APRENDIZAJE POR TRANSFERENCIA Y ENSAMBLES PARA CLASIFICACIÓN EN BIOINFORMÁTICA

Escudero, Sofía

Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, sinc(i), FICH-UNL, CONICET
Director/a: Stegmayer, Georgina
Codirector/a: Milone, Diego

Área: Ingeniería

Palabras claves: Inteligencia artificial, ensamble de modelos, clasificación de proteínas.

INTRODUCCIÓN

En bioinformática, la anotación computacional automática de proteínas es aún un desafío sin resolver, ya que la generación de datos ocurre a un ritmo mucho más rápido de lo que los expertos pueden revisar y anotar manualmente (Bateman et al, 2023). Por ejemplo, a la fecha (julio de 2024) hay más de 245 millones de entradas en la base de datos pública UniProt[1], de las cuales en menos del 1% se conoce su verdadera función. Esta brecha entre las capacidades actuales de generación de datos en bioinformática y su correspondiente anotación funcional puede limitar el avance en la aplicación de proteínas para, por ejemplo, la producción de vacunas y la cura de enfermedades.

La anotación automática de una proteína implica su clasificación en una de las familias o dominios funcionales depositados en la base de datos Pfam. Para esto se generan perfiles de modelos ocultos de Markov (HMM) a partir de un conjunto reducido de secuencias "semilla de familia", alineadas y curadas a mano. Luego, estos perfiles HMM son utilizados para clasificar familias directamente por máxima probabilidad. Sin embargo, un 25% de las proteínas de Pfam permanecen sin clasificación, ya que no pueden ser detectadas por perfiles HMM con suficiente probabilidad o no tienen ejemplos suficientes, alineados y curados manualmente, para construir un perfil adecuado (Mistry et al, 2021). Como alternativa a los HMM, se destacan los modelos basados en Deep Learning (DL), que son capaces de inferir patrones ocultos y compartidos por las secuencias que pertenecen a una familia (Bileschi et al, 2022). Sin embargo, es sabido que las técnicas de DL necesitan gran cantidad de datos para aprender, lo que representa una fuerte limitación en el caso de estas secuencias biológicas ya que de muchas familias sólo se posee un muy pequeño número de ejemplos. Una solución a este problema es el aprendizaje por transferencia (TL, por sus siglas en inglés), que ofrece representaciones genéricas aprendidas previamente por grandes modelos de lenguaje, y que, luego, pueden adaptarse para un problema específico. En el grupo de bioinformática del sinc(i) se ha probado TL junto con modelos de DL para la clasificación de proteínas, obteniendo mejoras con respecto al estado de arte (Vitale et al, 2023).

Título del proyecto: Estimación de distancias semánticas aprendizaje profundo para la predicción de nuevas funciones de genes
Instrumento: CAID 115
Año convocatoria: 2020
Organismo financiador: UNL
Director/a: Georgina Stegmayer

[1] https://www.uniprot.org/

Tomando estos resultados como punto de partida, en esta investigación se diseña una propuesta superadora basada en ensambles de modelos de DL. Los ensambles son una clase de estrategias de aprendizaje automático donde, en vez de construir un único modelo, se combinan múltiples modelos base para lograr un mejor desempeño en la generalización. En el ámbito de la bioinformática, los ensambles muestran gran potencial al poder tratar con tamaños de muestra pequeños, alta dimensionalidad, distribución desequilibrada de clases y datos ruidosos y heterogéneos generados por sistemas biológicos (Cao et al, 2020).

Adicionalmente, los ensambles propuestos utilizarán representaciones basadas en transferencia de aprendizaje, de modelos que se "congelan" pasando a ser la etapa inicial de otra arquitectura que es entrenada para la nueva tarea más específica. Cada nuevo modelo se entrena con aprendizaje supervisado en un pequeño conjunto de datos etiquetados para clasificación de secuencias en familias de proteínas. Para la parte supervisada se propone realizar distintos tipos de ensambles de varios modelos supervisados, con el objetivo de reducir el error de predicción y mejorar la clasificación de proteínas en familias, en relación a los métodos estándar y los modelos individuales.

OBJETIVOS

- Analizar comparativamente los métodos actuales basados en aprendizaje autosupervisado para aprender representaciones de secuencias de interés biológico.
- Investigar el potencial del ensamblado de modelos en la tarea de clasificación de secuencias de proteínas en dominios funcionales.
- Aplicar los modelos de predicción sobre datos reales y de libre acceso.

METODOLOGÍA

Para la primera etapa de este trabajo se utilizó un sub-conjunto de datos obtenido de la base de datos Pfam versión 32, que incluye 73.132 proteínas que deben ser clasificadas en una de 393 familias. Se utilizaron 55.029 de esas proteínas para el entrenamiento de los modelos base, 8.958 para la validación durante el entrenamiento y 9.145 para prueba, tanto de los modelos base como del ensamble final de los mismos.

La arquitectura general de los ensambles propuestos consiste en pre-entrenar varios modelos base de tipo DL y extraer sus capas iniciales para luego usarlas congeladas. Posteriormente, se lleva a cabo un entrenamiento individual para cada modelo que parte de una inicialización aleatoria de sus parámetros libres (pesos) (ver Figura 1). Además, cada miembro del ensamble puede tener diferentes hiperparámetros (p.e. número de entradas o tasa de aprendizaje). Para una proteína dada de prueba, cada modelo obtiene una puntuación para cada familia posible, y el ensamble determina la familia final de esa proteína como aquella con el mayor puntaje acumulado.

Se exploraron distintas formas de ensamblar las predicciones de los modelos base:

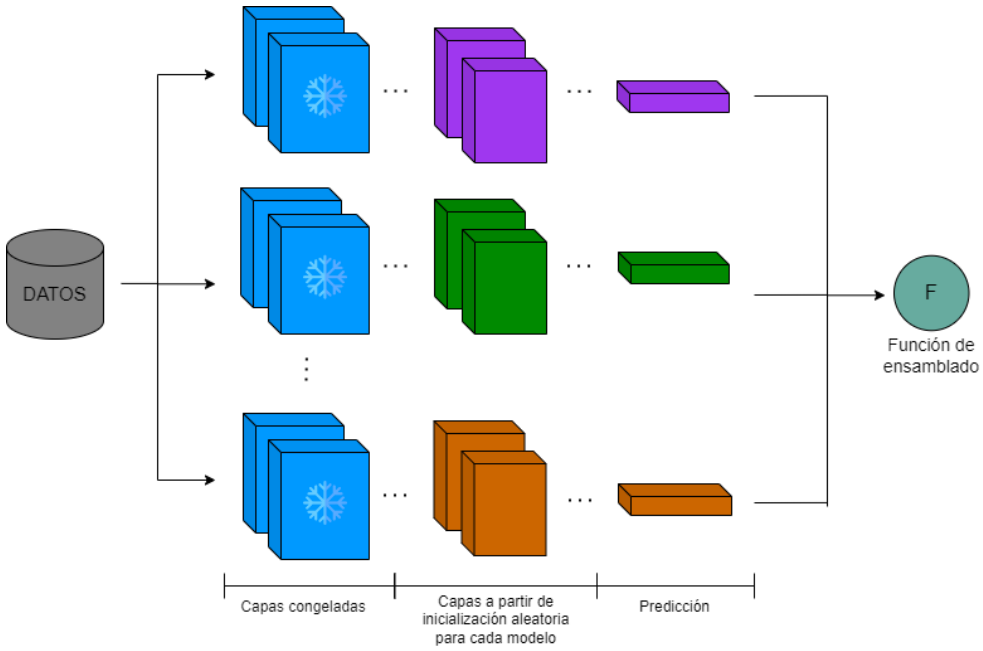

Figura 1. Arquitectura general propuesta para el ensamble de los modelos.

Votación: para cada proteína se obtienen las predicciones finales de los modelos base, donde cada predicción de familia es un voto a la misma. La predicción final del ensamble para esa proteína es la familia con mayor cantidad de votos y, en caso de empate, se elige la familia con el mayor puntaje.

Promedio: para cada proteína se suman los puntajes obtenidos por los modelos base para cada familia y luego se divide por la cantidad de modelos comprendidos en el ensamble. La proteína pertenece a la familia con el mayor puntaje resultante de la contribución de todos los modelos.

Promedio pesado con aprendizaje: en este método se asignan pesos a cada uno de los modelos base, para ser aplicados en la sumatoria de puntajes. En un principio, estos pesos se inicializan al azar y, luego, se entrenan para minimizar el error del ensamble de salida en un conjunto de datos de validación. Durante cada iteración del entrenamiento, se calculan las predicciones de los modelos base para todas las proteínas, se realiza el promedio utilizando los pesos actuales y se calcula el error del ensamble. Finalmente, se actualizan los pesos empleando el gradiente del error del ensamble con respecto a los mismos.

RESULTADOS Y CONCLUSIONES

Se entrenaron diez modelos de clasificación, comenzando con una configuración común y variando posteriormente los hiperparámetros para aumentar la diversidad entre ellos (ver Tabla 1). Luego se realizó un ensamble incremental para analizar la contribución de cada modelo en términos de su tasa de error y características específicas.

Las curvas de error muestran una clara tendencia decreciente para los tres métodos de ensamblado propuestos (ver Figura 2). Esto indica que, en general, se obtienen mejores resultados al aumentar el número de modelos ensamblados. Esto es válido incluso al añadir algunos modelos con un error individual mayor, como es el caso del Modelo 4 en comparación

a los Modelos 1, 2 y 3. Cabe destacar que el error final alcanzado por un ensamble por promedio pesado con aprendizaje de 10 modelos fue de 3,44%, en lugar de, por ejemplo, un 4,59% de error del Modelo 1.

Tabla 1. Configuración y tasa de error de los modelos base ensamblados.

Modelo	Ventana de entrada	Ventana de salida	Tasa de aprendizaje	Tasa de error
1	32	32	1e-6	4,59%
2				4,25%
3				4,68%
4				4,89%
5			1e-5	4,35%
6			1e-4	4,14%
7	64	64	1e-6	4,32%
8				4,34%
9	128	128		4,60%
10				4,05%

Entre los tres métodos de ensamblado propuestos, se determinó que, en general, el promedio pesado con aprendizaje es el de mejor desempeño, ya que presenta la curva más baja y registra el menor error (3,40% con ocho modelos ensamblados).

Como trabajo futuro se propone explorar de forma separada los efectos de la variación de las ventanas y la tasa de aprendizaje, así como investigar y proponer distintas arquitecturas de ensambles.

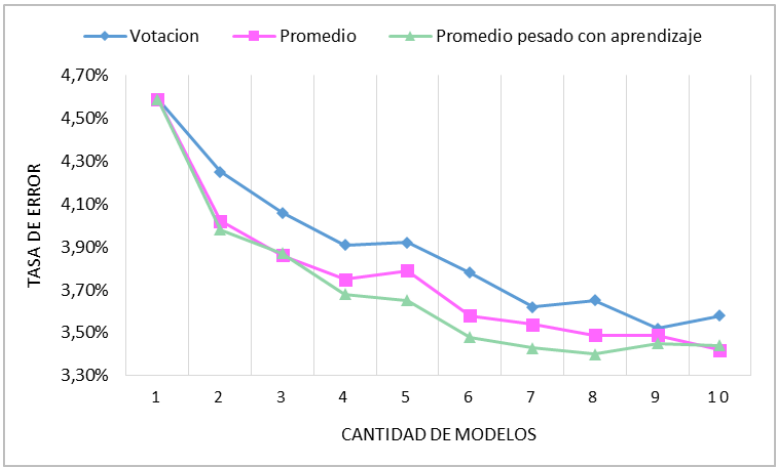

Figura 2. Tasa de error según cantidad de modelos ensamblados.

BIBLIOGRAFÍA BÁSICA

Bateman, A et al., 2023. Uniprot: the universal protein knowledgebase in 2023. Nucleic Acids Research, 51(D1), D523-D531.
Bileschi, M. L. et al., 2022. Using deep learning to annotate the protein universe. Nat Biotechnol. 40, 932–937.
Cao, Y. et al., 2020. Ensemble deep learning in bioinformatics. Nat Mach Intell 2, 500-508.
Mistry, J. et al., 2021. Pfam: The protein families database in 2021. Nucleic Acids Research. 49, D412–D419.
Vitale, R. et al., 2024. Evaluating large language models for annotating proteins. Brief in Bioinformatics, 25, 3, bbae177.