic Review of the Literature Regarding the Diagnosis of Sleep Apnea: Summary.* Agency for Healthcare Research and Quality (US), December 1998.

[8] Natasha Garg, Andrew J. Rolle, Todd A. Lee, and Bharati Prasad. Home-based diagnosis of obstructive sleep apnea in an urban population. *Journal of clinical sleep medicine: JCSM: official publication of the American Academy of Sleep Medicine*, 10(8):879–885, 2014.

[9] T. Young, M. Palta, J. Dempsey, J. Skatrud, S. Weber, and S. Badr. The occurrence of sleep-disordered breathing among middle-aged adults. *The New England Journal of Medicine*, 328(17):1230–1235, April 1993.

[10] Terry Young, Eyal Shahar, F. Javier Nieto, Susan Redline, Anne B. Newman, Daniel J. Gottlieb, Joyce A. Walsleben, Laurel Finn, Paul Enright, Jonathan M. Samet, and Sleep Heart Health Study Research Group. Predictors of sleep-disordered breathing in community-dwelling adults: the Sleep Heart Health Study. *Archives of Internal Medicine*, 162(8):893–900, April 2002.

[11] S. Redline, P. V. Tishler, T. D. Tosteson, J. Williamson, K. Kump, I. Browner, V. Ferrette, and P. Krejci. The familial aggregation of obstructive sleep apnea. *American Journal of Respiratory and Critical Care Medicine*, 151(3 Pt 1):682–687, March 1995.

[12] Anne B. Newman, Greg Foster, Rachel Givelber, F. Javier Nieto, Susan Redline, and Terry Young. Progression and regression of sleep-disordered breathing with changes in weight: the Sleep Heart Health Study. *Archives of Internal Medicine*, 165(20):2408–2413, November 2005.

[13] P. E. Peppard, T. Young, M. Palta, and J. Skatrud. Prospective study of the association between sleep-disordered breathing and hypertension. *The New England Journal of Medicine*, 342(19):1378–1384, May 2000.

[14] Mahssa Karimi, Jan Hedner, Henrike Häbel, Olle Nerman, and Ludger Grote. Sleep Apnea Related Risk of Motor Vehicle Accidents is Reduced by Continuous Positive Airway Pressure: Swedish Traffic Accident Registry Data. *Sleep*, 38(3):341–349, March 2015.

[15] Tratamiento médico del SAHS. *Archivos de Bronconeumología*, 41:43–50, 2005.

[16] J. Allan Hobson. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects: A. Rechtschaffen and A. Kales (Editors). (Public Health Service, U.S. Government Printing Office, Washington, D.C., 1968, 58

p., $4.00). *Electroencephalography and Clinical Neurophysiology*, 26(6):644, June 1969.

[17] Richard B. Berry, Rita Brooks, Charlene Gamaldo, Susan M. Harding, Robin M. Lloyd, Stuart F. Quan, Matthew T. Troester, and Bradley V. Vaughn. AASM Scoring Manual Updates for 2017 (Version 2.4). *Journal of Clinical Sleep Medicine : JCSM : Official Publication of the American Academy of Sleep Medicine*, 13(5):665–666, May 2017.

[18] H. Sala, C. Nigro, C. Rabec, A. Guardia, and M. Smurra. Consenso argentino de trastornos repiratorios vinculados al sueño. *Medicina (Buenos Aires)*, 61(3):351–363, 2001.

[19] Robert Thurnheer, Konrad E. Bloch, Irène Laube, Matthias Gugger, Markus Heitz, and Swiss Respiratory Polygraphy Registry. Respiratory polygraphy in sleep apnoea diagnosis. Report of the Swiss respiratory polygraphy registry and systematic review of the literature. *Swiss Medical Weekly*, 137(5-6):97–102, February 2007.

[20] R.E. Rolón, L.D. Larrateguy, L.E. Di Persia, R.D. Spies, and H.L. Rufiner. Discriminative methods based on sparse representations of pulse oximetry signals for sleep apnea–hypopnea detection. *Biomedical Signal Processing and Control*, 33:358–367, 2017.

[21] Thomas Penzel and AbdelKebir Sabil. The use of tracheal sounds for the diagnosis of sleep apnoea. *Breathe*, 13(2):e37–e45, June 2017.

[22] M. Deviaene, D. Testelmans, B. Buyse, P. Borzée, S. V. Huffel, and C. Varon. Automatic Screening of Sleep Apnea Patients Based on the SpO_2 Signal. *IEEE Journal of Biomedical and Health Informatics*, pages 1–1, 2018.

[23] Gastón Schlotthauer, Leandro E. Di Persia, Luis D. Larrateguy, and Diego H. Milone. Screening of obstructive sleep apnea with empirical mode decomposition of pulse oximetry. *Medical Engineering & Physics*, 36(8):1074–1080, August 2014.

[24] R. Casal and G. Schlotthauer. Sleep detection in heart rate signals from photoplethysmography. In *2017 XVII Workshop on Information Processing and Control (RPIC)*, pages 1–6, September 2017.

[25] Wu Huang, Bing Guo, Yan Shen, and Xiangdong Tang. A novel method to precisely detect apnea and hypopnea events by airflow and oximetry signals. *Computers in Biology and Medicine*, 88:32–40, 2017.

[26] Marcin Ciołek, Maciej Niedźwiecki, Stefan Sieklicki, Jacek Drozdowski, and Janusz Siebert. Automated detection of sleep apnea and hypopnea events based on robust airflow envelope tracking in the presence of breathing artifacts. *IEEE journal of biomedical and health informatics*, 19(2):418–429, 2015.

[27] C. Varon, A. Caicedo, D. Testelmans, B. Buyse, and S. Van Huffel. A Novel Algorithm for the Automatic Detection of Sleep Apnea From Single-Lead ECG. *IEEE Transactions on Biomedical Engineering*, 62(9):2269–2278, 2015.

[28] Ivan T. Ling, Alan L. James, and David R. Hillman. Interrelationships between body mass, oxygen desaturation, and apnea-hypopnea indices in a sleep clinic population. *Sleep*, 35(1):89–96, 2012.

[29] K. Sano, H. Nakano, Y. Ohnishi, Y. Ishii, T. Nakamura, K. Matuzawa, J. Maekawa, and N. Narita. [Screening of sleep apnea/hypopnea syndrome by home pulse oximetry]. *Nihon Kokyuki Gakkai Zasshi = the Journal of the Japanese Respiratory Society*, 36(11):948–952, November 1998.

[30] E. Chiner, J. Signes-Costa, J. M. Arriero, J. Marco, I. Fuentes, and A. Sergado. Nocturnal oximetry for the diagnosis of the sleep apnoea hypopnoea syndrome: a method to reduce the number of polysomnographies? *Thorax*, 54(11):968–971, November 1999.

[31] Juan-Carlos Vázquez, Willis H. Tsai, W. Ward Flemons, Akira Masuda, Rollin Brant, Eric Hajduk, William A. Whitelaw, and John E. Remmers. Automated analysis of digital oximetry in the diagnosis of obstructive sleep apnoea. *Thorax*, 55(4):302–307, April 2000.

[32] Evangelos Kaimakamis, Venetia Tsara, Charalambos Bratsas, Lazaros Sichletidis, Charalambos Karvounis, and Nikolaos Maglaveras. Evaluation of a Decision Support System for Obstructive Sleep Apnea with Nonlinear Analysis of Respiratory Signals. *PloS One*, 11(3):1–16, 2016.

[33] Trung Q. Le and Satish T. S. Bukkapatnam. Nonlinear Dynamics Forecasting of Obstructive Sleep Apnea Onsets. *PLOS ONE*, 11(11):1–12, November 2016.

[34] Shu-Han Fan, Chia-Ching Chou, Wei-Chen Chen, and Wai-Chi Fang. Real-time obstructive sleep apnea detection from frequency analysis of EDR and HRV using Lomb Periodogram. *Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, 2015:5989–5992, 2015.

[35] Ahnaf Rashik Hassan. Computer-aided obstructive sleep apnea detection using normal inverse gaussian parameters and adaptive boosting. *Biomedical Signal Processing and Control*, 29:22–30, 2016.

[36] M. O. Mendez, J. Corthout, S. Van Huffel, M. Matteucci, T. Penzel, S. Cerutti, and A. M. Bianchi. Automatic screening of obstructive sleep apnea from the ECG based on empirical mode decomposition and wavelet analysis. *Physiological Measurement*, 31(3):273–289, 2010.

[37] Harun Karamanli, Tankut Yalcinoz, Mehmet Akif Yalcinoz, and Tuba Yalcinoz. A prediction model based on artificial neural networks for the diagnosis of obstructive sleep apnea. *Sleep and Breathing*, 20(2):509–514, 2015.

[38] J.A. Tropp and A.C. Gilbert. Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007.

[39] M. Aharon, M. Elad, and A. Bruckstein. KSVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, November 2006.

[40] V. Papyan, Y. Romano, J. Sulam, and M. Elad. Theoretical Foundations of Deep Learning via Sparse Representations: A Multilayer Sparse Model and Its Connection to Convolutional Neural Networks. *IEEE Signal Processing Magazine*, 35(4):72–89, July 2018.

[41] David J. Klein, Peter König, and Konrad P. Körding. Sparse Spectrotemporal Coding of Sounds. *EURASIP Journal on Advances in Signal Processing*, 2003(7):659–667, December 2003.

[42] C.E. Martínez, J. Goddard, D.H. Milone, and H.L. Rufiner. Bioinspired sparse spectro-temporal representation of speech for robust classification. *Computer Speech and Language*, 26(5):336–348, October 2012.

[43] Christopher Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer-Verlag, New York, 2006.

[44] Stephan R Sain and V N Vapnik. The nature of statistical learning theory. *Technometrics*, 38(4):409, 1996.

[45] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, sep 2008.

[46] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Science & Business Media, aug.

[47] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1-3):489–501, 2006.

[48] Wanyu Deng, Qinghua Zheng, and Lin Chen. Regularized extreme learning machine. In *2009 IEEE Symposium on Computational Intelligence and Data Mining*, 2009.

[49] S. F. Quan, B. V. Howard, C. Iber, J. P. Kiley, F. J. Nieto, G. T. O'Connor, D. M. Rapoport, S. Redline, J. Robbins, J. M. Samet, and P. W. Wahl. The Sleep Heart Health Study: design, rationale, and methods. *Sleep*, 20(12):1077–1085, 1997.

[50] Bonnie K. Lind, James L. Goodwin, Joel G. Hill, Tauqeer Ali, Susan Redline, and Stuart F. Quan. Recruitment of healthy adults into a study of overnight sleep monitoring in the home: experience of the Sleep Heart Health Study. *Sleep & Breathing = Schlaf & Atmung*, 7(1):13–24, 2003.

[51] Stphane Mallat. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, Inc., Orlando, FL, USA, 3rd edition, 2008.

[52] J. A. Swets. ROC analysis applied to the evaluation of medical imaging techniques. *Investigative Radiology*, 14(2):109–121, April 1979.

[53] Karimollah Hajian-Tilaki. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian Journal of Internal Medicine*, 4(2):627–635, 2013.

[54] J. A. Swets. Indices of discrimination or diagnostic accuracy: their ROCs and implied models. *Psychological Bulletin*, 99(1):100–117, January 1986.

[55] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, April 1982.

[56] Rajeev Kumar and Abhaya Indrayan. Receiver operating characteristic (ROC) curve for medical researchers. *Indian Pediatrics*, 48(4):277–287, April 2011.

[57] C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[58] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[59] Naoki Saito and Ronald R. Coifman. Local discriminant bases and their applications. *Journal of Mathematical Imaging and Vision*, 5(4):337–358, 1995.

[60] W. Gersch, F. Martinelli, J. Yonemoto, M. D. Low, and J. A. Mc Ewan. Automatic classification of electroencephalograms:

Kullback-Leibler nearest neighbor rules. *Science (New York, N.Y.)*, 205(4402):193–195, 1979.

[61] Pedro J. Moreno, Purdy P. Ho, and Nuno Vasconcelos. A Kullback-Leibler Divergence Based Kernel for SVM Classification in Multimedia Applications. In S. Thrun, L. K. Saul, and P. B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 1385–1392. MIT Press, 2004.

[62] Harold Jeffreys. An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 186(1007):453–461, 1946.

[63] J. Lin. Divergence Measures Based on the Shannon Entropy. *IEEE Trans. Inf. Theor.*, 37(1):145–151, 2006.

[64] Charu C. Aggarwal. *Data classification: algorithms and applications*. CRC Press, 2014.

[65] George Francis Harpur. *Low Entropy Coding with Unsupervised Neural Networks*. PhD thesis, Department of Engineering, University of Cambridge, Queens' College, February 1997.

[66] Michael Elad. *Sparse and Redundant Representations*. Springer-Verlag New York, 2010.

[67] Federico Girosi. An Equivalence Between Sparse Approximation and Support Vector Machines. *Neural Computation*, 10(6):1455–1480, August 1998.

[68] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic Decomposition by Basis Pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1998.

[69] S. G. Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[70] Y.C. Pati, R. Rezaiifar, and P.S. Krishnaprasad. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In *Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, pages 40–44, November 1993.

[71] Ronald R. Coifman, Yves Meyer, Steven Quake, and M. Victor Wickerhauser. Signal processing and compression with wavelet packets. In *Wavelets and Their Applications*, pages 363–379. Springer, Dordrecht, 1994.

[72] Michael S. Lewicki and Bruno A. Olshausen. Probabilistic framework for the adaptation and comparison of image codes. *Journal of the Optical Society of America A*, 16(7):1587, 1999.

[73] M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12(2):337–365, February 2000.

[74] K. Engan, S. O. Aase, and J. Hakon Husoy. Method of optimal directions for frame design. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 2443–2446, 1999.

[75] Ke Huang and Selin Aviyente. Sparse representation for signal classification. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06, pages 609–616, Cambridge, MA, USA, 2006. MIT Press.

[76] Q. Zhang and B. Li. Discriminative K-SVD for dictionary learning in face recognition. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2691–2698, June 2010.

[77] D. S. Pham and S. Venkatesh. Joint learning and dictionary construction for pattern recognition. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[78] Zhuolin Jiang, Zhe Lin, and L.S. Davis. Label Consistent K-SVD: Learning a discriminative dictionary for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2651–2664, November 2013.

[79] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[80] Wei Zhang, Akshat Surve, Xiaoli Fern, and Thomas Dietterich. Learning non-redundant codebooks for classifying complex objects. pages 1–8. ACM Press, 2009.

[81] L. Yang, R. Jin, R. Sukthankar, and F. Jurie. Unifying discriminative visual codebook generation with classifier training for

object category recognition. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.

[82] Yuping Sun, Yuhui Quan, and Jia Fu. Sparse coding and dictionary learning with class-specific group sparsity. *Neural Computing and Applications*, 30(4):1265–1275, August 2018.

[83] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y. Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems 19*, pages 801–808. MIT Press, 2007.

[84] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *J. Mach. Learn. Res.*, 11:19–60, March 2010.

[85] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3360–3367, June 2010.

[86] Yann Lecun, L. D. Jackel, Leon Bottou, A. Brunot, Corinna Cortes, J. S. Denker, Harris Drucker, I. Guyon, U. A. Muller, Eduard Sackinger, Patrice Simard, and V. Vapnik. Comparison of learning algorithms for handwritten digit recognition. *International Conference on Artificial Neural Networks, Paris*, 1995.

[87] P. Bhattacharjee, S. Banerjee, M. Gulati, A. Majumdar, and S. S. Ram. Supervised analysis dictionary learning: Application in consumer electronics appliance classification. In *Proceedings of the*

Fourth ACM IKDD Conferences on Data Sciences, CODS '17, pages 2:1–2:10, New York, NY, USA, 2017. ACM.

[88] Sangwook Kim, Zhibin Yu, Rhee Man Kil, and Minho Lee. Deep learning of support vector machines with class probability output networks. *Neural Networks*, 64:19–28, 2015.

[89] P. de Chazal, J. Tapson, and A. van Schaik. A comparison of extreme learning machines and back-propagation trained feed-forward networks processing the mnist database. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2165–2168, 2015.

[90] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[91] Facundo Nogueira, Carlos Nigro, Hugo Cambursano, Eduardo Borsini, Julio Silio, and Jorge Ávila. Guías prácticas de diagnóstico y tratamiento del síndrome de apneas e hipopneas obstructivas del sueño. *Medicina (Buenos Aires)*, 73(4):349–362, August 2013.

[92] Laurens van der Maaten, E.O. Postma, and H.J. van den Herik. Dimensionality reduction: A comparative review. In *Tilburg University Technical Report*, pages TiCC–TR 2009–005, 2009.

[93] Bijoy Laxmi Koley and Debangshu Dey. Automatic detection of sleep apnea and hypopnea events from single channel measurement of respiration signal employing ensemble binary SVM classifiers. *Measurement*, 46(7):2082–2092, 2013.

[94] Cristina Garcia-Cardona and Brendt Wohlberg. Convolutional Dictionary Learning: A Comparative Review and New Algorithms. *IEEE Transactions on Computational Imaging*, 4(3):366–381, September 2018. arXiv: 1709.02893.

[95] J. Yang, Z. Wang, Z. Lin, S. Cohen, and T. Huang. Coupled Dictionary Training for Image Super-Resolution. *IEEE Transactions on Image Processing*, 21(8):3467–3478, August 2012.

Doctorado en Ingeniería
mención Inteligencia Computacional, Señales y Sistemas

Título de la obra:
Algoritmos avanzados para la detección del síndrome de apnea-hipopnea obstructiva del sueño

Autor: Román Emanuel Rolón

Lugar: Santa Fe, Argentina

Palabras Clave:

Síndrome de apnea,
Oximetría de pulso,
Procesamiento de señales,
Representaciones ralas,
Reconocimiento de patrones,
Aprendizaje maquinal,
Problemas inversos,
Medidas de discriminabilidad.