Información de la Linea de Investigación y Desarrollo:

Estimación de distancias semánticas y aprendizaje profundo para la predicción de nuevas funciones de genes.

Detalle:

La ciencia de datos ha experimentado un crecimiento exponencial en la última década. Cada día es más fácil adquirir y almacenar datos de todo tipo. Pero los desafíos ahora tienen que ver con la extracción de información útil de esos datos. La inteligencia artificial está proveyendo soluciones efectivas a gran cantidad de problemas de este tipo, especialmente desde el aprendizaje de máquina, que ha demostrado tener todo el potencial necesario para los desafíos actuales. En particular, el área de bioinformática presenta problemas en ciencia de datos cada vez más desafiantes. Por ejemplo, la predicción automática de la función de genes a partir de genomas completos y de mediciones experimentales de diferente naturaleza. Actualmente existen anotaciones semánticas con vocabulario controlado que describen a los genes en cualquier organismo en base a términos de la ontología de genes (GO). La curaduría (manual) de anotaciones para nuevos genes es un procedimiento muy costoso que requiere de conocimiento específico de parte del experto del dominio. Las herramientas computacionales basadas en aprendizaje de máquina pueden ayudar a encontrar rápidamente potenciales anotaciones para genes nuevos, e impulsar el descubrimiento de nuevo conocimiento en este dominio. En este proyecto se proponen nuevos modelos y algoritmos para predecir anotaciones de genes cuya potencial función es desconocida, es decir sin términos GO asociados, mediante el desarrollo de métodos novedosos de aprendizaje de máquina. En primer lugar se propone desarrollar un nuevo método a partir de factorización conjunta de matrices no negativas de distancias de expresión y distancias semánticas entre genes conocidos. Una vez realizada esta factorización, se propone utilizarla para reconstruir la información faltante en la matriz de distancia semántica a genes desconocidos. Una segunda etapa utilizará esta información semántica reconstruida para entrenar modelos probabilísticos y modelos de aprendizaje profundo que permitan predecir el conjunto de etiquetas GO que describen la función de cada gen desconocido.
Proyecto financiado CAI+D 2020.

Estado:

Financiado

Facultad:

Facultad de Ingeniería y Ciencias Hidricas