Propuesta de Proyecto

Seminario de Analítica y Ciencia de Datos

Presentación del Proyecto

La correcta vectorización de los datos de entrada para un modelo de clasificación puede significar en gran medida el éxito o fracaso de la creación e implementación del mismo. dado que para análisis de cadenas largas de texto es necesario la creación de un vocabulario y a su vez de un encoder que convierte las cadenas de texto a un input que el modelo pueda utilizar, la elección de un método errado de vectorización nos puede dejar con un vocabulario demasiado grande o un vocabulario muy pequeño que carecería de contexto. en múltiples procedimientos computacionales, el ADN suele ser representado como una cadena de caracteres consistiendo en las bases A, T, G y C, dichas cadenas para poder ser procesadas por algoritmos de ML tienen que pasar por un proceso de vectorización, y la implementación de métodos usados en aplicaciones de análisis de texto, podrían mejorar notablemente el desempeño de algoritmos de clasificación

Objetivo del Proyecto

[Descripción de alto nivel del objetivo general del proyecto y su propósito]

Evaluar los efectos del uso de una representación fisicoquímica de las secuencias de ADN sobre el rendimiento de los algoritmos de machine learning.

Contexto del Problema

[Definición detallada y clara del problema y sus aspectos más relevantes]

Los algoritmos de aprendizaje de máquina, mejor conocidos como Machine Learning, están siendo ampliamente usados en la actualidad para responder preguntas biológicas, por ejemplo para predecir la estructura tridimensional de las proteínas, para inferir su posible función, para tareas de clasificación taxonómica de secuencias de ADN, para predecir sitios de splicing o para la predicción de la afinidad de unión de factores de transcripción entre muchas otras aplicaciones.

Usar representaciones más informativas que la secuencia de ADN o aminoácidos en el caso de las proteínas ha probado ser un enfoque efectivo. Por ejemplo, el algoritmo de alineamiento de secuencias MAFFT transforma las secuencias en números que describen la polaridad, tamaño y otras propiedades de los monómeros que componen la secuencia.

Propuesta de Solución y Alcance

Identificación de la mejor representación digital para cadenas de ADN que permita el mejor rendimiento para algoritmos de clasificación

Definición del Alcance

Se pretende postular y evaluar diferentes métodos de encoding para cadenas de ADN y determinar bajo el mismo algoritmo de ML cual nos presenta los mejores resultados.

No se pretende evaluar diferentes arquitecturas o modelos de ML, solo se pretende evaluar los diferentes métodos de encoding

Riesgos e Impacto del Negocio

[Descripción breve de cómo la solución impacta positivamente el problema y a su vez los riesgos potenciales de su implementación. Use referencias de casos éxito que soporte sus argumentos]

Escoger el método de encoding adecuado puede representar un avance significativo en el desarrollo de futuros algoritmos de clasificación, dando una base sólida para investigar cuál sería la arquitectura más adecuada y con mejor rendimiento.

Por ejemplo, el algoritmo de alineamiento de secuencias MAFFT transforma las secuencias en números que describen la polaridad, tamaño y otras propiedades de los monómeros que componen la secuencia, mejorando considerablemente tiempos de computación [1].

[1] K. Katoh, K. Misawa, K.-i. Kuma, and T. Miyata, “MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform,” Nucleic Acids Research, vol. 30, pp. 3059–3066, July 2002.