UNIVERSIDAD NACIONAL DEL LITORAL

DOCTORADO EN INGENIERÍA

Índices de validación para algoritmos de agrupamiento

David Nazareno Campo

FICH
FACULTAD DE INGENIERÍA
Y CIENCIAS HÍDRICAS

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

INTEC
INSTITUTO DE DESARROLLO TECNOLÓGICO
PARA LA INDUSTRIA QUÍMICA

CIMEC
CENTRO DE INVESTIGACIÓN DE
MÉTODOS COMPUTACIONALES

Tesis de Doctorado **2019**

Doctorado en Ingeniería
Mención en Inteligencia Computacional, Señales y Sistemas

Título de la obra:

Índices de validación para algoritmos de agrupamiento

Autor: David Nazareno Campo

Lugar: Santa Fe, Argentina

Palabras Claves:

grupos solapados,
índices de validación, validación externa,
perturbación de grupos.

UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas
Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional

ÍNDICES DE VALIDACIÓN PARA ALGORITMOS DE AGRUPAMIENTO

David Nazareno Campo

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

2019

Comisión 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 Investigación en Señales, Sistemas e Inteligencia Computacional

ÍNDICES DE VALIDACIÓN PARA ALGORITMOS DE AGRUPAMIENTO

David Nazareno Campo

Lugar de Trabajo:
sinc(i)

Instituto de Señales, Sistemas e Inteligencia Computacional
Facultad de Ingeniería y Ciencias Hídricas
Universidad Nacional del Litoral

Director:
Dra. Georgina Stegmayer — **sinc**(i)-CONICET-UNL

Co-director:
Dr. Diego Humberto Milone — **sinc**(i)-CONICET-UNL

Jurado Evaluador:
Dr. Leandro Daniel Vignolo — **sinc**(i)-CONICET-UNL
Dr. Omar Chiotti — INGAR-CONICET-UTN
Dr. Javier Ivan Murillo — CIFASIS-CONICET
Dra. Silvia Schiaffino — ISISTAN-CONICET-UNICEN

2019

UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas

Santa Fe, 29 de Abril de 2019.

Como miembros del Jurado Evaluador de la Tesis de Doctorado en Ingeniería titulada ***"Índices de validación para algoritmos de agrupamiento"***, desarrollada por el Ing. David Nazareno CAMPO, 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. Omar Chiotti	Dr. Javier Murillo
.....	.....
Dra. Silvia Schiaffino (*)	Dr. Leandro Vignolo

Santa Fe, 29 de Abril 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. Diego Milone	Dra. Georgina Stegmayer
Codirector de Tesis	Directora de Tesis

(*) La Dra Silvia Schiaffino participó por videoconferencia

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

DECLARACIÓN LEGAL DEL AUTOR

Esta Tesis ha sido remitida como parte de los requisitos para la obtención del grado académico de Doctor en Ingeniería ante la Universidad Nacional del Litoral y ha sido depositada en la Biblioteca de la Facultad de Ingeniería y Ciencias Hídricas para que esté a disposición de sus lectores bajo las condiciones estipuladas por el reglamento de la mencionada Biblioteca.

Citaciones breves de esta Tesis son permitidas sin la necesidad de un permiso especial, en la suposición de que la fuente sea correctamente citada. Solicitudes de permiso para la citación extendida o para la reproducción parcial o total de ese manuscrito serán concebidos por el portador legal del derecho de propiedad intelectual de la obra, por medio escrito.

TESIS POR COMPILACIÓN

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."

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Resumen

Sin lugar a dudas la información generada por el ser humano, dirigida por las nuevas tecnologías tales como mensajes en redes sociales y conectividad constante, viene creciendo a pasos agigantados. Dicho volúmen de información no es útil si no se lo procesa o analiza de alguna manera inteligente. Una de las maneras de analizar grandes volúmenes de datos es a través de algoritmos de agrupamiento. Con la ayuda de los mismos es posible indagar sobre la estructura subyacente de los datos. Es en casos donde no se conoce la estructura del conjutno de datos donde se hace más importante contar con herramientas que permitan valorar la solución provista por dichos algoritmos de clustering. Muchas de estas herramientas se materializan a través de índices de calidad.

En esta tesis se hace foco en un tipo particular de índice: los de validación externa. Con la ayuda de los mismos es posible comparar distintas soluciones de agrupamientos entre sí, o comparar una solución con algún objetivo externo impuesto como referencia o "norma de oro". Éstos índices fueron evolucionando a lo largo del tiempo a la par de los algoritmos de agrupamiento. Sin embargo, empezó a notarse una brecha entre los algoritmos, cada vez más complejos, y los índices que los medían. En este estudio se ve cómo en los algoritmos de clustering que proveen soluciones solapadas, no es posible medir la calidad de éstas a partir de los índices clásicos que se mencionan en la literatura. Se demuestra cómo los mismos fallan ante tales situaciones y se abre camino para el desarrollo intuitivo de un nuevo índice que permita manejar dicho escenario. Se propone un nuevo índice a partir de un desarrollo basado en estimadores de probabilidad por ocurrencia de pares de objetos en los mismos grupos. Este índice permite medir tanto soluciones complejas con clusters solapados cómo las más tradicionales sin solapamiento. A través de distintos experimentos, con datos artificiales y reales, se muestra cómo el nuevo índice es capaz de medir adecuadamente la calidad.

Abstract

Without any doubt, information generated by humanity and driven by an ubiquitous new technology, social networks and ongoing connectivity, has been growing at an unprecedented rate. Such quantity of information is useless without a proper and systematic analysis through an intelligent way. A clever approach to analyze such a volume of data is through clustering algorithms. With such help on processing it is possible to discover internal properties of the data. Moreover, in most of the cases, where the relying structure of the data is mainly unknown, having tools that help to assess the quality of algorithms and its results is of primary importance. A large portion of these tools are validation indices.

In this thesis a particular type of indexes, those of external validation, are studied. With external indexes different solutions can be compared beetwen themselves or with a solution impossed arbitrary as a gold standard. With the advancement of clustering algorithms, these indexes have made progress. However, at certain point, algorithms became more complex and indexes became insufficient to fulfill the gap that algorithms created. This study shows how some solutions, called overlapped, provided by certain algorithms can not be measured properly by classical indexes mentioned in the literature. It is shown how those indexes fail, motivating the development of a new type of index capable of assessing such solutions. A new index based on probabilistic estimations is proposed. The idea behind the new index is to measure the probability that two objects are grouped together in both solutions. This new index allows measuring not only classical solutions but also overlapped ones. Throughout a wide set of experiments, with artificial and real datasets, it is shown how the new index is able to correctly measure the quality of solutions.

Introducción

Los algoritmos de agrupamiento, o clustering en inglés, particionan un conjunto de datos (patrones, casos, observaciones o instancias) de modo no supervisado, en un cierto número de grupos o clusters. Un cluster puede definirse básicamente como un conjunto de objetos que son similares entre sí pero distintos a otros objetos contenidos en otros clusters [1, 2]. Entre las técnicas de clustering más utilizadas hoy en día se pueden mencionar el agrupamiento jerárquico [3], k–medias [4] y mapas autoorganizativos (SOM, por su sigla en Inglés) [5, 6].

Una de las particularidades de los métodos de agrupamiento es que siempre encuentran grupos, incluso cuando éstos no existen, y por ello en los últimos años ha surgido lo que se se denominó análisis de soluciones de agrupamiento [7]. Además, no es claro qué tan estables son los resultados obtenidos al variar los parámetros de los métodos de entrenamiento (por ejemplo, el simple número de grupos). Otro punto a considerar es la coherencia y validez de los agrupamientos encontrados cuando los datos son perturbados, por ejemplo por el agregado de cierto nivel de ruido a las mediciones originales o por solapamiento de clusters.

Todo esto habla de la necesidad de contar con métodos computacionales para poder determinar objetivamente la calidad y robustez de los grupos encontrados. Una propuesta para realizar este tipo de análisis consiste en medir el promedio de las distancias entre todos los agrupamientos encontrados bajo algún tipo de perturbaciones de los datos [8], para lo cual se requiere una medida de comparación entre agrupaciones. Dentro de este tipo de enfoque, se toma como referencia el agrupamiento obtenido a partir de todos los datos disponibles (el conjunto completo). Luego se toma una submuestra del conjunto total de datos y se aplica el algoritmo de clustering sobre dicha submuestra, midiendo el grado de similaridad entre las soluciones. Otro enfoque posible es variar los parámetros del algoritmo de clustering sobre los mismos datos originales y repetir su aplicación varias veces sobre la muestra original y/o las submuestras. En todos los casos, para cada agrupamiento encontrado se debe calcular la similaridad con el agrupamiento de referencia. Se supone que si la estructura de los datos está bien representada, la partición de la muestra original será muy similar a las particiones encontradas en las submuestras [7].

Para evaluar la bondad de las soluciones de clustering se utilizan dos tipos de medidas de validación. Las medidas internas miden básicamente la homogeneidad y separación de los clusters. En cambio, las externas miden

los resultados de acuerdo al conocimiento de las clases o agrupaciones correctas para los datos bajo análisis [2, 9, 10]. Las propuestas de medidas de comparación de soluciones que pueden encontrarse en la literatura reciente se dividen en tres grandes grupos: i) las basadas en el conteo de a pares de datos conjuntamente agrupados en los cuales las soluciones de clustering coinciden; ii) las basadas en el análisis de coincidencias entre conjuntos; y iii) las basadas en la estadística y la teoría de la información [7, 11, 12]. Sin embargo, los métodos de validación externa poseen la limitación que en la mayoría de los casos de interés práctico no se cuenta con la información acerca de las clases reales en los datos. Aún en los métodos de validación interna, es complejo poder indicar con claridad una única medida objetiva capaz de evaluar la calidad de los clusters, y de este modo proporcionar grupos relevantes para ser analizados con el fin de descubrir nuevas relaciones entre los datos agrupados.

Más aún, en la actualidad, existe un auge importante en el procesamiento de información y se conoce que en varios escenarios prácticos la información creada a través de, por ejemplo: redes sociales, noticias, grupos de colaboración y otros medios de Internet, es generalmente de naturaleza solapada. Es por ello que recientemente han surgido importantes aplicaciones y teorías que se orientan al análisis de soluciones con clusters solapados. Por ejemplo, en [13] se compara la evolución de grupos de personas en redes sociales y se propone un algoritmo para computar nuevas distancias entre colecciones de grupos potencialmente solapados. Esto quiere decir que los datos pueden ser particionados de diferentes modos, todos válidos, dependiendo del punto de vista o criterio para la partición, o inclusive para un mismo criterio, podrían pertenecer a más de un grupo o cluster. Al particionar estos tipos de datos, no sólo los algoritmos deberían poder considerar grupos con cierto grado de solapamiento, sino también los índices de validación deberían poder medir y cuantificarlos adecuadamente.

Recientemente, con el auge de este tipo de datos en las redes sociales y de colaboración, se han propuesto nuevos algoritmos de detección de grupos solapados [14, 15, 16, 17, 18, 19, 20]. Sin embargo, y si bien existe una cantidad considerable de índices de validación externa para clusters no solapados [21, 22], se plantea actualmente la necesidad del desarrollo de uno que además pueda efectivamente medir clusters solapados.

En esta tesis se realizó un análisis detallado de los índices existentes, mostrando las fallas que presentan ante soluciones con clusters solapados. A partir de esta revisión, se aplicó un enfoque probabilístico para desarrollar un nuevo índice que puede ser aplicable tanto a soluciones no solapadas como a soluciones solapadas. La propuesta se basó en la idea de que las estimaciones de las probabilidades de encontrar dos o más objetos juntos en una u otra solución deben ser ajustadas teniendo en cuenta que hay objetos

que pertenecen a más de un cluster en ambas soluciones [23, 24].

1.1. Objetivo general

El objetivo general de esta tesis fue desarrollar un nuevo índice que permita analizar la calidad de las soluciones generadas con métodos de clustering, teniendo particular interés en soluciones que posean clusters solapados.

1.2. Objetivos específicos

A continuación se detallan los objetivos específicos que se lograron:

- Estudio de los índices comúnmente usados para validación en clustering y analizar sus limitaciones.
- Desarrollo de un nuevo índice de validación externa que es capaz de medir la calidad de los agrupamientos en los casos en los que pueda haber grupos solapados.
- Evaluación del índice de validación externa propuesto en casos en los que los datos hayan sido perturbados, y en relación a los índices clásicos en situaciones con y sin solapamiento.
- Validación de los resultados obtenidos comparando índices sobre diferentes conjuntos de datos, simulados y reales.

1.3. Metodología general

La metodología general aplicada en el desarrollo de esta investigación consistió principalmente de ciclos que comprendieron las siguientes etapas:

- Actualización bibliográfica: estudio profundo de los conceptos y análisis de trabajos relacionados con el tema, de forma continua durante todo el desarrollo de la tesis.
- Investigación exploratoria: análisis crítico de las principales propuestas disponibles actualmente para validación en clustering, que permitieron ajustar la definición del problema y hacer propuestas de solución.
- Desarrollo de un nuevo índice para la validación de soluciones de clustering.
- Implementación de los índices clásicos y el propuesto.
- Prueba de los métodos propuestos en esta tesis con datos sintéticos y en condiciones totalmente controladas.
- Diseño y experimentación numérica: prueba de los métodos propuestos, y comparación con los resultados obtenidos con otras técnicas del estado del arte aplicables al mismo conjunto de datos reales.

- Publicación de resultados: a través de informes periódicos, reportes sobre los resultados obtenidos en las etapas anteriores y publicaciones científicas.

Estas actividades no representan un lineamiento rígido ni establecen un órden único, y además las actividades de Desarrollo, Implementación y Diseño y experimentación se realizaron de forma repetida e incremental durante gran parte del desarrollo de esta tesis.

Nuevo índice propuesto

En esta sección se describe un nuevo índice para medición de estabilidad de soluciones de clustering que puede aplicarse tanto a soluciones con o sin grupos (clusters) solapados. Antes de pasar al desarrollo del mismo se procederá a la aplicación de los índices clásicos a un ejemplo sencillo, para mostrar cómo estos fallan ante soluciones con solapamiento.

En la Figura 2.1 se observan dos soluciones. En a), se ve cómo en la solución llamada C todos los objetos pertenecen a un mismo grupo. En b), se ve cómo en la solución llamada C' se observan dos grupos: uno posee todos los objetos y el otro posee todos los mismos objetos, excepto uno. En dicha solución, todos los objetos se encuentran solapados menos uno. Si se tiene en cuenta que cualquier posible par de objetos se puede encontrar agrupado junto en ambas soluciones, se podría esperar un valor muy cercano a 1 para cualquier índice de validación externa entre las soluciones a) y b). Sin embargo, al aplicar índices clásicos como Fowlkes and Mallows (FM)[25], Adjusted Rand Index (ARI)[26] y Jaccard (JAC)[27] se obtienen valores no esperados.

Para aplicar dichos índices al ejemplo de la Figura 2.1 se procede a definir la matriz de contingencia, necesaria para la aplicación de la fórmula de los mismos. La misma permite visualizar las coincidencias de los grupos de cada solución cuando se comparan de a pares. Por ejemplo, en la Tabla 2.1 se visualiza la matriz de contingencia del ejemplo de la Figura 2.1. En las filas se colocan los grupos de una solución (de referencia) y en las columnas los grupos de otra solución a evaluar. Por cada intersección se contabilizan los pares que hay en común entre el grupo de la solución de referencia con el grupo correspondiente de la solución a evaluar. Finalmente, en la última columna se sumarizan los elementos correspondientes a cada grupo de la solución de referencia. De la misma manera sucede para los elementos de la última fila y la solución a evaluar. Por último, la intersección de la última fila con la última columna sumariza la cantidad total de elementos.

El índice FM se define como

$$FM = \frac{T_k}{\sqrt{P_k Q_k}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k'} m_{ij}^2 - N}{\sqrt{\left(\sum_{i=1}^{k} m_{i*}^2 - N\right)\left(\sum_{j=1}^{k'} m_{*j}^2 - N\right)}}, \tag{2.1}$$

donde k es la cantidad de grupos de la solución de referencia, k' es la cantidad de grupos de la solución a evaluar, m_{ij} corresponde al elemen-

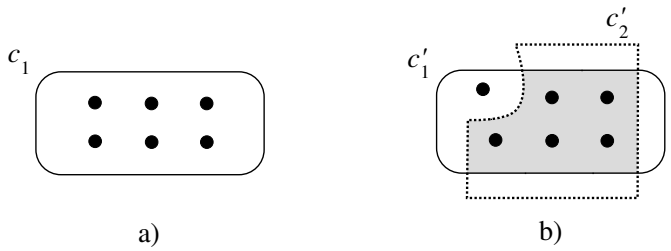


Figura 2.1: Ejemplo en el cual se observan dos soluciones con $N = 6$ elementos a agrupar: a) solución C con $k = 1$ y b) solución C' con $k' = 2$ y solapamiento.

Tabla 2.1: Matriz de contingencia correspondiente al ejemplo de la Figura 2.1

	c'_1	c'_2	Σ
c_1	6	5	11
Σ	6	5	11

to de la fila i y columna j de la matriz de contingencia, m_{i*} representa la sumatoria de la $i - ésima$ fila, m_{*j} representa la sumatoria de la $j - ésima$ columna y N representa la cantidad total de patrones a agrupar. Dichas definiciones se utilizan en el cálculo de todos los índices, teniendo el mismo significado en cada uno de ellos. Para calcular el índice FM sobre el ejemplo de la Figura 2.1 se procederá a calcular cada uno de sus factores. El valor del factor P_k se corresponde con la solución C y el de Q_k con la C'. Así $P_k = \sum_{i=1}^{1}(m_{i*}^2 - N) = 11^2 - 6 = 115$ y $Q_k = \sum_{j=1}^{2}(m_{*j}^2 - N) = 6^2 + 5^2 - 6 = 55$. Para el cálculo de T_k debemos obtener las intersecciones de los objetos del cluster de la solución a) con los de la b). Así, según la matriz de contingencia de la Tabla 2.1 obtenemos $T_k = \sum_{i=1}^{1}\sum_{j=1}^{2}(m_{ij}^2 - N) = 6^2 + 5^2 - 6 = 55$. De esta forma $FM = 0{,}692$,

El índice ARI es una versión ajustada del índice de Rand [28], considerando que su valor esperado para dos agrupamientos aleatorios no toma valor constante 0. En [26] se desarrolla el ARI y se comentan las mejoras y problemas que soluciona del índice de Rand. En dicho trabajo el ARI se define como

$$ARI = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k'}\binom{m_{ij}}{2} - \left[\sum_{i=1}^{k}\binom{m_{i*}}{2}\sum_{j=1}^{k'}\binom{m_{*j}}{2}\right]\Big/\binom{N}{2}}{\frac{1}{2}\left[\sum_{i=1}^{k}\binom{m_{i*}}{2} + \sum_{j=1}^{k'}\binom{m_{*j}}{2}\right] - \left[\sum_{i=1}^{k}\binom{m_{i*}}{2}\sum_{j=1}^{k'}\binom{m_{*j}}{2}\right]\Big/\binom{N}{2}}. \tag{2.2}$$

Aplicando esta ecuación al ejemplo de la Figura 2.1 se obtiene un ARI $= \frac{25-91{,}667}{40-91{,}667} = 1{,}290$.

Por último, el índice de Jaccard se define como

$$JAC = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k'} m_{ij}^2 - N}{\left(\sum_{i=1}^{k} m_{i*}^2 - N\right) + \left(\sum_{j=1}^{k'} m_{*j}^2 - N\right) - \left(\sum_{i=1}^{k}\sum_{j=1}^{k'} m_{ij}^2 - N\right)}. \tag{2.3}$$

Al igual que con los índices anteriores, se puede aplicar al ejemplo de la Figura 2.1 y se obtiene $JAC = \frac{55}{115+55-55} = 0{,}478$

Como se mencionó anteriormente, considerando el ejemplo de la Figura 2.1, es esperable que un índice devuelva un valor cercano a 1, dado que las soluciones son similares en cuanto a que cualquier par de patrones encontrado en una solución puede encontrarse en otra. Como se puede observar con los índices clásicos, ninguno se acerca a este valor esperado. Por un lado FM y JAC devuelven valores muy por debajo de lo esperado. Por otro lado, ARI devuelve un valor que está por encima de 1 (fuera del rango definido, $[-1, 1]$). Con este simple ejemplo queda claro que los índices clásicos no son capaces de medir soluciones que tengan algún tipo de solapamiento.

Motivado por el comportamiento no esperado de los índices clásicos de validación externa, el índice propuesto se desarrolló a partir de un cuidadoso análisis de los mismos en cuanto a su funcionamiento y limitaciones. El nuevo índice propuesto toma en cuenta la probabilidad de encontrar dos objetos cualesquiera agrupados juntos tanto en cada una de las soluciones como en ambas a la vez. Antes de detallar los elementos del nuevo índice, se introducirá la notación utilizada, así como algunas definiciones básicas. Cuando se consideren los datos que han sido agrupados, en cualquiera de las dos soluciones comparadas, al igual que con los índices clásicos, se denotará con N a la cantidad de objetos que se vayan a particionar en grupos. Se llamará c_i al grupo i de la la solución de referencia y k será la cantidad de grupos de dicha solución.

Suponiendo que cada uno de los N elementos tiene la misma probabilidad de pertenecer a un cluster cualquiera c_i de la solución de referencia, la misma se puede estimar para dos elementos $\mathbf{s_x}$ y $\mathbf{s_y}$ arbitrarios como

$$Pr((\mathbf{s}_x, \mathbf{s}_y) \in c_i) = \frac{\binom{|c_i|}{2}}{\binom{N}{2}} = \frac{|c_i|\,(|c_i|-1)}{N\,(N-1)}, \tag{2.4}$$

siendo $|c_i|$ el número de elementos agrupados bajo el grupo c_i. De esta forma, el numerador de (2.4) representa la cantidad de formas según las que se pueden agrupar de a pares $|c_i|$ elementos. Si a dicho valor lo normalizamos considerando la cantidad de formas que se pueden tomar N objetos de a 2 a la vez, obtenemos el denominador de dicha ecuación. De esta forma, el denominador representa la situación límite en la que absolutamente todos los

objetos bajo agrupamiento terminen juntos en un único grupo. Continuando con el análisis y aplicando el mismo razonamiento a los demás grupos de la misma solución se obtiene

$$\tilde{p} = \frac{\sum_{i=1}^{k} \binom{|c_i|}{2}}{k\binom{N}{2}} \tag{2.5}$$

que estima la probabilidad de que cualesquiera dos elementos se encuentren juntos en alguno (o varios) de los k grupos de referencia. Es decir, en el numerador se cuentan los posibles agrupamientos de cada uno de los grupos y el factor k del denominador contempla la situación límite en la que todos los grupos se encuentran solapados. Siguiendo un razonamiento análogo se puede arribar a una expresión similar para la probabilidad estimada del agrupamiento comparativo. De esta forma,

$$\tilde{p}' = \frac{\sum_{j=1}^{k'} \binom{|c'_j|}{2}}{k'\binom{N}{2}}. \tag{2.6}$$

estima la probabilidad de encontrar dos elementos cualesquiera en alguno de los k' grupos de la solución C'. Extendiendo este enfoque para considerar ambas soluciones a la vez, se puede arribar a la expresión

$$Pr((\mathbf{s}_x, \mathbf{s}_y) \in c_i \wedge (\mathbf{s}_x, \mathbf{s}_y) \in c'_j) = \frac{\binom{|c_i \cap c'_j|}{2}}{\binom{N}{2}} = \frac{\left|c_i \cap c'_j\right| \left(\left|c_i \cap c'_j\right| - 1\right)}{N\,(N-1)}, \tag{2.7}$$

que aproxima la probabilidad de que el par de patrones $(\mathbf{s_x}, \mathbf{s_y})$ se encuentren agrupados juntos tanto en un grupo arbitrario de la solución C como en alguno de la solución C'. Si se consideran ahora a todos los grupos de ambas soluciones se arriba a

$$\tilde{t} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{k'} \binom{|c_i \cap c'_j|}{2}}{\binom{N}{2} \frac{\text{máx}(n,n')}{N} \text{mín}(k,k')}, \tag{2.8}$$

donde n y n' representan el número de objetos que pueden contarse en las soluciones C y C', contando a los elementos solapados tantas veces como aparezcan en distintos grupos.

El numerador de dicha ecuación, al igual que como sucede con las ecuaciones (2.5) y (2.6), contabiliza los pares efectivos que se encuentran juntos en ambas soluciones a la vez. A su vez, el denominador nuevamente actúa como un factor de normalización, teniendo en consideración el caso extremo

en el que todos los objetos pueden agruparse juntos varias veces al mismo tiempo, a causa del solapamiento. Nuevamente, $\binom{N}{2}$ representa la cantidad de pares que pueden agruparse juntos con N elementos. Considerando los solapamientos en ambas soluciones, debería aplicarse un factor multiplicativo que permita representar la situación posible de varios solapamientos. Por un lado podrían encontrarse a lo sumo k solapamientos en C y k' en C', pero por otro lado hay que tener en cuenta que la coincidencia de grupos entre ambas soluciones puede causar a lo sumo mín(k,k') pares de grupos. Por último $\frac{\text{máx}(n,n')}{N}$ contabiliza el promedio de objetos que pueden encontrarse en ambas soluciones considerando solapamientos.

Con dichos elementos, puede darse forma al índice propuesto, denominado $\mathcal{OC}$, definido como la razón entre la probabilidad de encontrar un par de elementos juntos en ambas soluciones a la vez, y normalizado por el máximo de las probabilidades de encontrar dichos elementos juntos en una u otra solución. Es decir,

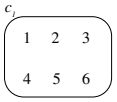
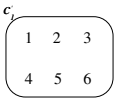
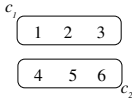
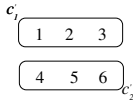
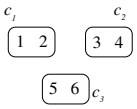
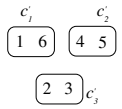
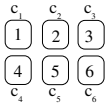
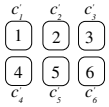
$$\mathcal{OC} = \frac{\tilde{t}}{\text{máx}(\tilde{p}, \tilde{p}')}. \tag{2.9}$$

Cuando el nuevo índice se aplica al ejemplo de la Figura 2.1, se obtienen los valores de $\tilde{p} = 1$ usando (2.5), $\tilde{p}' = 0{,}833$ cuando se utiliza (2.6) y $\tilde{t} = 0{,}909$ cuando se aplica (2.8). Finalmente, al aplicar la ecuación (2.9), se arriba a un valor de $\mathcal{OC} = 0{,}909/\,\text{máx}(1, 0{,}833) = 0{,}909$. Este valor se acerca más a la representación en la similitud de ambas soluciones de la Figura 2.1 b). Si se repite el caso de la Figura 2.1 pero llevando al extremo de 1000 objetos en vez de 6, se obtienen valores de FM, ARI y JAC de 0,707, 1,2 y 0,5, respectivamente; mientras que $\mathcal{OC}$ arroja un valor de 0,999, o sea, mucho más cercano a lo que se espera intuitivamente.

A continuación se muestra la aplicación del nuevo índice, junto con otros índices clásicos, y se comprueba su funcionamiento en casos triviales pero extremos. Con ello se muestra como es esperable que el índice $\mathcal{OC}$ funcione en casos solapados y los clásicos fallen. Las Tablas 2.2, 2.3 y 2.4 tienen la misma estructura: la primera columna sirve de enumeración del ejemplo; las columnas 2 y 3 representan los agrupamientos a comparar, tanto de la solución de referencia como la que se desea evaluar; y las columnas 4 a 7 representan los valores de los índices FM, ARI, Jaccard y $\mathcal{OC}$, respectivamente, para las soluciones de las columnas 2 y 3. En la Tabla 2.2 hay 4 ejemplos artificiales pensados para mostrar el funcionamiento del nuevo índice propuesto.

En cada una de las soluciones se agruparon 6 objetos. En los ejemplos I y II tenemos dos soluciones que son exactamente iguales entre sí. En el ejemplo I ambas soluciones agrupan todos los objetos en un único grupo.

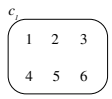
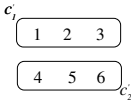
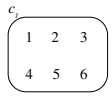
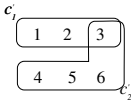
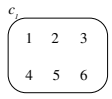
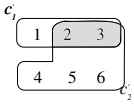
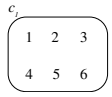
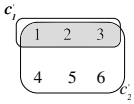
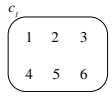
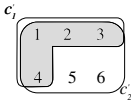
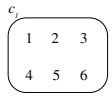
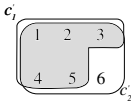
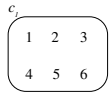
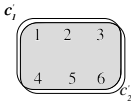
Tabla 2.2: Resultado de índices aplicados a ejemplos artificiales extremos

	Soluciones		Índices			
	C	C'	**FM**	**ARI**	**JAC**	$\mathcal{OC}$
I			1.000	—	1.000	1.000
II			1.000	1.000	1.000	1.000
III			0.000	-0.250	0.000	0.000
IV			0.000	—	—	0.000

En el ejemplo II, ambas soluciones dividen los objetos en dos grupos. Intuitivamente podría esperarse que los índices reporten una similitud total, de decir de 1,000, en ambos ejemplos ya que todos los objetos que podemos encontrar agrupados juntos en una solución se encuentran agrupados juntos en la otra. Como puede observarse, en ambos ejemplos se observa un valor idéntico de 1,000 para todos los índices, excepto ARI que no produce ningún valor en en el ejemplo I. Ésto es debido a que en la definición del índice en [26] el valor esperado y el máximo dan iguales; produciendo de esta manera una división por cero.

En los ejemplos III y IV se tienen situaciones en donde no se puede reflejar similitud alguna, y por ello se esperaría un valor de cero. En el ejemplo III, cualquier par de patrones agrupado junto que se tome de la solución C es imposible encontrarlo en la solución C'. Lo mismo sucede de forma inversa. Es por ello que ningún par encontrado junto en una solución se lo encuentra en la otra. En este ejemplo todos los índices responden como es esperado, excepto ARI. En este caso es debido a que el denominador del índice se calcula como un valor observado menos un valor esperado. El primero es menor que el segundo y ésto resulta en un valor negativo. En el ejemplo IV no hay forma de tomar pares de patrones agrupados juntos, en ninguna solución. Por ello se esperaría un valor de cero. FM y $\mathcal{OC}$ devuelven este valor, pero ARI y JAC no. Ésto es debido a que ambos producen un valor cero en el denominador y de esta forma no es posible computar el

Tabla 2.3: Resultados de índices aplicados a ejemplos artificiales gradualmente solapados

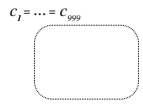
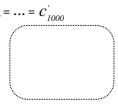
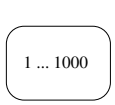
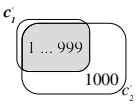
	Soluciones		Índices			
	C	C'	**FM**	**ARI**	**JAC**	$\mathcal{OC}$
I			0.632	0.000	0.400	0.400
II			0.665	-2.186	0.442	0.514
III			0.695	2.347	0.483	0.650
IV			0.721	1.444	0.520	0.800
V			0.700	1.324	0.489	0.840
VI			0.692	1.238	0.478	0.909
VII			0.692	1.179	0.478	1.000

resultado final. Con estos ejemplos se muestra que el índice propuesto puede manejar correctamente situaciones de soluciones sin solapamiento.

En la Tabla 2.3 se presentan varios ejemplos de soluciones con un solapamiento progresivo en la solución a evaluar C'. La tabla tiene una estructura análoga a la Tabla 2.2 y en todos los ejemplos también se agrupan 6 puntos de datos. La solución de referencia C es la misma para todos los ejemplos y corresponde a los 6 puntos agrupados en un único grupo c_1. La solución a evaluar siempre se comprende de los mismos 6 objetos agrupados en 2 grupos.

En el ejemplo I se comienza sin solapamiento y dividiendo los objetos de

Tabla 2.4: Resultados de índices para ejemplos extremos en cuanto al solapamiento o cantidad de objetos

	Soluciones		Índices			
	C	C'	**FM**	**ARI**	**JAC**	$\mathcal{OC}$
I			0.001	1.000	0.001	1.000
II			0.707	1.200	0.500	0.999

la solución C' en 2 grupos. De forma gradual en cada ejemplo se comienza a aumentar el solapamiento hasta llegar al último, el VII, en donde todos los objetos de la solución C' se encuentran agrupados en ambos grupos. De esta forma se llega a un solapamiento completo de ambos grupos. Por ejemplo, en el ejemplo II de C' el objeto 3 se encuentra en ambos grupos. Con esta pertenencia a ambos grupos no se genera ningún par repetido debido al solapamiento, pero sí se generan 3 nuevos pares debido a la nueva interacción del objeto 3 con los objetos del grupo c'_2 (los nuevos pares son: (3, 4), (3, 5), (3, 6)). Esta nueva información puede ser detectada por el índice propuesto con un incremento en su valor. Por el contrario, en los ejemplos III y IV, el incremento de solapamiento genera repetición de pares que se encuentran en ambos grupos de la solución a evaluar. Por ejemplo, el par formado por los objetos 2 y 3 puede encontrarse en ambos grupos de dicha solución. Así se puede visualizar que el incremento de solapamiento desde el ejemplo I al IV puede verse reflejado en un aumento de los índices bajo estudio, excepto ARI. El aumento de los índices se debe a que a más solapamiento, más pares de objetos encontrados en C pueden llegar a encontrarse en C', aumentando la similitud.

A partir del ejemplo IV en adelante, todos los objetos de la solución C' se pueden encontrar agrupados juntos. Además, a medida que se progresa en los ejemplos, se pueden encontrar nuevos pares de objetos agrupados juntos debido al solapamiento. Los índices clásicos, FM y JAC, aumentan de valor hasta el ejemplo V, en donde comienzan a decrementar. Sin embargo es esperable siguieran aumentando debido al incremento de solapamiento y de pares de patrones. ARI, por su parte, muestra valores en desacuerdo con lo esperado. Como se explicó en la Tabla 2.2 ARI puede llegar a arrojar valores incluso negativos en ciertos casos. Por otro lado, siguiendo la tendencia del solapamiento, el índice $\mathcal{OC}$ presenta una crecimiento monótono en su valor

a través de los ejemplos, hasta lograr un valor máximo de 1,000 en presencia de un solapamiento completo.

Finalmente, en la Tabla 2.4 se presentan 2 ejemplos extremos. La estructura de tabla es la misma que en los dos ejemplos anteriores.

En el primer ejemplo se plantea una situación similar al ejemplo VII de la Tabla 2.3 pero llevada al extremo, con más grupos. Dicha solución C posee 999 grupos y la C' posee 1000 grupos. En ambos casos todos los objetos están agrupados juntos en todos los grupos. Nuevamente $\mathcal{OC}$, y ahora además ARI, reflejan un valor de similitud de 1, mientras que FM y JAC reflejan valores que parecen tender a cero. Con esto se refleja que el índice es robusto frente a un alto grado de solapamiento, siendo esto lo esperado dado que independientemente al grado de solapamiento los pares de patrones pueden encontrarse juntos tanto en una como en otra solución. Los demás índices fallan frente al aumento de solapamiento.

Por último, en el ejemplo II se toma el ejemplo VI de la Tabla 2.3 y se lo lleva al extremo en el número de objetos agrupados. En C se agrupan 1000 objetos juntos en un solo grupo. En C' se agrupan los mismos 1000 objetos en dos grupos. En c'_2 están todos juntos y en c'_1 están todos excepto un objeto. Si bien todos los pares agrupados en la solución C pueden encontrarse en C', existen unos pocos pares que pueden encontrarse juntos solamente una vez en C' mientras que la mayoría puede encontrarse juntos 2 veces. Aquí $\mathcal{OC}$ representa un valor mucho más cercano a 1 que en el ejemplo VI de la Tabla 2.3. Ésto es lo que se espera dada la semejanza de ambos ejemplos, siendo que en la Tabla 2.3 se ha llevado a un extremo mucho menor la proporción de pares que se encuentran una vez en C' con respecto a los que se encuentran 2 veces. Los demás índices parecen estancarse en valores que no representan ni dan indicio alguno con respecto a la situación que se observa.

Con los ejemplos de las 3 tablas puede verse cómo el índice propuesto es capaz de representar correctamente tanto situaciones donde existen soluciones clásicas como soluciones con solapamiento. Más aún, puede observarse cómo el grado de solapamiento puede llegar a tener importantes variaciones tanto en la cantidad de grupos como en la cantidad de objetos a agrupar, y aún así el índice propuesto es capaz de trabajar naturalmente esta situación. Además, los índices clásicos tienen dificultades importantes al intentar medir similaridades en soluciones que puedan tener algún grado de solapamiento.

Resultados

En esta sección se continúa el análisis del índice propuesto junto con los índices clásicos, ahora con la aplicación a conjuntos de datos reales y a un conjunto de datos de una red social. Para facilitar la reproducibilidad de los resultados, se encuentra a disposición el código fuente[1] que genera los principales resultados de esta tesis. Más precisamente, la implementación del índice $\mathcal{OC}$ y los experimentos realizados sobre un conjunto de datos real (YouTube). Para detalles adicionales sobre la metodología propuesta, ver la sección de Anexos.

3.1. Evaluación sobre bases de datos reales

En esta subsección se presentan 4 conjuntos de datos tomados de UCI Machine Learning[2] Iris, Wine, Yeast y Glass. Son bases de datos bien conocidas por la comunidad científica del área y relativamente sencillas para trabajar en una primera experimentación con el índice.

El conjunto de datos Iris posee 150 patrones que representan la medición de 4 atributos de 3 clases distintas de la flor del Iris, de 50 patrones cada una. Los atributos que se miden de cada objeto son el ancho y el largo del sépalo y pétalo de la flor. Las tres especies de Iris medidas son *Iris Virgínica*, *Iris Versicolor* e *Iris Setosa*. Las primeras dos especies están relativamente cerca en el espacio de atributos mientras que la tercera, *I. Setosa*, está más alejada y por ello se la puede separar linealmente de las otras dos. Por su parte, el conjunto de datos Wine representa la medición y análisis de 13 atributos químicos realizados sobre vinos de una misma región de Italia, pero tomados de diferentes cultivos. Dicho conjunto de datos consta de 178 patrones distribuidos en 3 grupos: cultivo A con 59 casos, cultivo B con 71 y cultivo C con 48. El conjunto de datos de Yeast representa un estudio sobre la levadura en donde se busca determinar la localización de sus proteínas en las células. Posee 1484 casos distribuidos en 10 grupos; con 464, 429, 244, 163, 51, 44, 37, 30, 20 y 5 elementos en cada uno. A cada objetos se le realizaron 8 mediciones que representan sus atributos. Por último, el conjunto Glass posee 214 patrones distribuidos en 7 grupos o tipos de vidrios, en donde a cada objeto se le han medido 9 atributos. Se encuentran el índice de refracción del vidrio y el contenido de óxido de distintos elementos como el sodio, magnesio, aluminio, silicio, potasio, calcio, hierro y bario. Estas bases de datos se encuentran disponibles de forma

[1] https://github.com/dncampo/OC_index.

[2] http://archive.ics.uci.edu/ml/datasets/

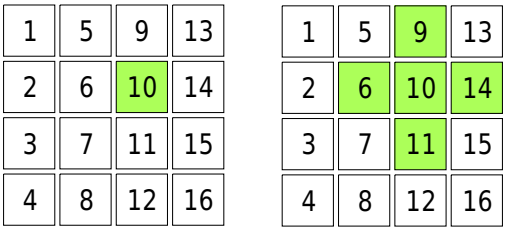


Figura 3.1: Mapa autoorganizativo de 4x4 neuronas. a) con vecindad de Von Neumann 0 ($V_n = 0$) (izq.), b) con vecindad de Von Neumann 1 ($V_n = 1$) (der.).

gratuita para uso público y, además, son ampliamente usadas en el ámbito académico para distintas pruebas. Es por ello que fueron seleccionadas para los experimentos propuestos a continuación.

Para el agrupamiento de los datos se utilizó el algoritmo de Mapas autoorganizativos (SOM, por su sigla en inglés)[29]. Se entrenaron mapas con distintas cantidades de neuronas (clusters). Para tener en cuenta clusters con solapamiento, se usó la vecindad de Von Neumann (V_n)[3]. De este modo, se toma cada neurona y sus vecinas inmediatas (norte, sur, este y oeste) como un mismo grupo. En la Figura 3.1 puede visualizarse un mapa autoorganizativo con 16 neuronas. En la Figura 3.1 a) la neurona 10 representa un grupo, ya que el mapa fue considerado con $V_n = 0$, o sin vecindad. En la Figura 3.1 b), el conjunto de neuronas 10 y sus vecinas (9, 11, 14 y 6) forman un único grupo, dado que el mapa fue considerado con $V_n = 1$. De esta manera, con $V_n = 1$ se obtienen clusters solapados, ya que varias neuronas forman parte de más de un grupo y, en consecuencia, también los patrones allí agrupados.

También se consideraron otros parámetros para el entrenamiento del mapa. Se utilizó topología rectangular en forma de grilla y la cantidad de iteraciones de entrenamiento fue fijada en 100. La inicialización de los mapas se hizo de forma determinística utilizando PCA [30] sobre los datos. Cada mapa fue entrenado con diferentes cantidades de neuronas para agrupar las bases de datos.

[3] http://mathworld.wolfram.com/vonNeumannNeighborhood.html

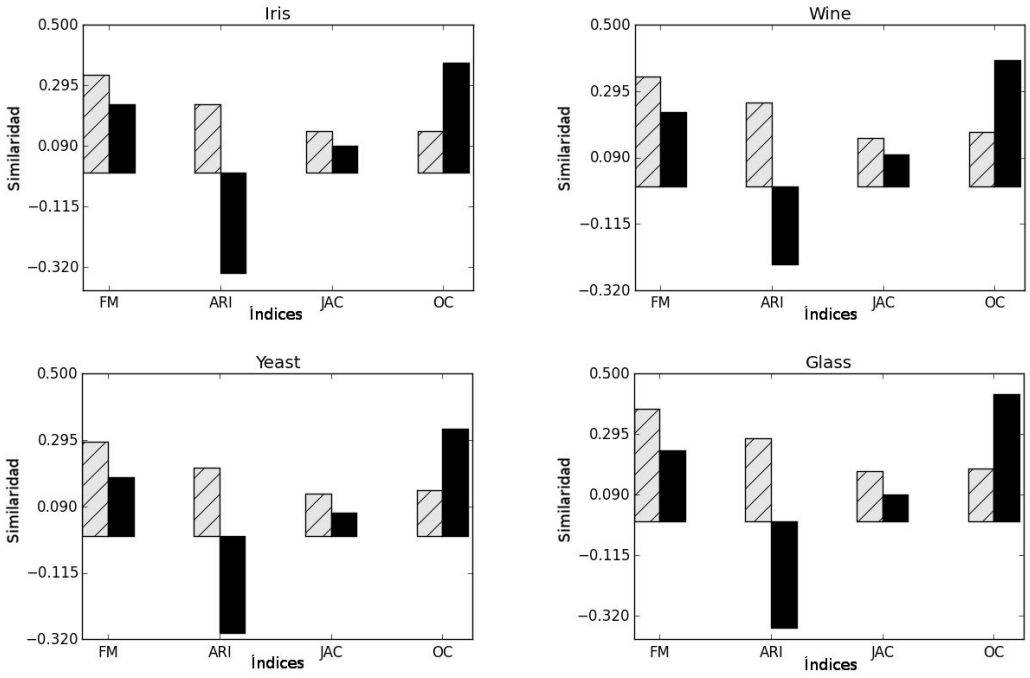


Figura 3.2: Gráfico de barras correspondiente a los índices FM, ARI, JAC y $\mathcal{OC}$ para los conjuntos de datos de Iris, Wine, Yeast y Glass. Las soluciones de referencia C poseen 25 grupos y sin solapamiento ($V_n = 0$). Soluciones C' poseen 100 grupos con $V_n = 0$ (barras grises) y $V_n = 1$ (barras negras).

Tabla 3.1: Resultados de los índices FM, ARI, JAC y $\mathcal{OC}$ aplicados a los conjunto de datos Iris, Wine, Yeast y Glass. Las soluciones de referencia C poseen 4 o 25 neuronas y solapamiento nulo ($V_n = 0$). Soluciones C' poseen 25 o 100 neuronas y considerando ($V_n = 1$) o no solapamiento ($V_n = 0$).

	grupos en C y C'	**FM**		**ARI**		**JAC**		$\mathcal{OC}$	
		$Vn = 0$	$Vn = 1$	$Vn = 0$	$Vn = 1$	$Vn = 0$	$Vn = 1$	$Vn = 0$	$Vn = 1$
Iris	$k = 4$ vs $k' = 25$	0,33	0,30	0,16	4,26	0,14	0,10	0,14	0,38
	$k = 4$ vs $k' = 100$	0,17	0,16	0,03	−0,66	0,03	0,03	0,03	0,11
	$k = 25$ vs $k' = 100$	0,33	0,23	0,23	−0,34	0,14	0,09	0,14	0,37
Wine	$k = 4$ vs $k' = 25$	0,40	0,32	0,23	9,33	0,17	0,12	0,17	0,48
	$k = 4$ vs $k' = 100$	0,19	0,18	0,06	−0,52	0,04	0,04	0,04	0,14
	$k = 25$ vs $k' = 100$	0,34	0,23	0,26	−0,24	0,15	0,10	0,17	0,39
Yeast	$k = 4$ vs $k' = 25$	0,32	0,23	0,15	6,61	0,13	0,08	0,13	0,35
	$k = 4$ vs $k' = 100$	0,16	0,14	0,04	−0,58	0,03	0,03	0,03	0,11
	$k = 25$ vs $k' = 100$	0,29	0,18	0,21	−0,30	0,13	0,07	0,14	0,33
Glass	$k = 4$ vs $k' = 25$	0,33	0,27	0,10	3,77	0,11	0,08	0,11	0,30
	$k = 4$ vs $k' = 100$	0,15	0,14	0,02	−0,75	0,02	0,02	0,02	0,09
	$k = 25$ vs $k' = 100$	0,38	0,24	0,28	−0,36	0,17	0,09	0,18	0,43

En la Tabla 3.1 se presentan los resultados obtenidas con 4 bases de datos reales. La tabla se divide de la siguiente forma: bajo la primera columna se visualiza el nombre de cada base de datos. La columna 2 indica la cantidad de grupos que se consideraron para la solución de referencia y para la que se desea comparar. Las columnas 3 a 6 presentan los valores obtenidos para cada uno de los índices estudiados: FM, ARI, JAC y $\mathcal{OC}$, respectivamente. Con respecto a éstas últimas 4 columnas, cada una posee una subdivisión en otras dos columnas con los valores de cada índice considerando la solución C' sin solapamiento y con solapamiento ($V_n = 0$ y $V_n = 1$). Sobre cada conjunto de datos se realizaron seis experimentos: $k = 4$ vs $k' = 25$, $k = 4$ vs $k' = 100$ y $k = 25$ vs $k' = 100$; considerando tanto $V_n = 0$ como $V_n = 1$ para la solución C'. Para la solución de referencia siempre se consideró $V_n = 0$.

Para el caso del conjunto de datos del Iris, se observa un descenso del valor del índice FM no sólo cuando $k = 4$ y k' pasa de 25 a 100, sino además cuando V_n pasa de 0 a 1. ARI produce un comportamiento más errático, aunque consistente, puesto que si bien disminuye su valor siempre que se pasa de $k' = 25$ a $k' = 100$, también se observa que al considerar solapamiento posee un valor mayor a 1,000 en el experimento de $k = 4$ vs $k' = 25$, pero con valores negativos en los otros dos casos probados. El índice JAC posee un comportamiento muy parecido al índice FM: disminuye cuando se considera solapamiento y cuando más grupos se consideran en C'. Finalmente, el índice propuesto, $\mathcal{OC}$, también muestra una disminución de sus valores cuando se toma en cuenta un mayor número de grupos en la solución C', pero aumenta cuando se consideran grupos solapados ($V_n = 1$). Este es un comportamiento esperado dado que considerar solapamiento es consistente con la idea de que la probabilidad de encontrar más pares de patrones agrupados juntos en ambas soluciones aumente.

El análisis para los demás conjuntos de datos es similar. Tanto para Wine, Yeast y Glass se puede arribar a las mismas conclusiones que para Iris. Esto es, los valores para los índices FM y JAC disminuyen cuando el número de grupos aumenta. Así también es el caso cuando se considera solapamiento. ARI también presenta un comportamiento similar al que presentaba para las pruebas con Iris. Con respecto a $\mathcal{OC}$, se puede observar un aumento notable de su valor en los casos en los que se considera $V_n = 1$ en C', como así también una disminución del mismo cuando más grupos son considerados en C'.

En la Figura 3.2 se observan las gráficas de los resultados obtenidos para los cuatro índices analizados en la Tabla 3.1, pero para el experimento específico de $k = 4$ vs $k' = 100$, con $V_n = 0$ y $V_n = 1$. Para todos los conjuntos de datos, cuando el solapamiento aumenta los índices clásicos muestran un decaimiento notable en su valor mientras que el índice propuesto muestra

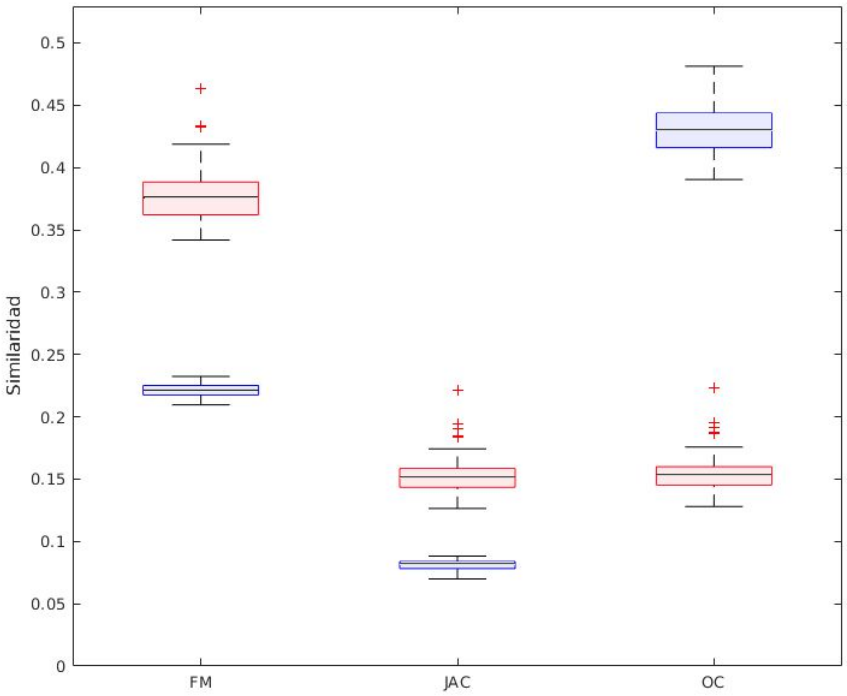


Figura 3.3: Diagrama de cajas de índices FM, JAC y $\mathcal{OC}$ para conjunto de datos Iris con los algoritmos k-medias y FCM. Para las pruebas realizadas se utilizó el algoritmo de k-medias con $k = 4$ grupos para la solución de referencia y el algoritmo FCM con $k = 25$ grupos para la solución solapada. Las cajas rojas corresponden a mediciones de soluciones sin solapamiento y las azules a mediciones de soluciones con solapamiento.

una tendencia opuesta.

En los experimentos realizados se puede presenciar un decremento en todos los índices cuando el número de grupos de C es más lejano al número de grupos de C'. Es decir, al haber más grupos los datos se distribuyen en más neuronas o grupos y, de esta manera, se reduce el valor de los índices que valoran la similitud. Ésto puede observarse cuando se analiza el experimento $k = 25$ vs $k' = 100$. En el caso de los índices clásico FM y JAC, sus valores disminuyen también como consecuencia de la consideración de la vecindad en los mapas C', mientras que el índice propuesto presenta siempre un aumento cuando aparece la consideración $V_n = 1$ con respecto a $V_n = 0$. El índice ARI acusa valores negativos cuando se considera vecindad. Esto se debe a que los índices clásicos no miden correctamente soluciones que consideran solapamiento, mientras que $\mathcal{OC}$ sí. Con dicha consideración en mente, debe notarse que con presencia de solapamiento los índices FM y

JAC no cuentan las coincidencias de los grupos apropiadamente. Con esto se explica por qué dichos índices disminuyen y $\mathcal{OC}$ aumenta en presencia de solapamiento.

Para mostrar la independencia del índice $\mathcal{OC}$ en relación al algoritmo de clustering, se realizó un experimento equivalente al de la Tabla 3.1 en el cual el equivalente al SOM con $V_n = 0$ es el algoritmo k-means (KM) estándar; y en lugar del SOM con clusters solapados ($V_n = 1$) se usó el algoritmo fuzzy c-means (FCM) [31]. En la Figura 3.3 se pueden observar los resultados de este experimento con el conjunto de datos de Iris. Para cada par de soluciones comparadas se realizaron 100 corridas. De esta manera se graficó el diagrama de cajas para cada grupo de corridas. Las cajas rojas corresponden a las comparaciones entre soluciones sin solapamiento. Las cajas azules corresponden a pruebas pero considerando soluciones solapadas. Como se puede ver, tanto el índice FM como el JAC, bajan al medir soluciones con solapamiento. En cambio, el índice $\mathcal{OC}$ aumenta su valor al ser aplicado a las soluciones de FCM con $V_n = 1$, es decir refleja correctamente que los clusters de esta solución son más parecidos a los de referencia, dado que tienen solapamiento.

Resumiendo, se pudo observar tanto con experimentos sobre conjuntos artificiales, en la sección anterior, o bases de datos reales, cómo el índice propuesto fue capaz de medir correctamente la similitud de soluciones de agrupamiento, sea en presencia o en ausencia de solapamiento de los grupos. Más aún, $\mathcal{OC}$ se mostró robusto y confiable con casos extremos de solapamiento, permitiendo un mejor entendimiento y comparación de los resultados de los algoritmos de agrupamiento.

3.2. Aplicación en una comunidad de red social

En esta sección se describe el experimento realizado sobre una base de datos real correspondiente a la representación de comunidades en una red social (YouTube), la cual es analizada y descrita en [32]. YouTube es una red social para compartir contenido en formato de videos con otros usuarios. Los usuarios del sistema pueden crear grupos, también llamados comunidades o simplemente canales, para compartir su contenido y otros usuarios pueden unirse o suscribirse a los mismos.

El conjunto de datos analizado muestra la relación de varios usuarios de la red social formando parte de distintas comunidades o canales. El conjunto de datos está compuesto de comunidades, las cuales son consideradas como tal cuando poseen 2 o más usuarios con intereses similares. Cada comunidad es considerada como un grupo de usuarios (como el grupo que se obtiene al aplicar un algoritmo de agrupamiento) y en el conjunto de datos se caracteriza como una lista de identificadores de los usuarios que lo forman.

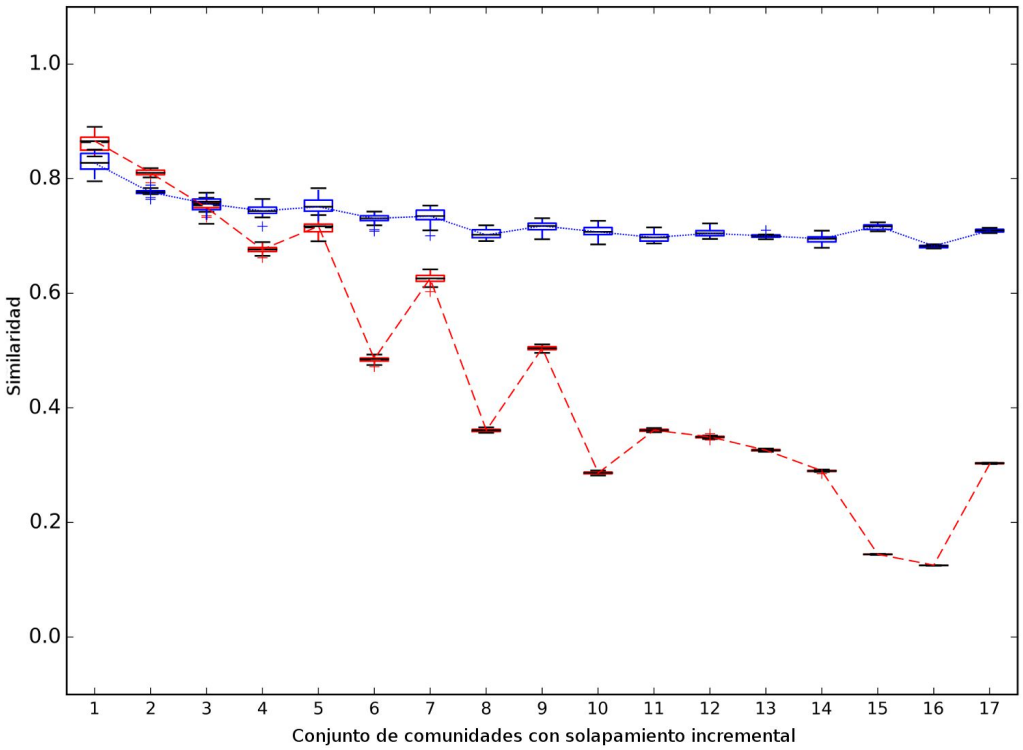


Figura 3.4: Diagrama de caja de los índices FM y $\mathcal{OC}$ para el conjunto de datos de YouTube. La línea roja (punteada) corresponde al índice FM mientras que la azul (continua) al índice $\mathcal{OC}$.

Un usuario puede pertenecer a una o más comunidades. Cuando un usuario pertenece a más de una comunidad, aquellas en las que dicho usuario pertenece quedan solapadas entre sí. El nivel de solapamiento de una comunidad dependerá de cuántos de sus usuarios pertenezcan a otras comunidades. Con esto, el conjunto de datos resultante, después de eliminar las comunidades que posean menos de 10 usuarios, contiene 37038 usuarios y 2087 comunidades. Las comunidades con menos de 10 usuarios se eliminaron dado que poseían, en general, poco solapamiento, lo cual hubiera desviado el foco de atención sobre el testeo de los índices cuando se considera solapamiento.

El experimento involucró el ordenamiento de los grupos C_i del conjunto de datos en función del solapamiento. Los grupos C'_j de la solución a evaluar fueron generados a partir del conjunto de datos original mediante la aplicación de perturbaciones aleatorias. Dichas perturbaciones corresponden a seleccionar usuarios al azar y agregarlos a distintas comunidades, también seleccionadas al azar. El conjunto de datos original se dividió en sub conjuntos. Dado que se realizó un ordenamiento creciente de las comunidades según su grado de solapamiento, el primer subconjunto consiste en las únicas 35 comunidades con un grado de solapamiento nulo. Es decir,

sus usuarios solamente pertenecen a una única comunidad. Cada uno de los siguientes subconjuntos considerados representa un nivel diferente de solapamiento, siempre incremental. De esta forma, las comunidades que forman un subconjunto poseen un grado de solapamiento similar entre sí. Cada subconjunto posee el triple del número de comunidades que el subconjunto inicial de solapamiento nulo. Esto garantiza que todos los subconjuntos tengan un número mínimo de elementos para calcular los índices. El último subconjunto incluye las comunidades con mayor grado de solapamiento. De esta manera el conjunto de datos original quedó dividido en 17 subconjuntos disjuntos ordenados desde un solapamiento nulo al máximo posible.

En la Figura 3.4 puede observarse el diagrama de cajas para los índices FM y $\mathcal{OC}$, para una perturbación del 10 % de los usuarios en la solución generada C'. Cada caja representa la mediana de 20 corridas del experimento sobre cada subconjunto. En el eje de las abscisas se representa el nivel de solapamiento creciente de cada subconjunto, desde solapamiento nulo (subconjunto etiquetado como 1) a solapamiento máximo (subconjunto etiquetado 17). Para los subconjuntos sin solapamiento ambos índices reportan valores entre 0,8 y 0,9. A medida que el grado de solapamiento crece se nota un decaimiento notable del índice FM hasta llegar a un valor apenas por encima de 0,1, cuando el solapamiento es muy alto. En cambio, los valores del índice $\mathcal{OC}$ se mantienen bastante estables aún a medida que crece el solapamiento. De esta forma el índice muestra una fuerte capacidad para mantenerse robusto. Más aún, en situación de solapamiento alto el índice FM falla de una forma fluctuante, mientras que el índice propuesto mantiene una curva estable y suave.

En consecuencia, se puede concluir que el índice propuesto es efectivo para la medición de similaridades en escenarios de soluciones de agrupamiento donde se presente solapamiento. Más aún, el índice $\mathcal{OC}$ presenta un comportamiento más estable que los índices clásicos, como FM, en situaciones tanto solapadas como no solapadas.

Conclusiones y trabajo futuro

4.1. Conclusiones

En esta tesis se estudiaron índices clásicos de validación de agrupamientos llegandose a la conclusión de que presentaban falencias con cierto tipo interesante de soluciones. Además se desarrolló un nuevo índice para medición de calidad en clustering, denominado $\mathcal{OC}$, para la evaluación externa de algoritmos de agrupamiento. Este índice puede ser usado tanto en el caso que exista solapamiento de los grupos como en la situación clásica de clustering, con particiones sin solapamiento. El índice propuesto fue diseñado desde un enfoque probabilístico y se lo comparó con otros índices clásicos estudiados como Fowlkes-Mallows, Adjusted Rand Index y Jaccard. Los índices clásicos mostraron comportamientos inesperados cuando midieron soluciones solapadas, tanto en conjuntos de datos artificiales como en conjuntos de datos reales. Se hicieron pruebas sobre una base de datos real, perteneciente a la red social de YouTube, donde se midieron comunidades de usuarios que compartían contenido en línea. Para éste experimento, además, se perturbaron los datos agregando usuarios en nuevas comunidades, incrementando el solapamiento. En este caso, el índice $\mathcal{OC}$ siempre mantuvo valores coherentes aún en aumentos más extremos en el grado de solapamiento. De esta forma, el índice propuesto se mostró inmune al solapamiento en todos los casos en los que se lo probó, midiendo de manera precisa las similitudes entre soluciones de clustering en tal situación. Más aún, $\mathcal{OC}$ se comportó de manera correcta también en situaciones clásicas donde no hay solapamiento. Así se concluye que el índice propuesto puede utilizarse en cualquier tipo de soluciones, haya o no solapamientos.

Con respecto a trabajos futuros, se realizarán experimentos usando el índice propuesto para analizar la estabilidad de soluciones de agrupamiento con distintos grados de solapamiento. Sería también interesante comparar el desempeñoo con otros índices clásicos basados en teoría de la información como, por ejemplo, Normalized Mutual Information. Asimismo, sería importante ampliar el sustento experimental y realizar un análisis de escalabilidad para su aplicación a grandes datos, con el correspondiente desarrollo de implementaciones computacionalmente eficientes.

REFERENCIAS

[1] D. Skillicorn, *Understanding Complex Datasets: Data Mining with Matrix Decompositions*. CRC Press, May 2007.

[2] R. Xu and D. C. Wunsch, *Clustering*. Wiley and IEEE Press, 2009.

[3] M. de Souto, I. Costa, D. de Araujo, T. Ludermir, and A. Schliep, "Clustering cancer gene expression data: a comparative study," *BMC Bioinformatics*, vol. 9, pp. 497–507, 2008.

[4] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, pp. 651–666, 2010.

[5] G. Stegmayer, D. Milone, L. Kamenetzky, M. Lopez, and F. Carrari, "Neural network model for integration and visualization of introgressed genome and metabolite data," in *IEEE International Joint Conference on Neural Networks*, pp. 3177–3183, IEEE Computational Intelligence Society, 2009.

[6] D. Milone, G. Stegmayer, L. Kamenetzky, M. Lopez, J. Giovannoni, J. M. Lee, and F. Carrari, "*omeSOM: a software for integration, clustering and visualization of transcriptional and metabolite data mined from interspecific crosses of crop plants," *BMC Bioinformatics*, vol. 11, pp. 438–448, 2010.

[7] X. V. Nguyen, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *Journal of Machine Learning Research*, vol. 11, pp. 2837–2854, 2010.

[8] X. V. Nguyen, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: is a correction for chance necessary?," in *ICML*, p. 135, 2009.

[9] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of Intelligent Informations Systems*, vol. 17, no. 1, pp. 107–145, 2001.

[10] J. Handl, J. Knowles, and D. B. Kell, "Computational cluster validation in post-genomic data analysis," *Bioinformatics*, vol. 21, no. 15, pp. 3201–3212, 2005.

[11] M. Meila and D. Heckerman, "An experimental comparison of model-based clustering methods," *Machine Learning*, vol. 42, pp. 9–29, January 2001.

[12] M. Meilă, "Comparing clusterings—an information based distance," *Journal of Multivariate Analysis*, vol. 98, pp. 873–895, May 2007.

[13] M. K. Goldberg, M. Hayvanovych, and M. Magdon-Ismail, "Measuring similarity between sets of overlapping clusters," in *2010 IEEE Second International Conference on Social Computing*, pp. 303–308, Aug 2010.

[14] Z. Wang, D. Zhang, X. Zhou, D. Yang, Z. Yu, and Z. Yu, "Discovering and Profiling Overlapping Communities in Location-Based Social Networks," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, pp. 499–509, Apr. 2014.

[15] H. Alvari, S. Hashemi, and A. Hamzeh, "Discovering overlapping communities in social networks: A novel game-theoretic approach," *AI Communications*, vol. 26, pp. 161–177, Apr. 2013.

[16] P. K. Gopalan and D. M. Blei, "Efficient discovery of overlapping communities in massive networks," *Proceedings of the National Academy of Sciences*, vol. 110, pp. 14534–14539, Sept. 2013.

[17] T. Gossen, M. Kotzyba, and A. Nürnberger, "Graph clusterings with overlaps: Adapted quality indices and a generation model," *Neurocomputing*, vol. 123, pp. 13–22, Jan. 2014.

[18] T. Chakraborty, "Leveraging disjoint communities for detecting overlapping community structure," *J. Stat. Mech.*, vol. 2015, p. P05017, May 2015.

[19] J. Xie, S. Kelley, and B. K. Szymanski, "Overlapping Community Detection in Networks: The State-of-the-art and Comparative Study," *ACM Comput. Surv.*, vol. 45, pp. 43:1–43:35, Aug. 2013.

[20] A. Amelio and C. Pizzuti, "Overlapping Community Discovery Methods: A Survey," in *Social Networks: Analysis and Case Studies* (Ş. Gündüz-Öğüdücü and A. Ş. Etaner-Uyar, eds.), Lecture Notes in Social Networks, pp. 105–125, Springer Vienna, 2014.

[21] M. Brun, C. Sima, J. Hua, J. Lowey, B. Carroll, E. Suh, and E. R. Dougherty, "Model-based evaluation of clustering validation measures," *Pattern Recognition*, vol. 40, pp. 807–824, Mar. 2007.

[22] J. Wu, J. Chen, H. Xiong, and M. Xie, "External validation measures for K-means clustering: A data distribution perspective," *Expert Systems with Applications*, vol. 36, pp. 6050–6061, Apr. 2009.

[23] D. Campo, G. Stegmayer, and D. Milone, "Stability analysis in overlapped clusters," *Iberoamerican Journal of Artificial Intelligence*, vol. 17, no. 53, pp. 79–89, 2014.

[24] D. Campo, G. Stegmayer, and D. Milone, "A new index for clustering validation with overlapped clusters," *Expert Systems with Applications*, vol. 64, pp. 549 – 556, 2016.

[25] E. B. Fowlkes and C. L. Mallows, "A Method for Comparing Two Hierarchical Clusterings," *Journal of the American Statistical Association*, vol. 78, pp. 553–569, Sept. 1983.

[26] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, pp. 193–218, Dec 1985.

[27] P. Jaccard, "Nouvelles recherches sur la distribution florale," 1908.

[28] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.

[29] T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, pp. 1–6, Nov. 1998.

[30] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[31] J. C. Bezdek, R. Ehrlich, and W. Full, "Fcm: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2, pp. 191 – 203, 1984.

[32] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," *CoRR*, vol. abs/1205.6233, 2012.

Anexos

El apartado de anexos se organiza de la siguiente manera.

- En el Anexo A se presenta el artículo publicado por Campo, D. N., Stegmayer, G., y Milone, D. H., "Nuevo índice para el análisis de estabilidad en clústers solapados", *14vo Simposio Argentino de Inteligencia Artificial, ASAI 2013, 42 JAIIO*, pp. 24–35, 2014.
- En el Anexo B se presenta el artículo publicado por Campo, D. N., Stegmayer, G., y Milone, D. H., "Análisis de estabilidad en clústers solapados", *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, vol. 17, núm. 53, enero-junio, 2014, pp. 79-89 Asociación Española para la Inteligencia Artificial.
- En el Anexo C se presenta el artículo publicado por Campo, D. N., Stegmayer, G., y Milone, D. H., "A New Index for Cluster Validation with Overlapped Clusters", *ELSEVIER Expert Systems With Applications*, doi: 10.1016/j.eswa.2016.08.021, enviado en Febrero de 2016, revisado en Julio de 2017, en prensa en Agosto de 2017.

El tesista declara haber implementado los algoritmos y llevado a cabo los experimentos descriptos para obtener los resultados que allí se presentan. Estas tareas fueron realizadas bajo la guía y supervisión del director Dr. Georgina Stegmayer y codirector de tesis Dr. D. H. Milone. 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 codirector de tesis. Los abajo firmantes avalan esta declaración.

Aval de los Directores:

.....	.....
Dr. Georgina Stegmayer	Dr. Diego H. Milone
Director	Co-Director

Apéndice A

Nuevo índice para el análisis de estabilidad en clústers solapados

Campo, D. N., Stegmayer, G., y Milone, D. H., "Nuevo índice para el análisis de estabilidad en clústers solapados", *14vo Simposio Argentino de Inteligencia Artificial, ASAI 2013, 42 JAIIO*, pp. 24–35, 2014.

En este trabajo se arribó a un enfoque alternativo del índice FM y se mostró cómo el mismo no podía medir adecuadamente clústers solapados. Además se propuso un nuevo índice que permitía medir dichos tipos de clústers.

Nuevo índice para el análisis de estabilidad en clusters solapados

†D. N. Campo[1,2], ⋆G. Stegmayer[1,2] y ‡D. H. Milone[2]

[1] CIDISI-UTN-FRSF, CONICET, Lavaisse 610 - Santa Fe (Argentina)
[2] SINC(I)-FICH-UNL, CONICET, Ciudad Universitaria - Santa Fe (Argentina)
†dncampo@santafe-conicet.gov.ar, ⋆gstegmayer@santafe-conicet.gov.ar, ‡d.milone@ieee.org.

Resumen.

Analizar la estabilidad de una solución de clustering implica medir la capacidad de un algoritmo para producir resultados similares dada una misma fuente de datos de entrada. Los índices de validación externa permiten cuantificar dicha similitud entre un par de soluciones de clustering. Dentro de los índices clásicos más utilizados es posible validar soluciones con clusters no solapados, en donde cada patrón sólo puede pertenecer a un cluster. Sin embargo, en aplicaciones prácticas, generalmente se dan situaciones en las que un patrón podría poseer más de una etiqueta. En este trabajo se analiza un índice de validación externa desde un enfoque probabilístico y se provee una reformulación aplicable a soluciones con clusters solapados. Luego de presentar el nuevo índice, se muestran y discuten resultados de experimentos realizados sobre ejemplos artificiales y una base de datos real. Los resultados muestran cómo el nuevo índice puede medir adecuadamente la similitud entre clusters solapados, permitiendo así analizar la estabilidad en ambos casos.

Palabras clave: análisis de estabilidad, clusters solapados, medida de validación.

1. Introducción

Los algoritmos de clustering reciben un conjunto de datos como entrada y mediante un proceso no-supervisado lo particionan en cierto número de clusters o grupos. Se puede definir un cluster como un grupo de objetos homogéneos que poseen alguna medida de similitud entre ellos y que se muestran diferentes a los objetos agrupados en otros clusters [1, 2]. Dado que la aplicación de estos algoritmos sobre una base de datos siempre devuelve algún resultado, incluso cuando no exista estructura en la misma, ha surgido el análisis de estabilidad de soluciones de clustering. Para realizar este análisis es importante medir la capacidad de un algoritmo de agrupamiento para producir grupos similares de forma repetida [3–5]. Aunque no hay un total acuerdo en cuanto a su definición, existen autores que relacionan el concepto de estabilidad con soluciones de agrupamiento que no se ven alteradas bajo alguna perturbación de los datos de entrada [5, 6]. Cuando se hace análisis de estabilidad en clustering, también se

suele hablar de que se mantengan las estructuras naturales "subyacentes en los datos" y no que aparezcan nuevas estructuras como producto artificial de un algoritmo concreto.

Luego de realizar el agrupamiento en distintas condiciones, se desea medir o cuantificar la similitud entre las soluciones para poder compararlas [1]. Ultimamente, con el crecimiento en el estudio de análisis de agrupamientos se han estado utilizando nuevos algoritmos y medidas de comparación de clusters [7]. Para medir la similitud entre soluciones de clustering se pueden utilizar dos tipos de medidas de validación: internas y externas. Las primeras miden atributos de homogeneidad y separación de los datos en los clusters. Las medidas externas hacen una comparación entre distintas soluciones de clustering, tomando una como referencia y comparándola con otras [8, 9]. Con respecto a las medidas externas, se dispone principalmente de tres tipos para realizar la comparación entre soluciones: las basadas en el conteo de pares de patrones, en las que las soluciones coinciden o no; las basadas en el análisis de correspondencia de conjuntos y las basadas en la estadística y teoría de la información. Con respecto a las primeras, la más utilizada y difundida es la de Fowlkes-Mallows (FM) [10] que trabaja con las frecuencias de pares de patrones que se detecten agrupados juntos en ambas soluciones. En [11] los autores evalúan las distintas soluciones obtenidas cuantificando la similaridad mediante este índice. De las métricas basadas en conjuntos, una muy utilizada es la de máxima coincidencia [12] que se basa en comparar los clusters de ambas soluciones analizadas, haciendo coincidir aquellos que tengan mayor cantidad de elementos en común. Por último, de las medidas basadas en estadística, la información mutua normalizada es una de las más representativas y se basa en cuantificar la cantidad de información compartida entre ambas soluciones, a través del concepto de entropía [7,13]. Sin embargo, ninguna de estas medidas es capaz de representar correctamente las similitudes al trabajar con soluciones que posean clusters solapados.

Ultimamente se ha suscitado interés por el análisis de soluciones con clusters solapados. En [14] se compara y estudia la evolución de grupos de personas en redes sociales y se propone un algoritmo para computar nuevas distancias entre colecciones de grupos potencialmente solapados. En [15] se propone una nueva medida de validación para clusters solapados en el contexto de recuperación de documentos, en particular tratando el tema de solapamiento entre soluciones con diferente número de clusters. En [16] se propone un conjunto de restricciones sobre métricas para validación externa y se extiende el análisis para tratar soluciones solapadas. En este trabajo, en cambio, se hace un aporte al área ya que presentamos un análisis detallado del índice FM. Con un enfoque probabilístico se analiza cada factor del mismo considerando el solapamiento de clusters y se muestra cómo falla este índice cuando se lo intenta aplicar a soluciones que posean clusters solapados. Luego, a partir del análisis anterior, se deriva una nueva propuesta de índice teniendo especial atención en la situación mencionada. Aplicando ambos índices a diferentes casos de estudio y una base de datos reales, se verifican experimentalmente las mejoras expresadas cuando existen clusters solapados.

La organización de este trabajo es la siguiente. En la Sección 2 se realiza un análisis detallado de los factores del índice FM. En la Sección 3 se deriva el nuevo índice propuesto para poder tratar con las soluciones solapadas. Luego, en la Sección 4, se presentan resultados de aplicar ambos índices sobre datos artificiales y reales, y se discuten dichos experimentos. Por último se presentan

las conclusiones del trabajo en la Sección 5.

2. Análisis conceptual del índice Fowlkes-Mallows

El índice de Fowlkes-Mallows es una medida de similaridad entre soluciones de clustering, generalmente utilizado para validación externa. Como se mencionó en la Sección 1, en el contexto de análisis de estabilidad se utiliza para poder comparar soluciones y verificar si las mismas son o no estables. FM recibe el etiquetado de ambas soluciones, sobre el mismo conjunto de datos, y devuelve un valor $B_k \in [0, 1]$ que cuantifica la similitud entre las mismas. El valor 1 representa que ambas son exactamente iguales, mientras que 0 denota completa desigualdad. FM está definido en [10] como

$$B_k = \frac{T_k}{\sqrt{P_k Q_k}}, \tag{1}$$

donde

$$T_k = \sum_{i=1}^{k}\sum_{j=1}^{k} m_{ij}^2 - N, \tag{2}$$

$$P_k = \sum_{i=1}^{k} m_{i*}^2 - N, \tag{3}$$

$$Q_k = \sum_{j=1}^{k} m_{*j}^2 - N, \tag{4}$$

$$m_{i*} = \sum_{j=1}^{k} m_{ij}, \tag{5}$$

$$m_{*j} = \sum_{i=1}^{k} m_{ij}, \tag{6}$$

siendo $M = \{m_{ij}\}$ la matriz de contingencia obtenida entre 2 soluciones de k clusters cada una con N datos en el conjunto a particionar. Esta matriz posee tantas filas como cantidad de clusters haya en la primer solución, que llamaremos C, y tantas columnas como clusters tenga la segunda, C'. En cada posición ij de la matriz se coloca la cantidad de patrones en comun que se pueden contar entre los elementos del cluster i de la solución C y los patrones del cluster j de la solución C'. Es decir, $m_{ij} = |c_i \cap c'_j|$.

A continuación se hará un análisis conceptual de cada factor del índice FM. Consideremos T_k, y reemplazando $\sum_{i=1}^{k}\sum_{j=1}^{k} m_{ij} = N$ en (2) se puede escribir

$$\begin{aligned} T_k &= \sum_{i=1}^{k}\sum_{j=1}^{k} m_{ij}^2 - N = \sum_{i=1}^{k}\sum_{j=1}^{k}(m_{ij}m_{ij}) - \sum_{i=1}^{k}\sum_{j=1}^{k} m_{ij} \\ &= \sum_{i=1}^{k}\sum_{j=1}^{k} m_{ij}(m_{ij}-1) = 2\sum_{i=1}^{k}\sum_{j=1}^{k}\binom{m_{ij}}{2}, \end{aligned} \tag{7}$$

que representa el doble de la sumatoria de cada uno de los elementos de la matriz de contingencia tomados de a 2. Es decir, los elementos en común entre cada par de clusters de ambas soluciones tomados de a 2. Esto representa la cantidad de formas de elegir dos elementos en dicha intersección.

Análogamente, a partir de (3) se puede obtener

$$\begin{aligned} P_k &= \sum_{i=1}^{k} m_{i*}^2 - N = \sum_{i=1}^{k}(m_{i*}m_{i*}) - \sum_{i=1}^{k}\sum_{j=1}^{k} m_{ij} \\ &= \sum_{i=1}^{k}(m_{i*}m_{i*}) - \sum_{i=1}^{k} m_{i*} = \sum_{i=1}^{k}(m_{i*}(m_{i*})-1) = 2\sum_{i=1}^{k}\binom{m_{i*}}{2}, \quad (8) \end{aligned}$$

siendo, de forma similar a como ocurre con T_k, la cantidad de formas que hay de tomar 2 elementos de cada cluster de la solución analizada. Idéntico razonamiento se aplica para la interpretación de Q_k.

Consideremos C y C', dos soluciones de clustering con k clusters. Sean x e y dos patrones cualesquiera. Podemos interpretar al primer factor, P_k, como la probabilidad de que los dos patrones se encuentren en un mismo cluster c_i de la solución C, considerando un muestreo uniforme de los datos. Dado que los patrones tienen la misma probabilidad de ser seleccionados, al tomar un segundo patrón bajo la misma hipótesis, la probabilidad de que los dos se agrupen en el cluster c_i es:

$$Pr(x \in c_i \wedge y \in c_i) = \frac{\binom{|c_i|}{2}}{\binom{N}{2}} = \frac{|c_i|\,(|c_i|-1)}{N(N-1)}, \quad (9)$$

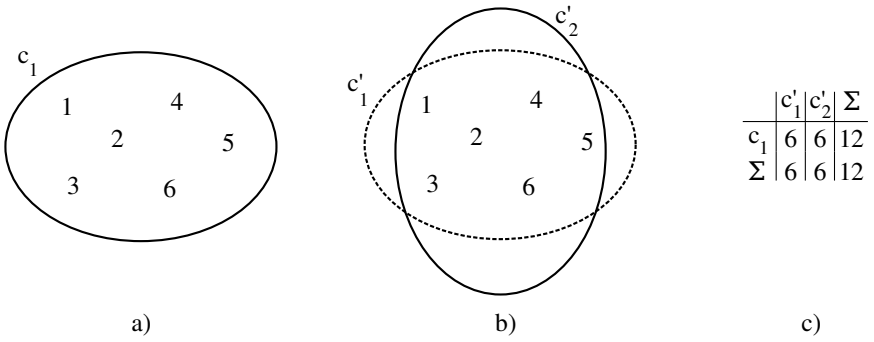
en donde el numerador representa la cantidad de formas de tomar de a 2 los patrones agrupados bajo el cluster c_i y el denominador considera todas las formas posibles de tomar 2 elementos del conjunto completo de datos. Esto último surge de considerar un caso extremo en el cual todos los patrones queden en un único cluster. Si se consideran los k clusters de C, ahora tenemos

$$\tilde{p_k} = \sum_{i=1}^{k} \frac{\binom{|c_i|}{2}}{\binom{N}{2}} = \frac{1}{N(N-1)}\sum_{i=1}^{k}|c_i|\,(|c_i|-1) = \frac{1}{N(N-1)}\sum_{i=1}^{k}|c_i|^2 - N, \quad (10)$$

que guarda relación directa con (3), dado que $|c_i| = m_{i*}$. La interpretación de Q_k es la misma que la de P_k, salvo que se analizan los clusters de la solución C'.

Con respecto a T_k, se puede considerar que representa la probabilidad de muestrear un par de patrones al azar y que éstos pertenezcan a un mismo cluster en C y C'. El razonamiento es similar al seguido para P_k, sólo que en vez de considerarse todos los patrones por cluster, se van considerando los pares de patrones comunes al emparejamiento entre cada par de clusters de ambas soluciones

$$Pr((x,y) \in c_i \wedge (x,y) \in c'_j) = \frac{\binom{|c_i \cap c'_j|}{2}}{\binom{N}{2}} = \frac{|c_i \cap c'_j|\,(|c_i \cap c'_j|-1)}{N(N-1)}. \quad (11)$$



	$\|c'_1\|$	$\|c'_2\|$	Σ
c_1	6	6	12
Σ	6	6	12

c)

Figura 1: Ejemplo ilustrativo con dos soluciones, a) C con $k = 1$, b) C' con solapamiento y $k = 2$ y c) matriz de contingencia correspondiente a las soluciones C y C'.

De la misma forma que en el desarrollo de P_k, el denominador representa la cantidad de formas de tomar 2 elementos de todo el conjunto de datos. Este sería el caso extremo en el que en ambas soluciones, todos los patrones hayan quedado en un único cluster. Si tomamos entonces todas las posibles comparaciones entre cada par de clusters de ambas soluciones, obtenemos la probabilidad

$$\tilde{t_k} = \sum_{i=1}^{k}\sum_{j=1}^{k}\frac{\binom{|c_i \cap c'_j|}{2}}{\binom{N}{2}} = \frac{1}{N(N-1)}\sum_{i=1}^{k}\sum_{j=1}^{k}|c_i \cap c'_j|\left(|c_i \cap c'_j| - 1\right) =$$

$$= \frac{1}{N(N-1)}\sum_{i=1}^{k}\sum_{j=1}^{k}|c_i \cap c'_j|^2 - N, \qquad (12)$$

que guarda relación directa con (2) dado que, como se observo antes, $m_{ij} = |c_i \cap c'_j|$.

Luego, el índice FM puede entenderse como la razón entre la probabilidad de obtener dos patrones juntos en las soluciones conjuntas sobre la media geométrica de dicha probabilidad considerando las soluciones por separado. Esto es:

$$B_k = \frac{\cancel{\frac{1}{\binom{N}{2}}}\sum_{i=1}^{k}\sum_{j=1}^{k}\binom{|c_i \cap c'_j|}{2}}{\sqrt{\cancel{\frac{1}{\binom{N}{2}}}\sum_{i=1}^{k}\binom{|c_i|}{2}\cancel{\frac{1}{\binom{N}{2}}}\sum_{j=1}^{k}\binom{|c'_j|}{2}}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k}(m_{ij}^2 - N)}{\sqrt{\sum_{i=1}^{k}(m_{i*}^2 - N)\sum_{j=1}^{k}(m_{*j}^2 - N)}} = \frac{T_k}{\sqrt{P_k Q_k}}. \qquad (13)$$

De esta manera se arriba a la formulación original de FM a traves del enfoque probabilístico alternativo.

Ahora bien, cuando este índice se aplica a clusters solapados, los resultados que se obtienen no son los intuitivamente esperados. Por ejemplo, en el gráfico de la Figura 1 se observan dos soluciones de clustering a) C con $k = 1$ y b) C' con $k = 2$. En c) se encuentra la matriz de contingencia para ambas soluciones. En la solución a), se puede observar cómo todos los patrones se agruparon juntos

en un único cluster. En la solución b) se puede observar cómo ambos clusters comparten todos los patrones. En este caso hay un solapamiento completo de ambos grupos.

Para calcular el índice FM sobre este ejemplo se procederá a calcular cada uno de sus factores. El valor del factor P_k se corresponde con la solución C y el de Q_k con la C'. Así $P_k = \sum_{i=1}^{1}(m_{i*}^2 - N) = 12^2 - 6 = 138$. En cambio $Q_k = \sum_{i=1}^{2}(m_{*j}^2 - N) = 6^2 + 6^2 - 6 = 66$. Para el cálculo de T_k debemos obtener las intersecciones de los objetos del cluster de la solución a) con los de la b). Así, según la matriz de contingencia de la Figura 1.c) obtenemos $T_k = \sum_{i=1}^{1}\sum_{j=1}^{2}(m_{ij}^2 - N) = 6^2 + 6^2 - 6 = 66$. De esta forma $B_k = 0{,}69$, valor que está por debajo de 1, aunque intuitivamente se esperaría que el índice pueda reflejar la similitud entre ambas soluciones. Esto es debido a que la probabilidad de que dos patrones cualesquiera se agrupen juntos en una de las soluciones es mucho menor que el valor obtenido.

3. Nuevo índice de estabilidad para clusters solapados

Como se ha observado a través del ejemplo anterior, cuando se trabaja con soluciones que poseen clusters solapados el índice FM es incapaz de producir resultados correctos. Ante la necesidad de contar con una herramienta que permita cuantificar la similitud de soluciones de clustering, ya sean estas solapadas o no, en esta sección se reinterpreta dicho índice y se propone uno nuevo denominado *overlapped FM* (oFM). La definición del nuevo índice parte de una forma más cuidadosa al calcular las probabilidades de que un patrón pueda agruparse junto con otro, ya sea en una solución, en la otra o en ambas a la vez. Particularmente, se debe tener especial cuidado en la normalización de las frecuencias, ya que se quiere evitar la contabilización repetida de los patrones sin considerar la cantidad de veces que aparezcan en otros clusters. Para que el nuevo índice pueda considerar la situación propuesta, se redefinirán cada uno de sus factores. Además, para ganar generalidad, ahora se considerarán explícitamente soluciones con distintos tamaños. Así se tomará k_1 para denotar la cantidad de clusters de C y k_2 para la de C'.

Con respecto a $\tilde{p_k}$, la nueva situación a ser contemplada admite escenarios en donde los patrones podrían llegar a estar en más de un cluster. Se debe tener especial consideración de no contabilizar dichos objetos más de una vez, o bien, normalizarlos para que su probabilidad de ocurrencia esté bien acotada. Para lograrlo, se debe tener en cuenta un caso extremo en el que todos los patrones estén en todos los clusters. A diferencia de la forma anterior de calcular este factor, ahora se lo debería normalizar por este caso extremo, dividiendo por algún valor que tenga relación directa con la cantidad de veces que podría llegar a contarse a todos los pares de patrones. Se redefine luego la nueva forma de calcular la probabilidad de que dos patrones se agrupen juntos en una misma solución como

$$\tilde{p_k} = \frac{\sum_{i=1}^{k_1}\binom{|c_i|}{2}}{k_1\binom{N}{2}}, \tag{14}$$

donde en el denominador se están considerando k_1 clusters con todos los patrones agrupados juntos, k_1 veces. En el numerador se cuentan los pares de patrones directamente tantas veces como clusters haya. Luego, para $\tilde{q_k}$, la situación es similar. Así se define

$$\tilde{q_k} = \frac{\sum_{j=1}^{k_2} \binom{|c'_j|}{2}}{k_2 \binom{N}{2}}. \tag{15}$$

El último factor a considerar es $\tilde{t_k}$, donde interviene la interacción de ambas soluciones. La cuenta de los pares de patrones agrupados es similar a la realizada en los casos anteriores. En cambio, la normalización de la misma involucra ciertos valores que surgen de ambas soluciones. Sean n_1 y n_2 la cantidad de elementos dentro de cada solución, considerando repeticiones por solapamiento. Luego, en la influencia recíproca de ambas soluciones, observamos intuitivamente que la cantidad de pares de elementos que podrían contarse está limitada por el valor más pequeño. Esto es, si $n_1 < n_2$, la cantidad de datos que puedan llegar a contarse en el emparejamiento entre ambas nunca podrá superar n_1, puesto que hay a lo sumo esa cantidad de objetos para contar en una de las soluciones. Así se tiene una noción de cómo se puede normalizar esta frecuencia, en donde la idea sigue el mismo camino que en las probabilidades de los factores anteriores. A su vez esta cantidad se la limita por la cantidad real de patrones N.

Por otro lado, hay que considerar también que la cantidad de clusters en cada solución puede ser diferente. Por ello, en el denominador se debe considerar también el caso límite en el que todos los patrones estén juntos en ambas soluciones, tantas veces como el máximo de solapamientos que haya entre las mismas. Este valor estará limitado por la proporción existente entre la cantidad máxima de clusters que existe en ambas soluciones multiplicado por la cantidad de formas de tomar todos los patrones de a pares. De esta forma se define $\tilde{t_k}$ como

$$\tilde{t_k} = \frac{\sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \binom{|c_i \cap c'_j|}{2}}{\binom{N}{2} \max(k_1, k_2) \frac{\min(n_1, n_2)}{N}}, \tag{16}$$

donde en el numerador se cuentan como antes las intersecciones de los patrones de cada uno de los clusters de una solución con los de la otra, tomados de a dos. El denominador, como se dijo anteriormente, representa el caso extremo en el que todos los patrones se agrupen juntos tantas veces como clusters haya en la solución.

Retomando el ejemplo de la Sección 2, el cálculo del nuevo índice arroja los siguientes valores. Para $\tilde{p_k} = \sum_{i=1}^{1} \binom{|c_i|}{2} / \binom{6}{2} = 15/15 = 1$. Luego $\tilde{q_k} = \sum_{j=1}^{2} \binom{|c'_j|}{2} / 2\binom{6}{2} = 30/30 = 1$. Y considerando $\max(1,2) = 2$ y $\min(6,12)/6 = 1$, se tiene $\tilde{t_k} = \sum_{i=1}^{1} \sum_{j=1}^{2} \binom{|c_i \cap c'_j|}{2} / 2\binom{6}{2} = 30/30 = 1$. Ahora sí se observa cómo el valor del nuevo índice oFM $= \tilde{t_k} / \sqrt{\tilde{p_k} \tilde{q_k}} = 1$ refleja la similitud esperada entre ambas soluciones, puesto que la probabilidad de encontrar dos patrones agrupados juntos en cualquiera de ellas es efectivamente 1.

4. Resultados y Discusión

En esta sección se presentan los resultados obtenidos en la evaluación de ambos índices. Primero se muestra y comenta una serie de ejemplos sencillos generados de forma artificial para observar el comportamiento de los índices en casos extremos y particulares. Luego se describe el conjunto de datos real utilizado en los experimentos y se presenta el algoritmo de clustering utilizado para agrupar los datos. A continuación se presentan los resultados obtenidos con ambos índices sobre dicho conjunto de datos y finalmente se contrastan los resultados de ambos.

Como se puede observar en la Tabla 1, se han creado 8 ejemplos artificiales que muestran situaciones de agrupamiento interesantes. En dicha tabla se numera cada caso con un número del I al $VIII$. Luego, se presentan dos columnas que representan las soluciones a ser comparadas: C y C'. Finalmente, las dos últimas columnas representan los valores de los índices FM y oFM, respectivamente. En el ejemplo I se observa claramente que ambas soluciones son exactamente las mismas. No existen clusters solapados y ambos índices logran mostrar correctamente las similitudas de C y C'. En II se observa cómo la solución C' posee sólo un cluster y cómo solamente la mitad de los patrones de la misma se podrían considerar similarmente agrupados en C, sean estos los que están en c_1 o c_2. De hecho, el valor de oFM es más cercano a 0,50, que es lo esperado.

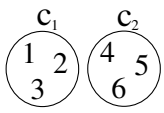
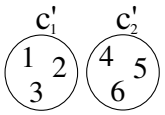
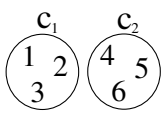
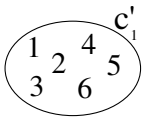
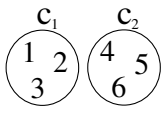
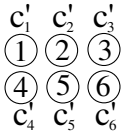
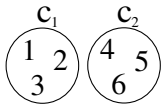
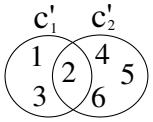
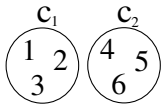
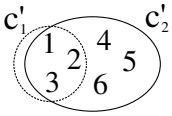
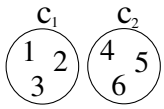
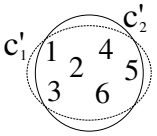
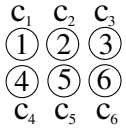
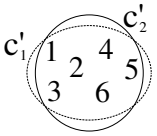
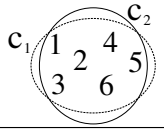
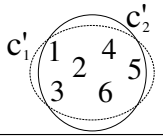
Ejemplos sobre datos artificiales				
	Soluciones		Índices	
	C	C'	**FM**	**oFM**
I			1,00	1,00
II			0,63	0,45
III			0,00	0,00
IV			0,68	0,82
V			0,54	0,87
VI			0,45	0,89
VII			0,17	0,00
VIII			0,49	1,00

Tabla 1: Resultados de los índices FM y oFM para algunos ejemplos de prueba artificiales.

clusters en C'	FM		oFM	
	$Vn = 0$	$Vn = 1$	$Vn = 0$	$Vn = 1$
$C = 4$ vs $C' = 25$	0,38	0,30	0,15	0,46
$C = 4$ vs $C' = 100$	0,17	0,16	0,03	0,13
$C = 25$ vs $C' = 100$	0,33	0,23	0,16	0,45

Tabla 2: Resultados de los índices FM y oFM en la base de datos Iris, para soluciones de referencia C de 4 y 25 clusters y sin solapamiento ($Vn = 0$) contra soluciones C' de 25 y 100 clusters tomando $Vn = 0$ y $Vn = 1$.

En el escenario mostrado en el ejemplo III, al igual que en I, ambos índices concuerdan con lo esperado. Ningún par de patrones puede agruparse junto en la solución C', y por ello esta solución no se asemeja en nada a la C. Con respecto a los ejemplo IV, V y VI se analizan en forma conjunta, ya que se observa cómo en cada una de las soluciones C' los clusters poseen progresivamente mayor solapamiento a medida que se avanza en cada ejemplo. Claramente se ve cómo FM decae cuando más solapamiento se considera, mientras que oFM se comporta de manera inversa. Este último comportamiento es el esperado, dado que al aumentar el nivel de solapamiento aumentan las posibilidades de obtener más pares de patrones agrupados juntos en la otra solución. En el ejemplo VII se observa una situación similar a la de III, en cuanto a que en una de las soluciones no se pueden agrupar patrones de a pares, y por ello oFM arroja una similitud nula; mientras que FM muestra algún grado de parecido que no se corresponde con la realidad. Por último, en el ejemplo $VIII$ se ve una situación similar a la de I ya que ambas soluciones son las mismas, aunque ahora con clusters totalmente solapados. Se observa como oFM refleja correctamente esta semejanza entre las soluciones con un valor = 1 y como FM falla debido al solapamiento.

Luego para los experimentos con datos reales se utilizó el conjunto de datos de la flor del Iris[1]. Este conjunto de datos con 4 atributos posee 50 registros de cada una de 3 especies distintas de la flor, haciendo un total de 150 patrones [17]. De las tres clases existentes en el conjunto, sólo una es linealmente separable de las otras dos, teniendo estas últimas patrones de distinta clase muy cercanos entre sí en el espacio de atributos. Se ha seleccionado este conjunto por ser sencillo y pequeño, además del acceso libre y su aceptación en el ámbito científico.

Para el agrupamiento de los datos se utilizó un mapa auto-organizativo (SOM, de su nombre en inglés *Self-Organizing Map*) [18,19]. Para el agrupamiento se entrenaron mapas con distintas cantidades de neuronas, es decir, clusters. Todos los mapas utilizados tienen una topología rectangular en forma de grilla y la cantidad de iteraciones de entrenamiento fue fijada en 100 épocas. La inicialización del mapa fue determínistica, utilizando PCA [20] sobre los datos. Para considerar solapamiento en las neuronas del mapa se tomaron las neuronas vecinas inmediatas, es decir la de la izquierda, derecha, arriba y abajo, de cada neurona. Así, al considerar vecindad $Vn = 0$ cada neurona es un único cluster

[1] http://archive.ics.uci.edu/ml/datasets/Iris

y $Vn = 1$ cada neurona y sus cuatro vecinas forman un mismo cluster. Por ello, varios patrones pueden llegar a pertenecer a más de un cluster y de esta forma pertenecer a clusters solapados.

En la Tabla 2 se observa una primer columna en donde se indican las cantidades de clusters considerados para las soluciones C y C'. Luego la tabla se divide en dos grandes columnas que representan los valores de los índices analizados, FM y oFM. Cada una de estas columnas se vuelven a dividir para mostrar los resultados de las pruebas con soluciones C' que poseen y no poseen solapamientos. La cantidad de clusters de las soluciones fueron de 4, 25 y 100. Para la solución tomada como referencia C no se consideró solapamiento. Cuando se toma $C = 4$ FM disminuye su valor no sólo al tomar solapamiento sino también al pasar de $C' = 25$ a $C' = 100$. Por otro lado oFM aumenta notablemente al considerar solapamiento, pero también disminuye al aumentar la cantidad de clusters en la solución C'. Por último, vemos como FM vuelve a decaer al considerar solapamiento cuando $C = 25$ y $C' = 100$. Se puede observar además cómo oFM mantiene la tendencia de aumentar al considerar solapamiento pero, al igual que FM, aumenta al pasar de $C = 4$ a $C = 25$ cuando $C' = 100$. En todos los casos el valor de FM decae cuando aumenta el solapamiento para la solución C'. De manera opuesta sucede para oFM. Se ve además que en todos los casos los dos índices mejoraron sus valores a medida que la cantidad de clusters de C se asemeja a la de C'. Esto es debido a que a mayor cantidad de clusters hay mayor dispersión de los patrones y por ello a mayor diferencia entre las cantidades de clusters entre C y C', menor es el valor de los índices calculados.

5. Conclusiones y trabajos futuros

En este trabajo se propuso un nuevo índice para análisis de estabilidad entre soluciones solapadas, parcial o completamente. Se realizó un análisis del índice FM para mostrar sus falencias al momento de aplicarlo sobre soluciones solapadas y a partir de éste se derivó uno nuevo. El índice propuesto refleja correctamente la similitud entre soluciones que pueden o no poseer clusters solapados. Esto se alcanzó teniendo especial cuidado en la forma de contar casos extremos en los que se pudiesen agrupar los patrones, permitiendo así una normalización más cuidadosa. Luego, en los experimentos se mostró como el índice FM es incapaz de tratar con soluciones solapadas, mientras que oFM puede sortear dicho problema.

Con respecto a trabajos futuros, se espera realizar mayor cantidad de pruebas de los índices utilizando otros algoritmos y otros conjuntos de datos. También sería deseable a futuro la incorporación al análisis de nuevos tipos de índices y derivar así nuevas métricas que se puedan comparar con las de este trabajo.

Referencias

[1] Xu, R., Wunsch, D.C.: Clustering. Wiley and IEEE Press (2009) 1

[2] Skillicorn, D.: Understanding Complex Datasets: Data Mining with Matrix Decompositions. CRC Press (May 2007) 1

[3] Ben-David, S., Luxburg, U.V., Pál, D.: A sober look at clustering stability. In: In COLT, Springer (2006) 5–19 1

[4] Shamir, O., Tishby, N.: Stability and model selection in -means clustering. Machine Learning (2010) 213–243 1

[5] Ben-Hur, A., Elisseeff, A., Guyon, I.: A stability based method for discovering structure in clustered data. In: Pacific Symposium on Biocomputing'02. (2002) 6–17 1

[6] Bayá, A.: Aplicación de algoritmos no supervisados a datos biológicos. PhD thesis, Universidad Nacional de Rosario (3 2011) Doctorado en ingeniería. 1

[7] Nguyen, X.V., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. Journal of Machine Learning Research **11** (2010) 2837–2854 1

[8] Halkidi, M., Batistakis, Y., Vazirgiannis, M.: On clustering validation techniques. Journal of Intelligent Informations Systems **17**(1) (2001) 107–145 1

[9] Handl, J., Knowles, J., Kell, D.B.: Computational cluster validation in post-genomic data analysis. Bioinformatics **21**(15) (2005) 3201–3212 1

[10] Fowlkes, E.B., Mallows, C.L.: A Method for Comparing Two Hierarchical Clusterings. Journal of the American Statistical Association **78**(383) (September 1983) 553–569 1, 2

[11] Ben-Hur, A., Guyon, I.: Detecting Stable Clusters Using Principal Component Analysis. In Brownstein, M., Khodursky, A., eds.: Functional Genomics. Number 224 in Methods in Molecular Biology. Humana Press (January 2003) 159–182 1

[12] Meila, M., Heckerman, D.: An experimental comparison of model-based clustering methods. Machine Learning **42** (January 2001) 9–29 1

[13] Meilă, M.: Comparing clusterings—an information based distance. Journal of Multivariate Analysis **98** (May 2007) 873–895 1

[14] Goldberg, M.K., Hayvanovych, M., Magdon-Ismail, M.: Measuring similarity between sets of overlapping clusters. In: 2010 IEEE Second International Conference on Social Computing. (Aug 2010) 303–308 1

[15] Wu, J., Yuan, H., Xiong, H., Chen, G.: Validation of overlapping clustering: A random clustering perspective. Information Sciences **180**(22) (2010) 4353 – 4369 1

[16] Amigó, E., Gonzalo, J., Artiles, J., Verdejo, F.: A comparison of extrinsic clustering evaluation metrics based on formal constraints. Inf. Retr. **12**(4) (August 2009) 461–486 1

[17] Fisher, R.A.: The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics **7**(2) (September 1936) 179–188 4

[18] Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics **43** (1982) 59–69 4

[19] Milone, D., Stegmayer, G., Kamenetzky, L., Lopez, M., Giovannoni, J., Lee, J.M., Carrari, F.: *omeSOM: a software for integration, clustering and visualization of transcriptional and metabolite data mined from interspecific crosses of crop plants. BMC Bioinformatics **11** (2010) 438–448 4

[20] F.R.S., K.P.: Liii. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **2**(11) (1901) 559–572 4

Apéndice B

Análisis de estabilidad en clústers solapados

Campo, D. N., Stegmayer, G., y Milone, D. H., "Análisis de estabilidad en clústers solapados", *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, vol. 17, núm. 53, enero-junio, 2014, pp. 79-89 Asociación Española para la Inteligencia Artificial.
En este trabajo se analizó un índice de valadación externa desde un enfoque probabilístico y se propuso una formulación alternativa aplicable a clústers solapados. Los resultados muestran cómo la nueva formulación puede medir conjuntos de datos tanto artificiales como reales y con características solapados.

Análisis de estabilidad en clusters solapados

†David N. Campo[1,2], ⋆Georgina Stegmayer[1,2] y ‡Diego H. Milone[2]

[1] CIDISI – Universidad Tecnológica Nacional – Facultad Regional Santa Fe
[2] sinc(i) – Universidad Nacional del Litoral – Facultad de Ingeniería y Ciencias Hídricas
†dncampo@santafe-conicet.gov.ar, ⋆gstegmayer@santafe-conicet.gov.ar, ‡d.milone@ieee.org.

Resumen Analizar la estabilidad de una solución de clustering implica medir la capacidad de un algoritmo para producir resultados similares dada una misma fuente de datos de entrada. Los índices de validación externa permiten cuantificar dicha similitud entre un par de soluciones de clustering. Dentro de los índices clásicos más utilizados es posible validar soluciones con clusters no solapados, en donde cada patrón sólo puede pertenecer a un cluster. Sin embargo, en aplicaciones prácticas, generalmente se dan situaciones en las que un patrón podría poseer más de una etiqueta. En este trabajo se analiza un índice de validación externa desde un enfoque probabilístico y se provee una reformulación aplicable a soluciones con clusters solapados. Luego de presentar el nuevo índice, se muestran y discuten resultados de experimentos realizados sobre ejemplos artificiales y bases de datos reales. Los resultados muestran cómo el nuevo índice puede medir adecuadamente la similitud entre clusters solapados, permitiendo así analizar la estabilidad en ambos casos.

Palabras Clave: análisis de estabilidad, clusters solapados, medida de validación.

1. Introducción

Los algoritmos de clustering reciben un conjunto de datos como entrada y mediante un proceso no-supervisado lo particionan en cierto número de clusters o grupos. Se puede definir un cluster como un grupo de objetos homogéneos que poseen alguna medida de similitud entre ellos y que se muestran diferentes a los objetos agrupados en otros clusters [21, 19]. Dado que la aplicación de estos algoritmos sobre una base de datos siempre devuelve algún resultado, incluso cuando no exista estructura en la misma, ha surgido el análisis de estabilidad de soluciones de clustering. Para realizar este análisis es importante medir la capacidad de un algoritmo de agrupamiento para producir grupos similares de forma repetida [4, 18, 5]. Aunque no hay un total acuerdo en cuanto a su definición, existen autores que relacionan el concepto de estabilidad con soluciones de agrupamiento que no se ven alteradas bajo alguna perturbación de los datos de entrada [5, 3]. Cuando se hace análisis de estabilidad en clustering, también se suele hablar de que se mantengan las estructuras naturales "subyacentes en los datos" y no que aparezcan nuevas estructuras como producto artificial de un algoritmo concreto.

Luego de realizar el agrupamiento en distintas condiciones, se desea medir o cuantificar la similitud entre las soluciones para poder compararlas [21]. Ultimamente, con el crecimiento en el estudio de análisis de agrupamientos se han

estado utilizando nuevos algoritmos y medidas de comparación de clusters [16]. Para medir la similitud entre soluciones de clustering se pueden utilizar dos tipos de medidas de validación: internas y externas. Las primeras miden atributos de homogeneidad y separación de los datos en los clusters. Las medidas externas hacen una comparación entre distintas soluciones de clustering, tomando una como referencia y comparándola con otras [10, 11]. Con respecto a las medidas externas, se dispone principalmente de tres tipos para realizar la comparación entre soluciones: las basadas en el conteo de pares de patrones, en las que las soluciones coinciden o no; las basadas en el análisis de correspondencia de conjuntos y las basadas en la estadística y teoría de la información. Con respecto a las primeras, la más utilizada y difundida es la de Fowlkes-Mallows (FM) [8] que trabaja con las frecuencias de pares de patrones que se detecten agrupados juntos en ambas soluciones. En [6] los autores evalúan las distintas soluciones obtenidas cuantificando la similaridad mediante este índice. De las métricas basadas en conjuntos, una muy utilizada es la de máxima coincidencia [14] que se basa en comparar los clusters de ambas soluciones analizadas, haciendo coincidir aquellos que tengan mayor cantidad de elementos en común. Por último, de las medidas basadas en estadística, la información mutua normalizada es una de las más representativas y se basa en cuantificar la cantidad de información compartida entre ambas soluciones, a través del concepto de entropía [16, 13]. Sin embargo, ninguna de estas medidas es capaz de representar correctamente las similitudes al trabajar con soluciones que posean clusters solapados.

Ultimamente se ha suscitado interés por el análisis de soluciones con clusters solapados. En [9] se compara y estudia la evolución de grupos de personas en redes sociales y se propone un algoritmo para computar nuevas distancias entre colecciones de grupos potencialmente solapados. En [20] se propone una nueva medida de validación para clusters solapados en el contexto de recuperación de documentos, en particular tratando el tema de solapamiento entre soluciones con diferente número de clusters. En [1] se propone un conjunto de restricciones sobre métricas para validación externa y se extiende el análisis para tratar soluciones solapadas. En este trabajo, en cambio, se hace un aporte al área ya que presentamos un análisis detallado del índice FM. Con un enfoque probabilístico se analiza cada factor del mismo considerando el solapamiento de clusters y se muestra cómo falla este índice cuando se lo intenta aplicar a soluciones que posean clusters solapados. Luego, a partir del análisis anterior, se deriva una nueva propuesta de índice teniendo especial atención en la situación mencionada. Aplicando ambos índices a diferentes casos de estudio y una base de datos reales, se verifican experimentalmente las mejoras expresadas cuando existen clusters solapados.

La organización de este trabajo es la siguiente. En la Sección 2 se realiza un análisis detallado de los factores del índice FM. En la Sección 3 se deriva el nuevo índice propuesto para poder tratar con las soluciones solapadas. Luego, en la Sección 4, se presentan resultados de aplicar ambos índices sobre datos artificiales y reales, y se discuten dichos experimentos. Por último se presentan las conclusiones del trabajo en la Sección 5.

2. Análisis conceptual del índice Fowlkes-Mallows

El índice de Fowlkes-Mallows es una medida de similaridad entre soluciones de clustering, generalmente utilizado para validación externa. Como se mencionó en la Sección 1, en el contexto de análisis de estabilidad se utiliza para poder comparar soluciones y verificar si las mismas son o no estables. FM recibe el etiquetado de ambas soluciones, sobre el mismo conjunto de datos, y devuelve un valor $B_k \in [0, 1]$ que cuantifica la similitud entre las mismas. El valor 1 representa que ambas son exactamente iguales, mientras que 0 denota completa desigualdad. FM está definido en [8] como

$$B_k = \frac{T_k}{\sqrt{P_k Q_k}}, \tag{1}$$

donde

$$T_k = \sum_{i=1}^{k}\sum_{j=1}^{k} m_{ij}^2 - N, \tag{2}$$

$$P_k = \sum_{i=1}^{k} m_{i*}^2 - N, \tag{3}$$

$$Q_k = \sum_{j=1}^{k} m_{*j}^2 - N, \tag{4}$$

$$m_{i*} = \sum_{j=1}^{k} m_{ij}, \tag{5}$$

$$m_{*j} = \sum_{i=1}^{k} m_{ij}, \tag{6}$$

siendo $M = \{m_{ij}\}$ la matriz de contingencia obtenida entre 2 soluciones de k clusters cada una con N datos en el conjunto a particionar. Esta matriz posee tantas filas como cantidad de clusters haya en la primer solución, que llamaremos C, y tantas columnas como clusters tenga la segunda, C'. En cada posición ij de la matriz se coloca la cantidad de patrones en comun que se pueden contar entre los elementos del cluster i de la solución C y los patrones del cluster j de la solución C'. Es decir, $m_{ij} = |c_i \cap c'_j|$.

A continuación se hará un análisis conceptual de cada factor del índice FM. Consideremos T_k, y reemplazando $\sum_{i=1}^{k}\sum_{j=1}^{k} m_{ij} = N$ en (2) se puede escribir

$$\begin{aligned} T_k &= \sum_{i=1}^{k}\sum_{j=1}^{k} m_{ij}^2 - N = \sum_{i=1}^{k}\sum_{j=1}^{k}(m_{ij}m_{ij}) - \sum_{i=1}^{k}\sum_{j=1}^{k} m_{ij} \\ &= \sum_{i=1}^{k}\sum_{j=1}^{k} m_{ij}(m_{ij}-1) = 2\sum_{i=1}^{k}\sum_{j=1}^{k}\binom{m_{ij}}{2}, \end{aligned} \tag{7}$$

que representa el doble de la sumatoria de cada uno de los elementos de la matriz de contingencia tomados de a 2. Es decir, los elementos en común entre cada par de clusters de ambas soluciones tomados de a 2. Esto representa la cantidad de formas de elegir dos elementos en dicha intersección.

Análogamente, a partir de (3) se puede obtener

$$\begin{aligned} P_k &= \sum_{i=1}^{k} m_{i*}^2 - N = \sum_{i=1}^{k}(m_{i*}m_{i*}) - \sum_{i=1}^{k}\sum_{j=1}^{k} m_{ij} \\ &= \sum_{i=1}^{k}(m_{i*}m_{i*}) - \sum_{i=1}^{k} m_{i*} = \sum_{i=1}^{k}(m_{i*}(m_{i*})-1) = 2\sum_{i=1}^{k}\binom{m_{i*}}{2}, \quad (8) \end{aligned}$$

siendo, de forma similar a como ocurre con T_k, la cantidad de formas que hay de tomar 2 elementos de cada cluster de la solución analizada. Idéntico razonamiento se aplica para la interpretación de Q_k.

Consideremos C y C', dos soluciones de clustering con k clusters. Sean x e y dos patrones cualesquiera. Podemos interpretar al primer factor, P_k, como la probabilidad de que los dos patrones se encuentren en un mismo cluster c_i de la solución C, considerando un muestreo uniforme de los datos. Dado que los patrones tienen la misma probabilidad de ser seleccionados, al tomar un segundo patrón bajo la misma hipótesis, la probabilidad de que los dos se agrupen en el cluster c_i es:

$$Pr(x \in c_i \wedge y \in c_i) = \frac{\binom{|c_i|}{2}}{\binom{N}{2}} = \frac{|c_i|\,(|c_i|-1)}{N(N-1)}, \quad (9)$$

en donde el numerador representa la cantidad de formas de tomar de a 2 los patrones agrupados bajo el cluster c_i y el denominador considera todas las formas posibles de tomar 2 elementos del conjunto completo de datos. Esto último surge de considerar un caso extremo en el cual todos los patrones queden en un único cluster. Si se consideran los k clusters de C, ahora tenemos

$$\tilde{p_k} = \sum_{i=1}^{k} \frac{\binom{|c_i|}{2}}{\binom{N}{2}} = \frac{1}{N(N-1)}\sum_{i=1}^{k} |c_i|\,(|c_i|-1) = \frac{1}{N(N-1)}\sum_{i=1}^{k} |c_i|^2 - N, \quad (10)$$

que guarda relación directa con (3), dado que $|c_i| = m_{i*}$. La interpretación de Q_k es la misma que la de P_k, salvo que se analizan los clusters de la solución C'.

Con respecto a T_k, se puede considerar que representa la probabilidad de muestrear un par de patrones al azar y que éstos pertenezcan a un mismo cluster en C y C'. El razonamiento es similar al seguido para P_k, sólo que en vez de considerarse todos los patrones por cluster, se van considerando los pares de patrones comunes al emparejamiento entre cada par de clusters de ambas soluciones

$$Pr((x,y) \in c_i \wedge (x,y) \in c'_j) = \frac{\binom{|c_i \cap c'_j|}{2}}{\binom{N}{2}} = \frac{|c_i \cap c'_j|\,(|c_i \cap c'_j|-1)}{N(N-1)}. \quad (11)$$

De la misma forma que en el desarrollo de P_k, el denominador representa la cantidad de formas de tomar 2 elementos de todo el conjunto de datos. Este sería el caso extremo en el que en ambas soluciones, todos los patrones hayan quedado

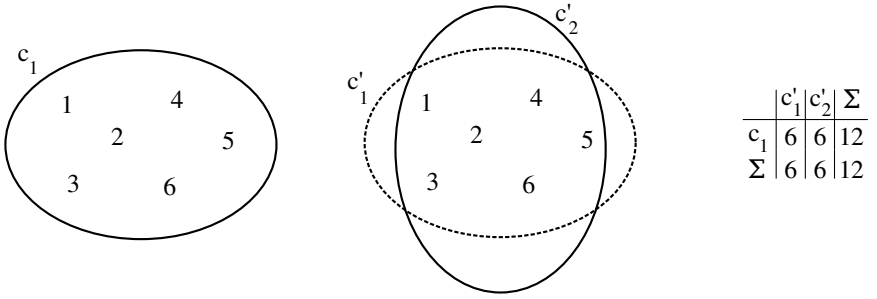


Figura 1: Ejemplo ilustrativo con dos soluciones, a) C con $k = 1$, b) C' con solapamiento y $k = 2$ y c) matriz de contingencia correspondiente a las soluciones C y C'.

en un único cluster. Si tomamos entonces todas las posibles comparaciones entre cada par de clusters de ambas soluciones, obtenemos la probabilidad

$$\tilde{t_k} = \sum_{i=1}^{k}\sum_{j=1}^{k} \frac{\binom{|c_i \cap c'_j|}{2}}{\binom{N}{2}} = \frac{1}{N(N-1)} \sum_{i=1}^{k}\sum_{j=1}^{k} |c_i \cap c'_j| \left(|c_i \cap c'_j| - 1\right) =$$

$$= \frac{1}{N(N-1)} \sum_{i=1}^{k}\sum_{j=1}^{k} |c_i \cap c'_j|^2 - N, \qquad (12)$$

que guarda relación directa con (2) dado que, como se observo antes, $m_{ij} = |c_i \cap c'_j|$.

Luego, el índice FM puede entenderse como la razón entre la probabilidad de obtener dos patrones juntos en las soluciones conjuntas sobre la media geométrica de dicha probabilidad considerando las soluciones por separado. Esto es:

$$B_k = \frac{\cancel{\frac{1}{\binom{N}{2}}} \sum_{i=1}^{k}\sum_{j=1}^{k} \binom{|c_i \cap c'_j|}{2}}{\sqrt{\cancel{\frac{1}{\binom{N}{2}}} \sum_{i=1}^{k} \binom{|c_i|}{2} \cancel{\frac{1}{\binom{N}{2}}} \sum_{j=1}^{k} \binom{|c'_j|}{2}}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} (m_{ij}^2 - N)}{\sqrt{\sum_{i=1}^{k} (m_{i*}^2 - N) \sum_{j=1}^{k} (m_{*j}^2 - N)}} = \frac{T_k}{\sqrt{P_k Q_k}}. \qquad (13)$$

De esta manera se arriba a la formulación original de FM a traves del enfoque probabilístico alternativo.

Ahora bien, cuando este índice se aplica a clusters solapados, los resultados que se obtienen no son los intuitivamente esperados. Por ejemplo, en el gráfico de la Figura 1 se observan dos soluciones de clustering a) C con $k = 1$ y b) C' con $k = 2$. En c) se encuentra la matriz de contingencia para ambas soluciones. En la solución a), se puede observar cómo todos los patrones se agruparon juntos en un único cluster. En la solución b) se puede observar cómo ambos clusters comparten todos los patrones. En este caso hay un solapamiento completo de ambos grupos.

Para calcular el índice FM sobre este ejemplo se procederá a calcular cada uno de sus factores. El valor del factor P_k se corresponde con la solución C y el de Q_k con la C'. Así $P_k = \sum_{i=1}^{1}(m_{i*}^2 - N) = 12^2 - 6 = 138$. En cambio $Q_k = \sum_{i=1}^{2}(m_{*j}^2 - N) = 6^2 + 6^2 - 6 = 66$. Para el cálculo de T_k debemos obtener las intersecciones de los objetos del cluster de la solución a) con los de la b). Así, según la matriz de contingencia de la Figura 1.c) obtenemos $T_k = \sum_{i=1}^{1}\sum_{j=1}^{2}(m_{ij}^2 - N) = 6^2 + 6^2 - 6 = 66$. De esta forma $B_k = 0{,}69$, valor que está por debajo de 1, aunque intuitivamente se esperaría que el índice pueda reflejar la similitud entre ambas soluciones. Esto es debido a que la probabilidad de que dos patrones cualesquiera se agrupen juntos en una de las soluciones es mucho menor que el valor obtenido.

3. Índice de estabilidad para clusters solapados

Como se ha observado a través del ejemplo anterior, cuando se trabaja con soluciones que poseen clusters solapados el índice FM es incapaz de producir resultados correctos. Ante la necesidad de contar con una herramienta que permita cuantificar la similitud de soluciones de clustering, ya sean estas solapadas o no, en esta sección se reinterpreta dicho índice y se propone uno nuevo denominado *overlapped FM* (oFM). La definición del nuevo índice parte de una forma más cuidadosa al calcular las probabilidades de que un patrón pueda agruparse junto con otro, ya sea en una solución, en la otra o en ambas a la vez. Particularmente, se debe tener especial cuidado en la normalización de las frecuencias, ya que se quiere evitar la contabilización repetida de los patrones sin considerar la cantidad de veces que aparezcan en otros clusters. Para que el nuevo índice pueda considerar la situación propuesta, se redefinirán cada uno de sus factores. Además, para ganar generalidad, ahora se considerarán explícitamente soluciones con distintos tamaños. Así se tomará k_1 para denotar la cantidad de clusters de C y k_2 para la de C'.

Con respecto a $\tilde{p_k}$, la nueva situación a ser contemplada admite escenarios en donde los patrones podrían llegar a estar en más de un cluster. Se debe tener especial consideración de no contabilizar dichos objetos más de una vez, o bien, normalizarlos para que su probabilidad de ocurrencia esté bien acotada. Para lograrlo, se debe tener en cuenta un caso extremo en el que todos los patrones estén en todos los clusters. A diferencia de la forma anterior de calcular este factor, ahora se lo debería normalizar por este caso extremo, dividiendo por algún valor que tenga relación directa con la cantidad de veces que podría llegar a contarse a todos los pares de patrones. Se redefine luego la nueva forma de calcular la probabilidad de que dos patrones se agrupen juntos en una misma solución como

$$\tilde{p_k} = \frac{\sum_{i=1}^{k_1}\binom{|c_i|}{2}}{k_1\binom{N}{2}}, \tag{14}$$

donde en el denominador se están considerando k_1 clusters con todos los patrones agrupados juntos, k_1 veces. En el numerador se cuentan los pares de patrones directamente tantas veces como clusters haya. Luego, para $\tilde{q_k}$, la situación es similar. Así se define

$$\tilde{q_k} = \frac{\sum_{j=1}^{k_2} \binom{|c'_j|}{2}}{k_2 \binom{N}{2}}. \tag{15}$$

El último factor a considerar es $\tilde{t_k}$, donde interviene la interacción de ambas soluciones. La cuenta de los pares de patrones agrupados es similar a la realizada en los casos anteriores. En cambio, la normalización de la misma involucra ciertos valores que surgen de ambas soluciones. Sean n_1 y n_2 la cantidad de elementos dentro de cada solución, considerando repeticiones por solapamiento. Luego, en la influencia recíproca de ambas soluciones, observamos intuitivamente que la cantidad de pares de elementos que podrían contarse está limitada por el valor más pequeño. Esto es, si $n_1 < n_2$, la cantidad de datos que puedan llegar a contarse en el emparejamiento entre ambas nunca podrá superar n_1, puesto que hay a lo sumo esa cantidad de objetos para contar en una de las soluciones. Así se tiene una noción de cómo se puede normalizar esta frecuencia, en donde la idea sigue el mismo camino que en las probabilidades de los factores anteriores. A su vez esta cantidad se la limita por la cantidad real de patrones N.

Por otro lado, hay que considerar también que la cantidad de clusters en cada solución puede ser diferente. Por ello, en el denominador se debe considerar también el caso límite en el que todos los patrones estén juntos en ambas soluciones, tantas veces como el máximo de solapamientos que haya entre las mismas. Este valor estará limitado por la proporción existente entre la cantidad máxima de clusters que existe en ambas soluciones multiplicado por la cantidad de formas de tomar todos los patrones de a pares. De esta forma se define $\tilde{t_k}$ como

$$\tilde{t_k} = \frac{\sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \binom{|c_i \cap c'_j|}{2}}{\binom{N}{2} \max(k_1, k_2) \frac{\min(n_1, n_2)}{N}}, \tag{16}$$

donde en el numerador se cuentan como antes las intersecciones de los patrones de cada uno de los clusters de una solución con los de la otra, tomados de a dos. El denominador, como se dijo anteriormente, representa el caso extremo en el que todos los patrones se agrupen juntos tantas veces como clusters haya en la solución.

Retomando el ejemplo de la Sección 2, el cálculo del nuevo índice arroja los siguientes valores. Para $\tilde{p_k} = \sum_{i=1}^{1} \binom{|c_i|}{2} / \binom{6}{2} = 15/15 = 1$. Luego $\tilde{q_k} = \sum_{j=1}^{2} \binom{|c'_j|}{2} / 2\binom{6}{2} = 30/30 = 1$. Y considerando $\max(1,2) = 2$ y $\min(6,12)/6 = 1$, se tiene $\tilde{t_k} = \sum_{i=1}^{1} \sum_{j=1}^{2} \binom{|c_i \cap c'_j|}{2} / 2\binom{6}{2} = 30/30 = 1$. Ahora sí se observa cómo el valor del nuevo índice oFM $= \tilde{t_k}/\sqrt{\tilde{p_k}\tilde{q_k}} = 1$ refleja la similitud esperada entre ambas soluciones, puesto que la probabilidad de encontrar dos patrones agrupados juntos en cualquiera de ellas es efectivamente 1.

4. Resultados y Discusión

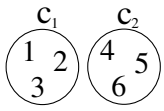
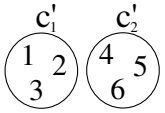
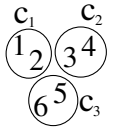
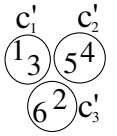
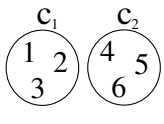
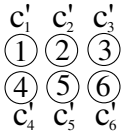
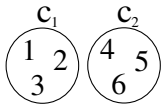
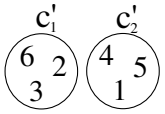
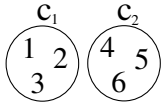
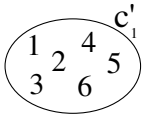
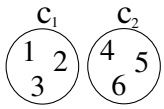
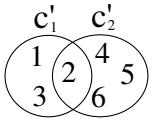
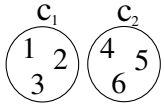
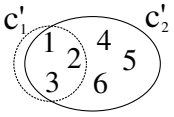
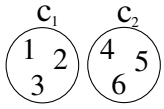
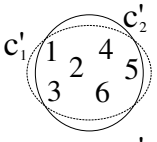
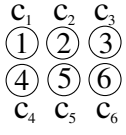
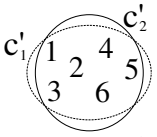
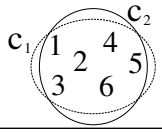
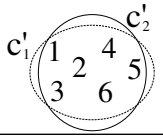
En esta sección se presentan los resultados obtenidos en la evaluación de ambos índices. Primero se muestra y comenta una serie de ejemplos sencillos generados de forma artificial para observar el comportamiento de los índices en casos extremos y particulares. Luego se describen los conjuntos de datos reales

utilizados en los experimentos y se presenta el algoritmo de clustering utilizado para agrupar los datos. A continuación se presentan los resultados obtenidos con ambos índices sobre dichos conjuntos de datos y finalmente se contrastan los resultados de ambos.

4.1. Casos de estudio

Como se puede observar en el Cuadro 1, se han creado 10 ejemplos artificiales que muestran situaciones de agrupamiento interesantes. En dicho cuadro se enumera cada caso del I al X. Luego, se presentan dos columnas que representan las soluciones a ser comparadas: C y C'. Finalmente, las dos últimas columnas representan los valores de los índices FM y oFM, respectivamente. En el ejemplo I se observa claramente que ambas soluciones son exactamente las mismas. No existen clusters solapados y ambos índices logran mostrar correctamente las similitudas de C y C'. En II se puede apreciar como cada solución posee 3 clusters, pero en ningún caso existen pares de patrones que se puedan encontrar simultáneamente en ambas. Así los dos índices logran, nuevamente, reflejar correctamente la situación observada. Algo similiar ocurre con el siguiente ejemplo, el III. En este escenario la solución C' presenta la particularidad de que no es posible agrupar patrones de a pares puesto que está cada uno en un cluster distinto, y por ello esta solución no se asemeja en nada a la C.

El ejemplo IV se plantea un caso similar a I pero con un par de patrones intercambiados en los clusters de la solución C'. Esto provoca que la cantidad total de pares de patrones que se pueden contabilizar en ambas soluciones a la vez decaiga en una proporción de 3 con respecto al ejemplo I anteriormente mencionado, y esto se refleja perfectamente tanto en el valor de FM como en el de oFM. En V se observa cómo la solución C' posee sólo un cluster y cómo solamente la mitad de los patrones de la misma se podrían considerar similarmente agrupados en C, sean estos los que están en c_1 o c_2. De hecho, el valor de oFM es más cercano a 0,50, que es lo esperado.

Ejemplos sobre datos artificiales				
	Soluciones		Índices	
	C	C'	**FM**	**oFM**
I			1,00	1,00
II			0,00	0,00
III			0,00	0,00
IV			0,33	0,33
V			0,63	0,45
VI			0,68	0,82
VII			0,54	0,87
VIII			0,45	0,89
IX			0,17	0,00
X			0,49	1,00

Cuadro 1: Resultados de los índices FM y oFM para algunos ejemplos de prueba artificiales.

	clusters en C'	**FM**		**oFM**	
		$Vn = 0$	$Vn = 1$	$Vn = 0$	$Vn = 1$
Iris	$k_1 = 4$ vs $k_2 = 25$	0,38	0,30	0,15	0,46
	$k_1 = 4$ vs $k_2 = 100$	0,17	0,16	0,03	0,13
	$k_1 = 25$ vs $k_2 = 100$	0,33	0,23	0,16	0,45
Wine	$k_1 = 4$ vs $k_2 = 25$	0,40	0,31	0,16	0,48
	$k_1 = 4$ vs $k_2 = 100$	0,19	0,18	0,04	0,15
	$k_1 = 25$ vs $k_2 = 100$	0,34	0,23	0,17	0,43
Yeast	$k_1 = 4$ vs $k_2 = 25$	0,32	0,23	0,13	0,39
	$k_1 = 4$ vs $k_2 = 100$	0,16	0,14	0,03	0,12
	$k_1 = 25$ vs $k_2 = 100$	0,29	0,18	0,14	0,38
Glass	$k_1 = 4$ vs $k_2 = 25$	0,33	0,27	0,13	0,41
	$k_1 = 4$ vs $k_2 = 100$	0,15	0,14	0,03	0,12
	$k_1 = 25$ vs $k_2 = 100$	0,38	0,24	0,19	0,50

Cuadro 2: Resultados de los índices FM y oFM en la base de datos Iris, Wine, Yeast y Glass para soluciones de referencia C de 4 y 25 clusters y sin solapamiento ($Vn = 0$) contra soluciones C' de 25 y 100 clusters tomando $Vn = 0$ y $Vn = 1$.

Los ejemplo VI, VII y $VIII$ se analizan en forma conjunta, ya que se observa cómo en cada una de las soluciones C' los clusters poseen progresivamente mayor solapamiento a medida que se avanza en cada ejemplo. Claramente se ve cómo FM decae cuando más solapamiento se considera, mientras que oFM se comporta de manera inversa. Este último comportamiento es el esperado, dado que al aumentar el nivel de solapamiento aumentan las posibilidades de obtener más pares de patrones agrupados juntos en la otra solución. En el ejemplo IX se observa una situación similar a la de III, en cuanto a que en una de las soluciones no se pueden agrupar patrones de a pares, y por ello oFM arroja una similitud nula; mientras que FM muestra algún grado de parecido que no se corresponde con la realidad. Por último, en el ejemplo X se ve una situación similar a la de I ya que ambas soluciones son las mismas, aunque ahora con clusters totalmente solapados. Se observa como oFM refleja correctamente esta semejanza entre las soluciones con un valor = 1 y FM falla debido al solapamiento.

4.2. Datos reales

Para los experimentos con datos reales se utilizaron 4 bases de datos: el conjuntos de datos de la flor del Iris[1], de Wine[2], de Yeast[3] y de Glass[4] [2] El conjunto de datos de Iris posee 4 atributos y 50 registros de cada una de 3 especies distintas de la flor, haciendo un total de 150 patrones [7]. De las tres clases existentes en el conjunto, sólo una es linealmente separable de las otras dos, teniendo estas últimas patrones de distinta clase muy cercanos entre sí en el espacio de atributos. Se ha seleccionado este conjunto por ser sencillo y pequeño, además del acceso libre y su aceptación en el ámbito científico. Por su parte, Wine representa la medición y análisis de 13 atributos químicos realizados sobre vinos de una misma región de Italia, pero tomados de diferentes cultivos. Este conjunto de datos consta de 178 patrones distribuidos en 3 grupos: cultivo A con 59 patrones, el B con 71 y el C con 48. La base de datos de Yeast, por su parte, representa un estudio sobre la levadura en donde se busca determinar la localización de sus proteínas en las células. Posee 1484 patrones distribuidos en 10 grupos; con 463, 429, 244, 163, 51, 44, 37, 30, 20 y 5 instancias en cada uno. A cada patrón se le realizaron 8 tipos de mediciones. Por último, el conjunto Glass posee 214 patrones distribuidos en 7 grupos o tipos de vidrios, en dónde a cada objeto se le han medido 9 atributos. Entre ellos se encuentran en índice de refracción del vidrio y el contenido de óxido de distintos elementos como ser sodio, magnesio, aluminio, silicio, potasio, calcio, bario y hierro.

Para el agrupamiento de los datos se utilizó un mapa auto-organizativo (SOM, de su nombre en inglés *Self-Organizing Map*) [12, 15]. Para el agrupamiento se entrenaron mapas con distintas cantidades de neuronas, es decir, clusters. Todos los mapas utilizados tienen una topología rectangular en forma de grilla y la cantidad de iteraciones de entrenamiento fue fijada en 100 épocas. La inicialización del mapa fue determínistica, utilizando PCA [17] sobre los datos. Para considerar solapamiento en las neuronas del mapa se tomaron las neuronas vecinas inmediatas, es decir la de la izquierda, derecha, arriba y abajo, de cada neurona. Así, al considerar vecindad $Vn = 0$ cada neurona es un único cluster y $Vn = 1$ cada neurona y sus cuatro vecinas forman un mismo cluster. Por ello, varios patrones pueden llegar a pertenecer a más de un cluster y de esta forma pertenecer a clusters solapados.

En el Cuadro 2 se observa una primer columna en donde se indican las cantidades de clusters considerados para las soluciones C y C'. Luego el cuadro se divide en dos grandes columnas que representan los valores de los índices analizados, FM y oFM. Cada una de estas columnas se vuelven a dividir para mostrar los resultados de las pruebas con soluciones C' que poseen y no poseen solapamientos. La cantidad de clusters de las soluciones fueron de 4, 25 y 100. Para la solución tomada como referencia C no se consideró solapamiento. En el caso de Iris, cuando se toma $k_1 = 4$ FM disminuye su valor no sólo al tomar solapamiento sino también al pasar de $k_2 = 25$ a $k_2 = 100$. Por otro lado oFM aumenta notablemente al considerar solapamiento, pero también disminuye al aumentar la cantidad de clusters en la solución C'. Se observa como FM vuelve a decaer al considerar solapamiento cuando $k_1 = 25$ y $k_2 = 100$. Se puede

[1] `http://archive.ics.uci.edu/ml/datasets/Iris`
[2] `http://archive.ics.uci.edu/ml/datasets/Wine`
[3] `http://archive.ics.uci.edu/ml/datasets/Yeast`
[4] `http://archive.ics.uci.edu/ml/datasets/Glass+Identification`

visualizar además cómo oFM mantiene la tendencia de aumentar al considerar solapamiento pero, al igual que FM, aumenta al pasar de $k_1 = 4$ a $k_1 = 25$ cuando $k_2 = 100$. Para la base de datos Wine se percibe, de la misma forma, que en todas las pruebas el valor del índice FM experimenta un descenso de su valor cuando se pasa de no considerar solapamiento a considerarlo. Se observa además cómo el valor de oFM empieza a subir cuando se toma en cuenta el solapamiento. Al observar los valores de Yeast, se percibe cómo baja el índice de FM cuando se toma $Vn = 1$ en vez de $Vn = 0$ en C' y cómo, opuestamente, el índice propuesto comienza lentamente a crecer cuando se pasa a considerar vecindad en las soluciones C'. Por el último, al ver la evolución de los valores para Glass, se puede apreciar, nuevamente, cómo los valores que arroja para FM decrecen en todos los casos que se aplicó solapamiento, y una tendencia opuesta en el índice oFM.

En todos los experimentos, considerando los distintos conjuntos de datos, el valor de FM decae cuando aumenta el solapamiento para la solución C', de forma totalmente opuesta a lo que sucede con los valores de oFM. Se ve además que siempre los dos índices mostraron mayor valor, es decir mayor semejanza de las soluciones comparadas, a medida que la cantidad de clusters de C se asemejaba a la de C'. En el caso de FM este valor disminuía cuando se tenía en cuenta solapamiento, mientras que se mostraba un incremento notorio en el valor que arrojaba el índice que se propone. Esto es debido a que a mayor cantidad de clusters hay mayor dispersión de los patrones en los mismos y por ello menor es el valor de los índices calculados. Esto último es consistente con lo que se observa en todas las comparaciones dónde $k_1 = 4$ y $k_2 = 100$. En dicho caso los valores no disminuyen o aumentan prácticamente, sea para FM o para oFM respectivamente, puesto que existe mucha dispersión de los patrones en el mapa C'. De esta forma, el efecto de considerar vecindad hace que el solapamiento sea más pequeño en comparación con los casos en dónde los mapas no tienen tanta diferencia en cuanto a tamaño. Es por ello que tanto FM apenas baja y oFM incrementa poco en la consideración de solapamiento con soluciones con tanta diferencia en cantidad de clusters.

5. Conclusiones y trabajos futuros

En este trabajo se propuso un nuevo índice para análisis de estabilidad entre soluciones solapadas, parcial o completamente. Se realizó un análisis del índice FM para mostrar sus falencias al momento de aplicarlo sobre soluciones solapadas y a partir de éste se derivó uno nuevo. El índice propuesto refleja correctamente la similitud entre soluciones que pueden o no poseer clusters solapados. Esto se alcanzó teniendo especial cuidado en la forma de contar casos extremos en los que se pudiesen agrupar los patrones, permitiendo así una normalización más cuidadosa. Luego, en los experimentos se mostró como el índice FM es incapaz de tratar con soluciones solapadas, mientras que oFM puede sortear dicho problema.

Con respecto a trabajos futuros, se espera realizar mayor cantidad de pruebas de los índices utilizando otros algoritmos y posiblemente nuevos conjuntos de datos. También se desea incorporar pruebas de significancia estadística a los resultados para poder arribar a conclusiones más sólidas. También sería deseable a futuro la incorporación al análisis de nuevos tipos de índices y derivar así

nuevas métricas que se puedan comparar con las de este trabajo.

Referencias

[1] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo. A comparison of extrinsic clustering evaluation metrics based on formal constraints. 12:461–486. doi:10.1007/s10791-008-9066-8.

[2] K. Bache and M. Lichman. UCI machine learning repository, 2013.

[3] A Bayá. *Aplicación de algoritmos no supervisados a datos biológicos*. PhD thesis, Universidad Nacional de Rosario, Marzo 2011.

[4] S. Ben-David, U.V. Luxburg, and D. Pál. A sober look at clustering stability. *In COLT, Springer*, pages 5–19, 2006. doi:10.1007/11776420_4.

[5] A. Ben-Hur, A. Elisseeff, and I. Guyon. A stability based method for discovering structure in clustered data. pages 6–17, 2002.

[6] A. Ben-Hur and I. Guyon. Detecting stable clusters using principal component analysis. pages 159–182, 2003. doi:10.1385/1-59259-364-X:159.

[7] R.A. Fisher. The use of multiple measurements in taxonomic problems. pages 179–188, 1936.

[8] E. Fowlkes and C. Mallows. A method for comparing two hierarchical clusterings. 78:553–569, 1983.

[9] Mark K. Goldberg, Mykola Hayvanovych, and Malik Magdon-Ismail. Measuring similarity between sets of overlapping clusters. pages 303–308, 2010. doi:10.1109/SocialCom.2010.50.

[10] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. pages 107–145, 2001.

[11] J. Handl, J. Knowles, and D.B. Kell. Computational cluster validation in post-genomic data analysis. pages 3201–3212, 2005. doi:10.1093/bioinformatics/bti517.

[12] T. Kohonen. Self-organized formation of topologically correct feature maps. 43:59–69, 1982.

[13] M. Meilă. Comparing clusterings—an information based distance. pages 873–895, May 2007. doi:10.1016/j.jmva.2006.11.013.

[14] M. Meilă and D. Heckerman. An experimental comparison of model-based clustering methods. pages 9–29, January 2001.

[15] D. Milone, G. Stegmayer, L. Kamenetzky, M. Lopez, J. Giovannoni, J.M. Lee, and F. Carrari. *omesom: a software for integration, clustering and visualization of transcriptional and metabolite data mined from interspecific crosses of crop plants. pages 438–448, 2010. doi:10.1186/1471-2105-11-438.

[16] X. V. Nguyen, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. pages 2837–2854, 2010.

[17] K. Pearson. On lines and planes of closest fit to systems of points in space. pages 559–572, 1901.

[18] Ohad Shamir and Naftali Tishby. Model selection and stability in k-means clustering. pages 213–243, 2010. doi:10.1007/s10994-010-5177-8.

[19] David Skillicorn. *Understanding complex datasets. Data mining with matrix decompositions*. Ed. Chapman & Hall / CRC, 2007.

[20] J. Wu, H. Yuan, H. Xiong, and G. Chen. Validation of overlapping clustering: A random clustering perspective. 180:4354–4369, November 2010. doi:10.1016/j.ins.2010.07.028.

[21] R. Xu. and D. Wunsch. *Clustering*. IEEE Press Series on Computational Intelligence. Ed. Wiley, 2009.

Anexo B

Apéndice C

A New Index for Cluster Validation with Overlapped Clusters

Campo, D. N., Stegmayer, G., y Milone, D. H., "A New Index for Cluster Validation with Overlapped Clusters", *ELSEVIER Expert Systems With Applications*, doi: 10.1016/j.eswa.2016.08.021, enviado en Febrero de 2016, revisado en Julio de 2017, en prensa en Agosto de 2017.
En este trabajo se desarrollaró un nuevo índice para la medición de grupos cuando los mismos pueden estar solapados. Los resultados muestran que el índice propuesto midió con valores esperados y expresó mejor, sobre todo en soluciones solapadas, que índices clásicos y ampliamente utilizados.

A new index for clustering validation with overlapped clusters

†David N. Campo[1,2], ⋆Georgina Stegmayer[1,2] y ‡Diego H. Milone[2]

[1] CIDISI – Universidad Tecnológica Nacional – Facultad Regional Santa Fe

[2] sinc(i) – Universidad Nacional del Litoral – Facultad de Ingeniería y Ciencias Hídricas

†dncampo@santafe-conicet.gov.ar,⋆gstegmayer@santafe-conicet.gov.ar,‡d.milone@ieee.org

Abstract External validation indexes allow similarities between two clustering solutions to be quantified. With classical external indexes, it is possible to quantify how similar two disjoint clustering solutions are, where each object can only belong to a single cluster. However, in practical applications, it is common for an object to have more than one label, thereby belonging to overlapped clusters; for example, subjects that belong to multiple communities in social networks. In this study, we propose a new index based on an intuitive probabilistic approach that is applicable to overlapped clusters. Given that recently there has been a remarkable increase in the analysis of data with naturally overlapped clusters, this new index allows to comparing clustering algorithms correctly. After presenting the new index, experiments with artificial and real datasets are shown and analyzed. Results over a real social network are also presented and discussed. The results indicate that the new index can correctly measure the similarity between two partitions of the dataset when there are different levels of overlap in the analyzed clusters.
keywords: overlapped clusters, validation index, external validation, cluster perturbation.

1 Introduction

Clustering algorithms take a dataset as input and, through a non–supervised process, partition the data into a set of clusters or groups. A cluster can be defined as a group of objects that are similar given a relative measure and that are dissimilar to objects grouped in others clusters [22, 29]. The application of clustering algorithms always returns a solution, even when there is not a clear structure in the data. Therefore, a reliable mechanism for measuring similarities between partitions is desirable to detect which ones are, for example, more stable when several solutions are considered.

Furthermore, in current practice, most information that is created through social networks, news tags, collaboration networks and other Internet media, is naturally overlapped. As a result, overlapped solutions are expected to be found in the analysis of such type of data. An index for measuring similarities between these partitions would therefore be a valuable tool to study them. Recently, with the spread of social and collaboration networks, the use of clustering with overlapping properties has increased and new algorithms have been proposed [25, 1, 9, 10, 6, 28, 2].

Two types of validation measures can be used for measuring similarities between clustering solutions: internal and external. The first type of metrics measures attributes taken from the data itself and the clusters formed, such as data compactness and separability. The second one makes a comparison between clustering solutions, taking one as a reference and comparing it with other groupings [11, 12]. Considering external metrics only, three types of measures are available: pair counting measures, set matching measures, and information theory measures. One of the most used and widely known pair counting measure is the Fowlkes-Mallows index (FM), which works with the frequency of pairs of patterns found in two clustering solutions that are being compared [3, 8]. A representative set matching measure is the Maximum Match [21], which analyses the most similar clusters from both solutions and counts the elements in common in such paired groups. Finally, regarding measures based on information theory,

Normalized Mutual Information is extensively used and works by quantifying the information shared between both solutions, through the concept of entropy [20, 24]. However, none of these metrics was designed for evaluating similarities between solutions when overlapped clusters are considered.

Lately, given the overwhelming amount of information created through different social and collaboration networks, interest has emerged in clustering analysis to process such amount of data when there are overlapping clusters [2]. For example, in [31], the authors proposed an ant-based algorithm to detect communities in networks, where clusters are formed by nodes that may be considered as overlapped. In [18], the authors developed a method for characterizing the structure of real–world affiliation networks composed of groups of fully connected, generally overlapped communities. Also, in [1] overlapped clusters in social networks are studied and a framework based on a game theory approach is proposed for detecting communities. Similarly, in [9], a new method based on a Bayesian model is presented. This method enables the detection of large overlapped communities in massive synthetic datasets and in large–scale, real life social, biological and citation networks. In [14], the authors present an R package that extends an existing algorithm for clustering, which handles directed and weighted links between nodes in a biological network. These networks would naturally contain nested or overlapped links. In [19], data acquired from protein networks is clustered and the results are integrated with chemical databases using ontologies. Hence, based on the principle of guilt–by–association [26, 16, 23], the author studies new types of cellular functions. The objects are grouped together in either overlapped or non–overlapped clusters. With an index that enables the evaluation of overlapped clusters, the author would be able to evaluate the resulting groups, according to the study of diseases related to diverse cellular functions. However, although there are plenty of external non-overlapped indexes [4, 27] and a significant amount of research in overlapping clustering, there is a lack of external validation indexes for assessing and comparing overlapped solutions. Finally, taken into account indexes for overlapped clusters, [5] presented a preliminary study about stability analyses in the context of overlapped clusters and developed an initial index to assess overlapped solutions. However, it failed to show the expected values in some basic cases.

In this study a novel index is presented based on an intuitive probabilistic approach. The new index works with the probability of finding any pair of objects in each solution and in both solutions simultaneously. The behavior of the proposed index is shown in the presence of overlapped and disjoint clusters, when two clustering solutions for a same dataset are analyzed. Comparisons with classical external indexes such as FM, Jaccard (JAC) and Adjusted Rand Index (ARI) are performed on artificial and real datasets. Also, a real–life case from YouTube is presented, in which classical indexes fail because they show false differences when clusters become more overlapped.

The remainder of this paper is organized as follows. Section 2 presents a detailed analysis of the new index and an explanation of its factors. Section 3 describes the experiments performed with artificial and real datasets, and with a particular social network, and discusses the results obtained. Finally, conclusions are drawn and future research is suggested in Section 4.

2 A probabilistic approach for designing the new index

In this section, we introduce a new index for evaluating overlapped clustering solutions. First, notation and basic definitions are outlined. Next, a probabilistic approach for analyzing and designing the proposed index is presented, as well as an application example. The new index is also compared with some classical indexes to show its advantages when measuring overlapped solutions.

Given a set $S = \{\mathbf{s}_1, ..., \mathbf{s}_N\}$, which is comprised of N objects, a clustering algorithm partitions them into a collection of subsets $C = \{c_1, ..., c_k\}$ called clusters. The union of clusters in such partition forms a covering of the original set of objects: $\cup_{i=1}^{k} c_i = S$. Similarly, another algorithm or an equivalent one with different parameters over the same dataset could generate an alternative partition of k' clusters: $C' = \{c'_1, ..., c'_{k'}\}$. Since each individual object could be grouped into more than one cluster, it is important to note that the number of elements in the clusters could be greater than or equal to N. For example, two clustering solutions are depicted in Figure 1: C with $k = 1$ and C' with $k' = 2$. In Figure 1.a), the solution is composed of a single cluster, c_1, that groups all of the objects together. In Fig. 1.b) there are two clusters. Cluster c'_1 groups all objects and c'_2 groups all but one. In this scenario, both clusters of C' share $N - 1$ objects and are said to be *overlapped*. Given that every pair of objects that

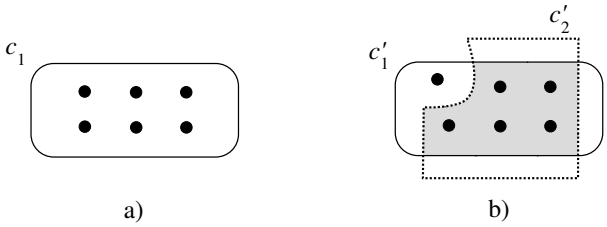


Figure 1: This illustrative example depicts two solutions: a) C with $k = 1$, and b) C' with overlapping and $k' = 2$. The shaded area includes the common elements between the overlapped clusters.

exists in one solution could be found in the other one, and vice versa, a similarity value close to 1 should be expected if an external index is applied. However, the values obtained by classical indexes such as FM, ARI and JAC are 0.692, 1.238 and 0.478, respectively.

To overcome the evident misbehavior of classical indexes when overlapped clusters are present, a new index is proposed considering the probability that any pair of objects could be found in a given solution or in both solutions. Consequently, consider the cluster c_i of a given solution. Assuming that all of the objects have the same chance of being grouped into any cluster, this probability can be estimated as

$$Pr((\mathbf{s}_x, \mathbf{s}_y) \in c_i) = \frac{\binom{|c_i|}{2}}{\binom{N}{2}} = \frac{|c_i|\,(|c_i| - 1)}{N\,(N-1)}, \tag{1}$$

where $|c_i|$ is the number of elements in cluster c_i. The numerator represents the number of pairs that can be found with $|c_i|$ elements. In order to normalize it, the denominator represents a similar situation where all of the objects are grouped together in a single cluster; hence any possible pair could be found.

Taking into account the previous analysis, consider the solution C, where

$$\tilde{p} = \frac{\sum_{i=1}^{k} \binom{|c_i|}{2}}{k\binom{N}{2}} \tag{2}$$

estimates the probability of finding a pair of elements in any cluster c_i for all of the existing clusters k. The numerator accumulates all of the pairs found in each cluster. The denominator represents a normalization factor, which acts as if all of the objects were grouped together. The k factor considers the situation where the overlapping is complete up to all k clusters. An identical reasoning could be applied to obtain a comparable expression for C',

$$\tilde{p}' = \frac{\sum_{j=1}^{k'} \binom{|c'_j|}{2}}{k'\binom{N}{2}}. \tag{3}$$

The same analysis described for $\tilde{p}$ and $\tilde{p}'$ can be performed for both solutions together. Therefore,

$$Pr((\mathbf{s}_x, \mathbf{s}_y) \in c_i \wedge (\mathbf{s}_x, \mathbf{s}_y) \in c'_j) = \frac{\binom{|c_i \cap c'_j|}{2}}{\binom{N}{2}} = \frac{|c_i \cap c'_j|\,(|c_i \cap c'_j| - 1)}{N\,(N-1)}, \tag{4}$$

can be seen as an approximation to the probability that the pair of data points ($\mathbf{s}_x$, $\mathbf{s}_y$) is present in both solutions. In this equation, $|c_i \cap c'_j|$ represents the number of elements in common in clusters c_i and c'_j. The whole expression stands for the event of drawing two objects that are in both clusters c_i and c'_j.

Now suppose that the same analysis is made for every possible pairing between clusters of C and C'. The probability of finding ($\mathbf{s}_x$,$\mathbf{s}_y$) in both solutions can be estimated as

$$\tilde{t} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{k'}\binom{|c_i \cap c'_j|}{2}}{\binom{N}{2}\frac{\max(n,n')}{N}\min(k,k')}, \tag{5}$$

where n and n' represent the number of objects that can be counted in solutions C and C', respectively, considering every overlap. For example, in Figure 1.a), $n = 6$, and in Figure 1.b), $n' = 11$. Similarly to $\tilde{p}$ and $\tilde{p}'$, the numerator of (5) counts all of the effective pairs of objects that can be found in both solutions simultaneously. The denominator acts once again as a normalization term. It basically covers the extreme scenario where all of the objects are clustered together several times. Just as in (2) and (3), $\binom{N}{2}$ counts the number of pairs that can be arranged given all N objects. Since there could be overlaps in both solutions, the given number of pairs should be multiplied by a factor. On the one hand, there could be as many overlaps as k in C and k' in C'. On the other hand, it was found that the matching between clusters of both solutions produces at most $\min(k, k')$ pairs of clusters in the comparison. Finally, $\max(n, n')/N$ is the average number of objects that can be found considering overlaps.

With these elements in mind, the new index for overlapped clusters ($\mathcal{OC}$) could be defined as the ratio between the probability of finding two items grouped together in both solutions and the maximum probability of finding them in one of the given solutions. That is,

$$\mathcal{OC} = \frac{\tilde{t}}{\max(\tilde{p}, \tilde{p}')}. \tag{6}$$

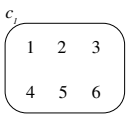
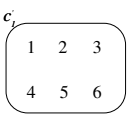
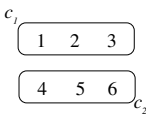
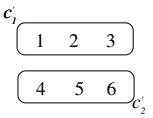
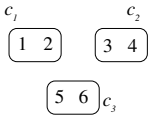
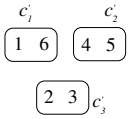
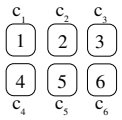
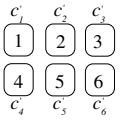
For the example in Figure 1, the new index produces the following values. When (2) is applied to the solution in Figure 1.a), $\tilde{p} = 1$ is obtained. Then, using (3) in Figure 1.b), $\tilde{p}' = 0.833$, and using (5), $\tilde{t} = 0.909$. Finally, when (6) is employed the new index is $\mathcal{OC} = 0.909/\max(1, 0.833) = 0.909$. The same experiment was performed using 1000 objects and the values obtained for FM, ARI and JAC were 0.707, 1.200 and 0.500, respectively, whereas $\mathcal{OC} = 0.999$. These examples show that the proposed index obtains an intuitively expected similarity between similar solutions with overlapped clusters, given that the probability of finding two objects grouped together in any of them tends effectively to 1.

The FM index tries to reflect the similarity of the two evaluated solutions considering the probability of randomly finding a pair of objects together, for each or both solutions at the same time. The problem is that it does not consider the existence of a pair of objects more than once, when the objects are overlapped in several clusters. With respect to ARI, the behavior with overlapped clusters is inconsistent. It fails to narrow the index score below 1. This behavior is observed because ARI is a corrected–for–chance version of the Rand Index, in which an expected value is substracted in both the numerator and denominator. In practical applications, when overlapped clusters are present, such adjustment could produce values either below 0 or above 1. As is the case with the FM index, the Jaccard index, is the result of the ratio between a count of objects found in both solutions over the objects found in any of both solutions. As a result, the index does neither consider the overlapping situation nor counts the occurrences of repeated pairs of objects. This is why it is expected to fail in an overlapping scenario. Finally, the proposed index was carefully designed considering the overlap in the solutions and non–overlapping situations, which is an aspect that has not been considered in the design of the other indexes.

3 Experimental results and discussion

In this section, we present the results of the application of FM, ARI, JAC and the new $\mathcal{OC}$ index. First, a set of artificial examples of extreme[1] situations is given. Next, tests are shown in which the overlap is gradually introduced. These examples show the behavior of the new index compared with standard measures in trivial cases. Then, the application of the index in several real datasets is described. Finally, the application of the $\mathcal{OC}$ index for the analysis of social network data is presented.

Table 1: Index results for extreme artificial examples

	Solutions		Indexes			
	C	C'	**FM**	**ARI**	**JAC**	$\mathcal{OC}$
I			1.000	—	1.000	1.000
II			1.000	1.000	1.000	1.000
III			0.000	-0.250	0.000	0.000
IV			0.000	—	—	0.000

3.1 Performance with artificial datasets

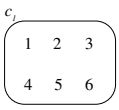
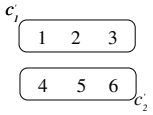
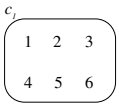
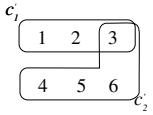
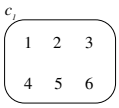
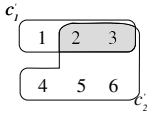
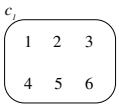
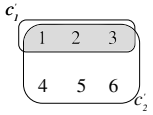
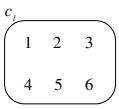
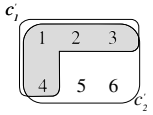
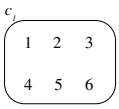
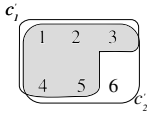
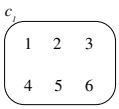
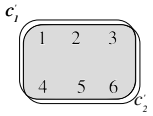
The first set of tests was performed over artificial clustering situations, where some extreme cases are analyzed. Also, examples with a gradual degree of overlap are given. The three tables in this subsection present basic tests that can help to better understand the behavior of the indexes under different types of overlap. In Table 1, the first column enumerates the examples given. Columns 2 and 3 depict the solutions that are compared. Finally, columns $4-7$ show the values for FM, ARI, JAC and $\mathcal{OC}$, respectively. All of the examples in Table 1 have six data points that were clustered through one to six clusters.

Examples I and II show a pair of identical solutions with different configurations. In Example I, there is only one cluster in each solution, and every pair of objects that can be found in one cluster can also be found in the other one. In Example II, there are two clusters in each solution, and every pair of objects found in one cluster can also be found in a cluster from the other solution. In all of these cases, a value of 1.00 is expected, since the complete equivalence of both solutions is evident. In fact, all of the indexes can detect such situation, except for ARI, which produces no value at all in Example I. This is because the expected and maximum values of the denominator in the definition of ARI [13] are equal and a division by zero is returned. The last two examples in Table 1 present situations where no similarity between both solutions exists. In Example III, none of the possible pairs of objects found in solution C can be found in solution C'. In Example IV, both solutions cannot form any pair of objects at all since each data point is in a different cluster. In this case, a value of 0.00 is expected for each example because no pairs of data could be found in the first and second solutions simultaneosly. Once again, the only index that disagrees with this intuition is ARI in Example III. In this case, the numerator of the index is defined as a difference between an observed and an expected value. Therefore, a negative score is computed when the observed value is lower than the expected one. In addition, in Example IV, ARI and JAC present a division by zero since no pairs are formed at all, and the indexes cannot provide a value. The results show clearly that the proposed index can measure basic situations without overlap.

The examples in Table 2 represent several scenarios with gradual overlap. The table has the same

[1] In the sense of expected 0/1 values for indexes.

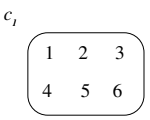
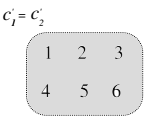
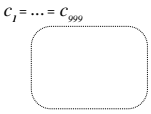
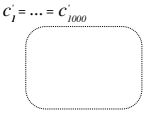
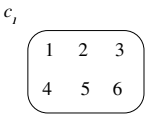
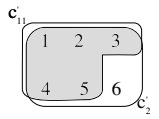
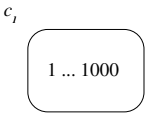
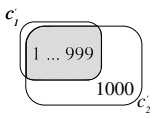
Table 2: Index results for some gradually overlapped artificial examples

	Solutions		Indexes			
	C	C'	**FM**	**ARI**	**JAC**	$\mathcal{OC}$
I			0.632	0.000	0.400	0.400
II			0.665	-2.186	0.442	0.514
III			0.695	2.347	0.483	0.650
IV			0.721	1.444	0.520	0.800
V			0.700	1.324	0.489	0.840
VI			0.692	1.238	0.478	0.909
VII			0.692	1.179	0.478	1.000

columns as in the previous case and all of the examples given in it have six data points clustered. Solutions labeled with C are always identical (the reference solution) and all of the objects are always grouped into a single cluster c_1. Solutions named C' have two clusters that are incrementally overlapped between them, ranging from no overlap in Example I to a full double overlap in Example VII. For instance, in Example II the element 3 appears in both clusters of C', but a repeated pair is not generated. In this case, three new pairs arise from the interaction between object 3 and each object from cluster c'_2, and the proposed index can identify such situation. By contrast, in Examples III to VII the increasing overlapping effect allows repeated pairs to be produced. Such pairs are formed by objects that belong to both clusters. For example, the pair formed by objects 2 and 3 can be found in both clusters c'_1 and c'_2. Thus, given the slowly increasing overlap in examples I to IV, all of the indexes but ARI show a corresponding incremental behavior.

Hence, since there is more overlap, more pairs of objects from solution C can be found in C', with some extra repeated pairs due to the overlap itself. It is expected that, while one solution increases its overlapped clusters, the index tends to raise its value as there are more matchings between solutions. In Example IV and successive ones, all of the possible pairs of objects among six data points can be

Table 3: Index results for some extremely overlapped artificial examples

	Solutions		Indexes			
	C	C'	**FM**	**ARI**	**JAC**	$\mathcal{OC}$
I			0.692	1.179	0.478	1.000
II			0.001	1.000	0.001	1.000
III			0.692	1.238	0.478	0.909
IV			0.707	1.200	0.500	0.999

actually found in both solutions, in addition to some repeated ones in C'. Classical indexes (FM and JAC) increase up to Example V, where they begin to decrease. However, it is expected that the indexes continue to rise given the increasing overlap. The ARI once again shows disagreeing values in these examples. As explained earlier for Table 1, ARI might produce negative values in certain cases. By contrast, as the overlap progresses and duplicated pairs of points are found more frequently, the $\mathcal{OC}$ index follows this behavior with a monotonic increasing value through all of the last examples, achieving the top 1.00 score when a perfect overlap of both clusters exists.

Finally, Table 3 presents more extreme situations. The structure of this table is the same as the previous ones. In the first example, solution C is composed of a single cluster with all of the objects contained in it. By contrast, solution C' has two overlapped clusters with all of the objects grouped together. The proposed index shows a complete equivalence between both solutions since any pair of objects can be found in them. None of the other indexes present this similarity. In the second example, something similar occurs but with more overlapped clusters. Solution C has 999 complete overlaps and solution C' has 1000. Once again, $\mathcal{OC}$ and this time ARI present a complete equivalence, while others decrease to almost zero. These cases demonstrate that the proposed index does not change as the overlap increases with a high number. The index maintains a value of 1.00 under any number of complete overlaps, which is the expected behavior since the pairs of objects are maintained through the overlaps. Other indexes fail to show this and their values decrease with higher overlaps.

Example III is exactly the same as Example VI from Table 2. In solution C, a complete overlap is observed among all of the objects, while solution C' contains one cluster with a complete overlap and one cluster that groups all of the objects but one. The proposed index shows a value relatively close to 1.00. This is because almost all of the pairs formed in solution C can be found twice in solution C', but a few others can be found only once. This is consistent with the fact that not all of the pairs have the same proportion of appearance in the second solution, and the $\mathcal{OC}$ index can reflect this irregular situation.

Finally, Example IV takes the previous example to the limit, where solution C has a thousand objects grouped all together in one cluster. Solution C' has cluster c'_1, which groups all of the objects, and c'_2, which groups all but one. In the last two examples (III and IV), solution C has one cluster with all of

the elements grouped together. By contrast, solution C' contains two clusters: in the first one, all of the elements are grouped together, but in the second one all of the elements but one are grouped. The difference between Examples III and IV is that the number of elements considered in the latter tends to be high. As demonstrated by these two experiments, the $\mathcal{OC}$ index accurately reflects the fact that all pairs of data can be found in both solutions, obtaining a velue close to 1.00, as expected. Other indexes fail when an almost complete overlap is presented. Our proposed index tends to obtain a maximum score when the number of elements is relatively high.

3.2 Benchmarking with real datasets

Four well–known databases[2] were used for performing the experiments on real datasets: Iris, Wine, Yeast and Glass [17]. The Iris dataset has four attributes and 150 patterns distributed in three classes of 50 patterns each [7]. Only one of the three classes is linearly separable from the others, which have many patterns that are very close in the attribute space. The Wine dataset represents the measure and analysis of 13 chemical attributes of an Italian wine taken from different vineyards. This dataset of 178 patterns is distributed in three groups: A, B and C, with 59, 71 and 48 patterns each, respectively. The Yeast dataset is based on a study of yeast and it is intended to determine the location of its proteins in the cell. It has 1484 patterns distributed in 10 groups with 463, 429, 244, 163, 51, 44, 37, 30, 20 and 5 elements each, and 8 attributes have been measured. Finally, the Glass dataset has 9 attributes and 214 patterns distributed in 7 groups. These datasets are freely available for general purpose use, and they are widely used in the academic community. They were selected for their small size and adequacy for the detailed analysis of the proposed measure.

[2] `http://archive.ics.uci.edu/ml/datasets/`

Table 4: Results for FM, ARI, JAC and $\mathcal{OC}$ indexes using Iris, Wine, Yeast and Glass databases. The reference solutions C have 4 or 25 clusters and zero overlap ($V_n = 0$). Solutions C' have 25 and 100 clusters, taking $V_n = 0$ and $V_n = 1$

	clusters in C and C'	**FM**		**ARI**		**JAC**		$\mathcal{OC}$	
		$Vn = 0$	$Vn = 1$	$Vn = 0$	$Vn = 1$	$Vn = 0$	$Vn = 1$	$Vn = 0$	$Vn = 1$
Iris	$k = 4$ vs $k' = 25$	0.33	0.30	0.16	4.26	0.14	0.10	0.14	0.38
	$k = 4$ vs $k' = 100$	0.17	0.16	0.03	-0.66	0.03	0.03	0.03	0.11
	$k = 25$ vs $k' = 100$	0.33	0.23	0.23	-0.34	0.14	0.09	0.14	0.37
Wine	$k = 4$ vs $k' = 25$	0.40	0.32	0.23	9.33	0.17	0.12	0.17	0.48
	$k = 4$ vs $k' = 100$	0.19	0.18	0.06	-0.52	0.04	0.04	0.04	0.14
	$k = 25$ vs $k' = 100$	0.34	0.23	0.26	-0.24	0.15	0.10	0.17	0.39
Yeast	$k = 4$ vs $k' = 25$	0.32	0.23	0.15	6.61	0.13	0.08	0.13	0.35
	$k = 4$ vs $k' = 100$	0.16	0.14	0.04	-0.58	0.03	0.03	0.03	0.11
	$k = 25$ vs $k' = 100$	0.29	0.18	0.21	-0.30	0.13	0.07	0.14	0.33
Glass	$k = 4$ vs $k' = 25$	0.33	0.27	0.10	3.77	0.11	0.08	0.11	0.30
	$k = 4$ vs $k' = 100$	0.15	0.14	0.02	-0.75	0.02	0.02	0.02	0.09
	$k = 25$ vs $k' = 100$	0.38	0.24	0.28	-0.36	0.17	0.09	0.18	0.43

A self–organizing map (SOM) [15] was used for clustering the data. Given that several neurons in a region of the map may be considered as a single group, incrementally overlapped clusters can be easily analyzed with different levels of neighborhood between neurons. To process the datasets, each map was trained with different numbers of neurons (clusters). All of the experiments were performed with a rectangular topology, grid shape, principal component analysis initialization and training iterations set to 100 epochs. To consider overlap between clusters, the topological closeness between neurons in the map has been taken into account with a Von Neumann neighborhood. When it is equal to zero ($V_n = 0$), each neuron represents a single cluster. When $V_n = 1$ is considered, each neuron and its four adjacent neighboring neurons (north, south, east and west) are considered as part of the same cluster. Thus, when $Vn = 1$ is used, each neuron and its neighbors may overlap between them, and some patterns are associated with more than one cluster. This is how overlapped clusters are formed in a SOM.

Table 3.2 presents the results obtained over the real datasets. The table is divided as follows: column 1 contains the name of each dataset, column 2 shows the number of clusters considered for C and C', and columns $3-6$ present the values obtained for each experiment and for each index. The last four columns are divided into values with and without overlap ($Vn = 0$ and $Vn = 1$) in the solution C'. Six different experiments were performed for each dataset: $k = 4$ vs. $k' = 25$, $k = 4$ vs. $k' = 100$ and $k = 25$ vs. $k' = 100$. This is with $V_n = 0$ and $V_n = 1$ for solution C'. In all of the cases, no overlap was considered for the reference solution C. For the Iris dataset, a decrease is observed in the FM index, not only when $k = 4$ and k' change from 25 to 100, but also when overlap ($V_n = 1$) is considered. The ARI produces a value over 1.00 when overlap is considered in the experiment $k = 4$ vs. $k' = 25$. The opposite behavior is observed when other sizes of clusters are considered and overlap is taken into account, showing values below 0. The JAC index exhibits a similar behavior to the FM index: it decreases when overlap is considered and when more clusters are taken into consideration in solution C'. Finally, the $\mathcal{OC}$ index decreases when a higher number of clusters is considered in C', but increases when overlapped clusters are analyzed. This is an expected behavior since it is consistent with the fact that, when overlapped clusters are used, the probability of finding more matching pairs of points between solutions is higher.

The analyses for Wine, Yeast and Glass datasets are very similar to the previous one. In these experiments, the values of FM and JAC indexes decrease when a higher number of clusters is considered. This is also the case when overlap is taken into account. The ARI shows exactly the same behavior as in the previous dataset. With respect to the $\mathcal{OC}$ index, a remarkable increase is observed when overlapped clusters are analyzed. However, it decreases when solution C' has more clusters.

Figure 2 presents the results of the experiment where the reference solution C have $k = 4$ and $V_n = 0$, and the C' solutions has $k = 100$ and either $V_n = 0$ or $V_n = 1$. For all of the datasets, when the overlap increases, classical indexes show a notable decrement, while the proposed index shows an increment. This behavior is consistent with the intuition that, given the existence of overlapped clusters, it should be more likely to find a pair of objects in common in both solutions.

In these experiments, an increment is observed in all of the indexes but ARI when the number of clusters of C is close to the number of clusters of C'. This is due to the data dispersion in C': when there are more clusters, the data patterns are spread through more neurons, thereby reducing the value of the index. This is observed when the experiment $k = 25$ vs. $k' = 100$ is analyzed. In the case of FM and JAC, the scores also decrease when overlap is considered, while the proposed index always shows an increment for $V_n = 1$ with respect to $V_n = 0$. This is because the classical indexes do not handle overlapped clusters properly, whereas $\mathcal{OC}$ does. With this in mind, should be noted that when there are overlapped clusters, both FM and JAC indexes do not count the matchings between groups appropriately. This explains why FM and JAC barely decrease, and $\mathcal{OC}$ hardly increases with overlap.

In summary, with artificial or real datasets, the proposed index is effective for assessing clustering solutions in which there can be overlapped clusters. Moreover, $\mathcal{OC}$ shows reliable and confident results with extreme overlapping cases, thus enabling a better understanding and comparison of the outcome of clustering algorithms.

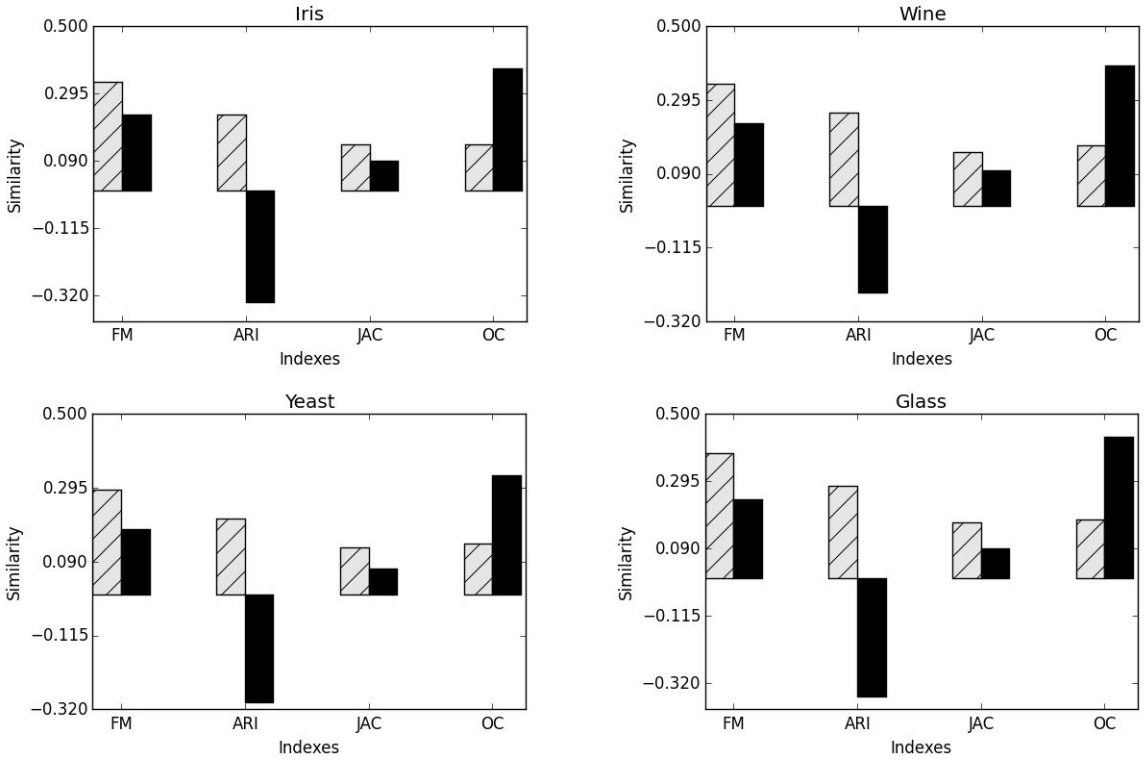


Figure 2: Bar plots of FM, ARI, JAC and $\mathcal{OC}$ indexes for Iris, Wine, Yeast and Glass databases. The reference solutions C have 25 clusters and zero overlap ($V_n = 0$). Solutions C' have 100 clusters with $V_n = 0$ (gray diagonal striped bars) and $V_n = 1$ (black bars).

3.3 Social networking application

This subsection describes the experiments performed over a real dataset taken from a social network. The results of the application of the proposed measure over the YouTube dataset are analyzed and discussed [30]. YouYube is a video–sharing social network. Users can create groups or communities to share their videos, and other users can join them. This dataset captures the relation of a group of users of the social network through communities. The dataset is comprised of communities that are defined as groups of two or more users who share similar interests. Each community, which is considered as a cluster of users, is described in the dataset as a list of user IDs. One user may belong to one or more communities. When a user belongs to several communities, those communities are said to be overlapped. The level of overlap of a community depends on how many of its participants also belong to other communities. The resulting dataset, after preprocessing and removing communities with less than 10 users, contains 37038 users and 2087 communities. Communities with less than 10 users showed an almost non–existent overlapping behavior, which would affect the focus and interpretability of indexes when overlap is tested.

The experiment performed over this dataset involved sorting the groups C_j by their degree of overlap. Solution C'_j was taken from solution C_j with different levels of perturbation. Random modifications were applied and users were added to random communities. The original dataset was then divided into several subsets. Since the communities are arranged by levels of overlap, the first subset of 35 communities has zero overlap. Each of the following subsets considered in this study represents a different level of increasing overlap. Communities are grouped into subsets with a similar level of overlap until no more communities are available. Each subset has at least three times the number of communities as the first one (with zero overlap) in order to ensure that all of the solutions have a minimum number of elements

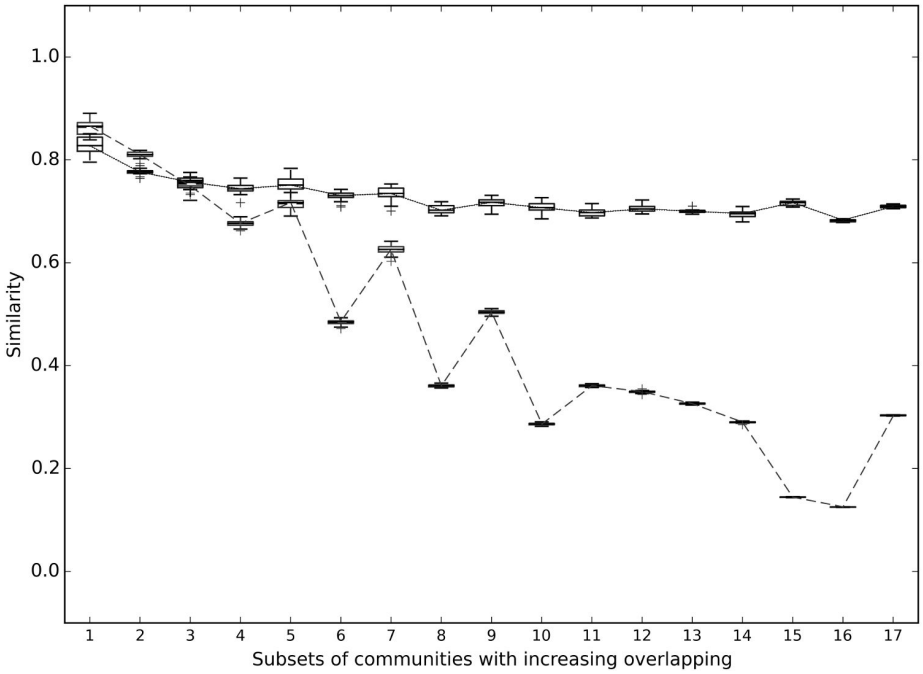


Figure 3: Boxplot of FM and $\mathcal{OC}$ indexes for the social network (YouTube) dataset. The dashed line corresponds to the FM index and the continuous one, to the $\mathcal{OC}$ index.

for calculating the indexes. The last subset includes the communities with a higher level of overlap. Therefore, the original dataset was divided into 17 disjoint subsets ranging from zero to the maximum overlap.

Figure 3 shows boxplots for the FM and $\mathcal{OC}$ indexes for a 10% random perturbation of users within communities. Each box corresponds to the median of 20 runs over the corresponding subset of the dataset. The abscissa axis shows the increasing level of overlap of each subset from zero (subset labeled 1) to the maximum overlap (subset labeled 17). Communities with no overlap reach similar values, for both indexes, between 0.8 and 0.9. As the overlap increases, a remarkable decrease is observed in the FM index, reaching values barely upon 0.1 when the overlap is very high. The values obtained by the $\mathcal{OC}$ index are more stable when the overlap increases, thereby demonstrating the capability of $\mathcal{OC}$ to be immune to the overlap. Moreover, with a higher overlap in the subsets, the FM curve falls in a fluctuating manner. By contrast, the $\mathcal{OC}$ curve shows a smooth behavior, maintaining high values. Therefore, we conclude that the proposed index is effective for measuring similarities in real scenarios where there are overlapped clusters. Furthermore, the $\mathcal{OC}$ index exhibits a more stable behavior than classical measures such as FM, irrespective of the presence of overlap between subsets.

4 Conclusions and future work

In this study, we proposed a new index ($\mathcal{OC}$) for comparing solutions that may have a certain degree of overlap. The proposed index was designed from an intuitive probabilistic approach and was then compared with classical approaches, such as Fowlkes-Mallows, Adjusted Rand and Jaccard indexes. For simple artificial examples, these indexes showed unexpected behaviors, while a more reliable situation was observed with the $\mathcal{OC}$ index. Experiments performed with benchmark datasets and real data from a social network confirmed these findings. On the one hand, classical indexes tended to show fewer similarities between solutions as the overlap increased. On the other hand, the proposed index was immune to the overlap and performed accurately, showing the level of similarity between clustering solutions. It should be noted that the $\mathcal{OC}$ index also performed well when there was no overlap. Thus, the proposed index can be applied to any type of solution, regardless of the presence of overlapped clusters.

In future research, we will perform experiments using the proposed index in order to analyze the stability of clustering solutions with any degree of overlap.

5 Funding and acknowledgments

This study was supported by the National Scientific and Technical Research Council (CONICET) [PIP 2013-2015 117], Universidad Nacional del Litoral (UNL) [CAI+D 2011 548] and National Agency of Science and Technology Promotion (ANPCyT) [PICT 2014 2627].

6 References

References

[1] Hamidreza Alvari, Sattar Hashemi, and Ali Hamzeh. Discovering overlapping communities in social networks: A novel game-theoretic approach. *AI Communications*, 26(2):161–177, April 2013.

[2] Alessia Amelio and Clara Pizzuti. Overlapping Community Discovery Methods: A Survey. In Şule Gündüz-Öğüdücü and A. Şima Etaner-Uyar, editors, *Social Networks: Analysis and Case Studies*, Lecture Notes in Social Networks, pages 105–125. Springer Vienna, 2014.

[3] Asa Ben-Hur and Isabelle Guyon. Detecting Stable Clusters Using Principal Component Analysis. In MichaelJ. Brownstein and ArkadyB. Khodursky, editors, *Functional Genomics*, number 224 in Methods in Molecular Biology, pages 159–182. Humana Press, January 2003.

[4] Marcel Brun, Chao Sima, Jianping Hua, James Lowey, Brent Carroll, Edward Suh, and Edward R. Dougherty. Model-based evaluation of clustering validation measures. *Pattern Recognition*, 40(3):807–824, March 2007.

[5] David Campo, Georgina Stegmayer, and Diego Milone. Stability analysis in overlapped clusters. *Iberoamerican Journal of Artificial Intelligence*, 17(53):79–89, 2014.

[6] Tanmoy Chakraborty. Leveraging disjoint communities for detecting overlapping community structure. *J. Stat. Mech.*, 2015(5):P05017, May 2015.

[7] R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2):179–188, September 1936.

[8] E. B. Fowlkes and C. L. Mallows. A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, 78(383):553–569, September 1983.

[9] Prem K. Gopalan and David M. Blei. Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences*, 110(36):14534–14539, September 2013.

[10] Tatiana Gossen, Michael Kotzyba, and Andreas Nürnberger. Graph clusterings with overlaps: Adapted quality indices and a generation model. *Neurocomputing*, 123:13–22, January 2014.

[11] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On Clustering Validation Techniques. *J. Intell. Inf. Syst.*, 17(2-3):107–145, December 2001.

[12] Julia Handl, Joshua Knowles, and Douglas B. Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3212, August 2005.

[13] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[14] Alex T. Kalinka and Pavel Tomancak. linkcomm: an R package for the generation, visualization, and analysis of link communities in networks of arbitrary size and type. *Bioinformatics (Oxford, England)*, 27(14):2011–2012, July 2011.

[15] Teuvo Kohonen. The self-organizing map. *Neurocomputing*, 21(1–3):1–6, November 1998.

[16] V. Lacroix, L. Cottret, P. Thébault, and M. F. Sagot. An Introduction to Metabolic Networks and Their Structural Analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5(4):594–617, October 2008.

[17] M. Lichman. UCI machine learning repository, 2013.

[18] Dajie Liu, Norbert Blenn, and Piet Van Mieghem. Characterising and modelling social networks with overlapping communities. *International Journal of Web Based Communities*, 9(3):371–391, 2013.

[19] Ken McGarry. Discovery of functional protein groups by clustering community links and integration of ontological knowledge. *Expert Systems with Applications*, 40(13):5101–5112, October 2013.

[20] Marina Meilă. Comparing clusterings – an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, May 2007.

[21] Marina Meilă and David Heckerman. An Experimental Comparison of Model-Based Clustering Methods. *Machine Learning*, 42(1-2):9–29, January 2001.

[22] David Skillicorn. *Understanding Complex Datasets: Data Mining with Matrix Decompositions*. CRC Press, May 2007.

[23] Björn Usadel, Takeshi Obayashi, Marek Mutwil, Federico M. Giorgi, George W. Bassel, Mimi Tanimoto, Amanda Chow, Dirk Steinhauser, Staffan Persson, and Nicholas J. Provart. Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant, Cell & Environment*, 32(12):1633–1651, December 2009.

[24] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *J. Mach. Learn. Res.*, 11:2837–2854, December 2010.

[25] Zhu Wang, Daqing Zhang, Xingshe Zhou, Dingqi Yang, Zhiyong Yu, and Zhiwen Yu. Discovering and Profiling Overlapping Communities in Location-Based Social Networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 44(4):499–509, April 2014.

[26] Cecily J. Wolfe, Isaac S. Kohane, and Atul J. Butte. Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC Bioinformatics*, 6:227, 2005.

[27] Junjie Wu, Jian Chen, Hui Xiong, and Ming Xie. External validation measures for K-means clustering: A data distribution perspective. *Expert Systems with Applications*, 36(3, Part 2):6050–6061, April 2009.

[28] Jierui Xie, Stephen Kelley, and Boleslaw K. Szymanski. Overlapping Community Detection in Networks: The State-of-the-art and Comparative Study. *ACM Comput. Surv.*, 45(4):43:1–43:35, August 2013.

[29] Rui Xu and Don Wunsch. *Clustering*. John Wiley & Sons, November 2008.

[30] Jaewon Yang and Jure Leskovec. Defining and evaluating network communities based on ground-truth. *Knowl Inf Syst*, 42(1):181–213, October 2013.

[31] Xu Zhou, Yanheng Liu, Jindong Zhang, Tuming Liu, and Di Zhang. An ant colony based algorithm for overlapping community detection in complex networks. *Physica A: Statistical Mechanics and its Applications*, 427:289–301, June 2015.