Aplicación de Deep Symbolic Learning en NGS

Ángel Canal-Alonso[1], Pedro Jiménez[1] and Noelia Egido[1], Javier Prieto[1], Juan Manuel Corchado[1]

[1] Departamento de Bioinformática y Biología Computacional, AIR Institute, Carbajosa de la Sagrada, España

E-mail: acanal@air-institute.com

Resumen

La aplicación de Deep Symbolic Learning en el análisis genómico ha comenzado a ganar tracción como un enfoque prometedor para interpretar y comprender vastos conjuntos de datos derivados de la secuenciación del ADN. Las técnicas de secuenciación de nueva generación (NGS) han revolucionado el campo de la genética clínica y la biología humana, generando volúmenes masivos de datos que requieren herramientas avanzadas para su análisis. Sin embargo, los métodos tradicionales a menudo resultan demasiado abstractos o complicados para el personal clínico. Este trabajo se centra en explorar cómo el Deep Symbolic Learning, un subcampo de la inteligencia artificial explicable (XAI), puede ser aplicado efectivamente a los datos de NGS. Se llevará a cabo una evaluación detallada de la adecuación de diferentes arquitecturas, con el objetivo final de ofrecer recomendaciones para su implementación en flujos de trabajo de secuenciación clínica.

Palabras Clave: Next-Generation sequencing, Explainable Artificial Intelligence, Deep Symbolic Learning

Introducción

Deep Learning

El Deep Learning, también conocido como aprendizaje profundo, representa una subcategoría de técnicas dentro del aprendizaje automático que ha revolucionado múltiples campos de estudio y aplicación, desde la visión por computadora hasta el procesamiento del lenguaje natural, pasando por supuesto por la bioinformática y genómica computacional. Específicamente, esta rama del aprendizaje automático se basa en arquitecturas de redes neuronales artificiales con numerosas capas ocultas, que son denominadas "profundas". Estas capas permiten al modelo aprender jerarquías de características a partir de los datos, desde las más básicas hasta las más complejas.

El fundamento del Deep Learning es la simulación de estructuras neuronales análogas a las presentes en el cerebro humano, aunque de manera considerablemente más simplificada. Cada neurona en estas redes está conectada a otras y puede transmitir información entre ellas. A medida que los datos pasan por cada capa, se lleva a cabo una transformación no lineal que permite la detección y construcción gradual de características. Por ejemplo, en el contexto de la visión por computadora, las primeras capas podrían identificar bordes, mientras que las capas más profundas podrían identificar estructuras complejas como caras o patrones específicos.

Un aspecto clave que ha propulsado el éxito del Deep Learning es la capacidad de estas redes neuronales profundas para realizar el aprendizaje automático de características. A diferencia de los enfoques tradicionales, donde las características se extraen manualmente, en el aprendizaje profundo la red es capaz de aprender de manera autónoma las características más relevantes directamente de los datos brutos.

La eficiencia y precisión de estas técnicas, sin embargo, vienen con el costo de la necesidad de grandes volúmenes de datos y una capacidad computacional significativa. La retroalimentación y ajuste constante de los pesos de las conexiones en estas redes durante el entrenamiento requiere de potentes unidades de procesamiento, siendo las Unidades de Procesamiento Gráfico (GPU) las más empleadas debido a su capacidad para manejar operaciones matriciales en paralelo, esenciales en el entrenamiento de redes neuronales.

En el contexto de la genómica y la secuenciación de nueva generación (NGS), el Deep Learning ha mostrado potencial para mejorar la precisión y velocidad en tareas como la anotación genómica, la predicción de estructura de proteínas

y, de relevancia para nuestro estudio, en la fase de Variant Calling, donde se busca identificar variantes genéticas a partir de datos de secuenciación. La capacidad del Deep Learning para manejar grandes conjuntos de datos y aprender características complejas lo convierte en una herramienta valiosa para abordar los desafíos inherentes a la genómica de alta resolución.

Symbolic Learning

El aprendizaje simbólico, o Symbolic Learning, es un enfoque clásico en el campo de la inteligencia artificial (IA) que se centra en la representación y manipulación del conocimiento en forma de símbolos y reglas. En lugar de depender exclusivamente de cálculos numéricos o estadísticos, como en otros enfoques de aprendizaje automático, el aprendizaje simbólico se basa en la construcción de representaciones simbólicas de la información, permitiendo razonamientos lógicos y deducciones basadas en estas representaciones.

La esencia del aprendizaje simbólico reside en su capacidad para modelar relaciones complejas y estructuradas en los datos. Estos modelos son generalmente interpretables, ya que están formados por conjuntos de reglas, hechos o estructuras lógicas, que pueden ser fácilmente comprendidos y examinados por los humanos. Por ejemplo, un sistema basado en aprendizaje simbólico podría expresar el conocimiento en forma de reglas del tipo "Si A, entonces B", permitiendo razonar sobre estas reglas y llegar a conclusiones específicas.

El aprendizaje simbólico ha sido fundamental en la evolución de la IA, especialmente en los años iniciales de la disciplina, y ha dado lugar a sistemas expertos, motores de inferencia, y bases de conocimiento. Estos sistemas son especialmente eficaces en dominios donde el conocimiento previo es esencial y puede ser claramente definido y estructurado, como en la medicina, el derecho o la ingeniería.

En el ámbito de la genómica y la bioinformática, el aprendizaje simbólico ofrece una perspectiva única al proporcionar herramientas para representar y manipular el conocimiento biológico de forma estructurada. Las relaciones genéticas, las vías metabólicas o las interacciones proteína-proteína, por ejemplo, pueden ser codificadas en estructuras simbólicas que faciliten su análisis y comprensión.

Cuando se combina con enfoques como el Deep Learning, el aprendizaje simbólico permite a los sistemas beneficiarse tanto de la capacidad de generalización y aprendizaje automático de características de las redes neuronales, como de la precisión, transparencia e interpretabilidad del razonamiento simbólico. Esta combinación, denominada "Deep Symbolic Learning", integra lo mejor de ambos mundos, y en el contexto del Variant Calling en pipelines NGS, puede ofrecer soluciones robustas y altamente interpretables para la identificación y clasificación de variantes genéticas.

Deep Symbolic Learning

El Deep Symbolic Learning (DSL) es una aproximación emergente en el campo de la inteligencia artificial que busca combinar las fortalezas del Deep Learning y del aprendizaje simbólico. Esta integración se propone resolver una de las principales críticas al Deep Learning, que es la falta de interpretabilidad y transparencia de sus modelos, ofreciendo soluciones que no solo sean potentes en términos de rendimiento, sino también comprensibles y justificables.

El Deep Symbolic Learning opera a través de la conjunción de redes neuronales, capaces de aprender representaciones ricas y jerárquicas de los datos, con estructuras simbólicas que permiten la construcción de modelos lógicos y semánticamente coherentes. En lugar de considerar únicamente patrones numéricos o estadísticos, como lo hace el Deep Learning puro, el DSL incorpora símbolos, reglas y relaciones lógicas en su proceso de aprendizaje, brindando un mayor contexto y estructura al conocimiento adquirido.

Un aspecto destacado de DSL es su capacidad para aprovechar el conocimiento previamente establecido, codificado en representaciones simbólicas, para informar y guiar el aprendizaje de las redes neuronales. Esto es especialmente útil en dominios donde existe una rica base de conocimientos, como es el caso de la biología y genómica.

En el contexto del Variant Calling en pipelines NGS, el DSL ofrece un enfoque prometedor. Las redes neuronales pueden aprender patrones complejos y sutilezas en los datos de secuenciación, mientras que el componente simbólico puede incorporar reglas y hechos conocidos sobre variantes genéticas, mutaciones y su relevancia biológica. Esto no solo podría mejorar la precisión de la identificación de variantes, sino también proporcionar explicaciones lógicas y basadas en el conocimiento sobre por qué se considera una determinada secuencia como una variante.

La naturaleza híbrida del Deep Symbolic Learning también permite una mayor flexibilidad en el modelado. Mientras que las redes neuronales pueden ajustarse a las peculiaridades y ruidos de los datos, el componente simbólico puede actuar como un regulador, garantizando que las predicciones y conclusiones sean coherentes con el conocimiento biológico establecido. De esta forma, el DSL se posiciona como una herramienta robusta y a la vanguardia para enfrentar los desafíos intrínsecos de la fase de Variant Calling y de la genómica en general.

Aplicación del DSL en NGS

La secuenciación de nueva generación (NGS) ha revolucionado el campo de la genómica, permitiendo la obtención de grandes volúmenes de datos genómicos en tiempos y costos significativamente reducidos en comparación con las técnicas tradicionales. Estos avances, si bien proveen una riqueza de información sin precedentes,

también presentan desafíos significativos en cuanto a procesamiento, análisis e interpretación de los datos. Es en este contexto donde el Deep Symbolic Learning (DSL) emerge como una solución potencial para abordar y superar dichos desafíos.

Ventajas de su aplicación

La unión del aprendizaje simbólico y el aprendizaje profundo en el contexto de la secuenciación de nueva generación (NGS) ofrece un abanico de ventajas que capitalizan las fortalezas de ambos enfoques:

- Interpretabilidad y Transparencia: Uno de los principales desafíos del Deep Learning es su naturaleza "caja negra", lo que significa que, aunque el modelo pueda tener un alto rendimiento, puede ser difícil entender cómo llega a una decisión particular. Al integrar el aprendizaje simbólico, se introduce una capa de transparencia y explicabilidad al modelo. Las decisiones basadas en reglas simbólicas pueden ser inspeccionadas, rastreadas y justificadas, facilitando la comprensión y validación de los resultados en el contexto de la NGS.
- Incorporación de Conocimiento Previo: En genómica, hay un vasto cuerpo de conocimiento acumulado sobre genética, mutaciones y relaciones genómicas. El aprendizaje simbólico permite la incorporación explícita de este conocimiento en forma de reglas y relaciones. Esto no solo informa y guía al modelo, sino que también puede aumentar la precisión y robustez, al asegurarse de que el sistema no contradiga principios genómicos bien establecidos.
- Generalización y Adaptabilidad: Mientras que el Deep Learning es excelente para detectar y aprender patrones en grandes conjuntos de datos, el aprendizaje simbólico le otorga al sistema la capacidad de generalizar a partir de ejemplos específicos y de adaptarse a nuevos datos o contextos. Esto es esencial en NGS, donde los datos pueden variar según la técnica de secuenciación, el organismo estudiado o las condiciones experimentales.
- Robustez ante el Ruido: Los datos de NGS pueden ser ruidosos debido a errores en la secuenciación o variaciones biológicas. Mientras que las redes neuronales profundas pueden ser susceptibles a sobreajustar a este ruido, la naturaleza estructurada y lógica del aprendizaje simbólico puede actuar como un moderador, evitando conclusiones precipitadas basadas en información ruidosa o atípica.
- Optimización Computacional: La integración de conocimiento simbólico puede dirigir y enfocar el proceso de aprendizaje, reduciendo potencialmente la necesidad de iteraciones computacionalmente costosas. Al tener una estructura guía basada en reglas y relaciones conocidas, el sistema puede converger más rápidamente a soluciones óptimas, ahorrando tiempo y recursos computacionales.

Integración de Multi-Modalidad: En genómica, a menudo se combinan diferentes tipos de datos, como secuencias genómicas, expresión génica y datos de proteómica. Mientras que el aprendizaje profundo puede manejar eficientemente la integración de múltiples modalidades de datos, el aprendizaje simbólico puede proporcionar un marco coherente y estructurado para comprender y razonar sobre cómo estos diferentes tipos de datos se relacionan entre sí.

Aplicabilidad en las fases de un pipeline

La interpretación de secuencias genéticas es una tarea crítica en genómica y bioinformática, ya que implica la identificación y comprensión de variantes y mutaciones que pueden tener implicaciones clínicas, evolutivas o funcionales. El Deep Symbolic Learning (DSL), al combinar el aprendizaje profundo y el aprendizaje simbólico, tiene un potencial significativo para mejorar y enriquecer esta interpretación. Aquí se exploran algunas aplicaciones concretas del DSL en el ámbito de la interpretación genómica:

- Identificación de Variantes: Una de las principales tareas en la interpretación genómica es identificar variantes, como SNPs y mutaciones estructurales, a partir de secuencias de NGS. El DSL puede ser de particular utilidad aquí, ya que las redes neuronales pueden identificar patrones complejos en los datos, mientras que el componente simbólico puede validar estas identificaciones contra reglas y conocimientos previamente establecidos. Esta combinación puede reducir significativamente los falsos positivos y negativos.
- Análisis Funcional: No todas las variantes identificadas tienen una repercusión funcional. El DSL puede ayudar a predecir el impacto de una variante, combinando el aprendizaje automático basado en datos de expresión génica, estructura proteica y otras modalidades, con reglas simbólicas que codifican el conocimiento previo sobre sitios funcionales, dominios proteicos y vías biológicas.
- Interpretación Clínica: Para las variantes con potencial importancia clínica, es esencial interpretar su significado en términos de enfermedades, fenotipos o respuesta a tratamientos. Aquí, el componente simbólico del DSL puede aprovechar bases de datos de variantes clínicas y literatura científica, mientras que el aprendizaje profundo puede identificar patrones sutiles en los datos que

correlacionan variantes específicas con outcomes clínicos.

- Comprensión Evolutiva: El DSL también puede ser aplicado para entender las implicaciones evolutivas de las variantes, combinando la capacidad del aprendizaje profundo para analizar grandes conjuntos de datos genómicos de diferentes especies, con reglas y teorías evolutivas codificadas simbólicamente.
- Integración de Datos Multi-Omics: La genómica moderna va más allá de solo secuencias de ADN, incorporando también datos transcriptómicos, proteómicos y metabolómicos. El DSL es especialmente adecuado para esta tarea integrativa, ya que puede aprender representaciones unificadas de diferentes tipos de datos mientras razona sobre ellos en un marco simbólico coherente.
- Automatización y Escalabilidad: A medida que la cantidad de datos genómicos crece exponencialmente, es esencial que los sistemas de interpretación sean automáticos y escalables. El DSL, al combinar la eficiencia computacional del aprendizaje profundo con la estructura y coherencia del aprendizaje simbólico, ofrece una solución que puede procesar grandes volúmenes de datos de manera eficiente y precisa.

Propuestas actuales de arquitecturas de DSL

Recientemente, investigadores de IBM Research Zürich y ETH Zürich diseñaron una arquitectura que combina redes neuronales profundas y modelos vector-simbólicos, conocida como arquitectura neuro-vector-simbólica (NVSA). Esta arquitectura supera limitaciones anteriores, proporcionando un marco unificado para resolver tareas que involucran percepción y razonamiento de alto nivel. NVSA ha demostrado ser eficaz en la resolución de matrices progresivas de Raven, una tarea de razonamiento abstracto, con una eficiencia y precisión notables en comparación con otras arquitecturas

La arquitectura neuro-vector-simbólica (NVSA) propuesta por investigadores de IBM Research Zürich y ETH Zürich representa un paso innovador en la evolución de los sistemas de inteligencia artificial. A continuación, se describe con mayor detalle cómo funciona y qué la hace especial.

Las redes neuronales profundas (DNN, por sus siglas en inglés) son una subclase de las redes neuronales que tienen múltiples capas ocultas entre la entrada y la salida. Estas capas permiten a las DNN modelar y aprender patrones complejos y no lineales. Se han utilizado con éxito en una amplia variedad de tareas, especialmente en aquellas relacionadas con la percepción, como el reconocimiento de imágenes y el procesamiento de voz.

Por otro lado, los modelos vector-simbólicos se basan en representaciones simbólicas, lo que significa que trabajan con conceptos abstractos y relaciones entre ellos en lugar de patrones directos de datos. Estos modelos son especialmente útiles para tareas que requieren razonamiento y manipulación de símbolos, ya que pueden representar y trabajar con estructuras lógicas y semánticas.

La arquitectura NVSA combina la potencia de las DNN y los modelos vector-simbólicos. Mientras que las DNN se encargan de la percepción y la extracción de características de los datos de entrada, los modelos vector-simbólicos se ocupan del razonamiento de alto nivel y de la manipulación simbólica.

Este diseño híbrido permite a la NVSA superar limitaciones anteriores al proporcionar un marco unificado. En lugar de depender únicamente de las DNN para todas las tareas o de confiar solo en los sistemas simbólicos, esta arquitectura utiliza las fortalezas de ambos enfoques donde son más relevantes.

Un buen ejemplo de su eficacia es la resolución de matrices progresivas de Raven. Estas matrices son pruebas psicométricas diseñadas para evaluar el razonamiento abstracto de un individuo. Requieren tanto la percepción (identificar patrones visuales) como el razonamiento lógico (deducir la relación entre diferentes elementos y predecir el siguiente en la secuencia). La NVSA ha demostrado ser notablemente eficiente y precisa en esta tarea, superando a otras arquitecturas que solo utilizan uno de los dos enfoques.

La arquitectura neuro-vector-simbólica representa una integración prometedora de la percepción basada en DNN y el razonamiento basado en modelos vector-simbólicos. Su capacidad para abordar tareas que combinan ambas necesidades muestra su potencial para llevar a la inteligencia artificial a nuevos horizontes en términos de versatilidad y eficiencia.

References

Garcia-Retuerta D, Canal-Alonso A, Casado-Vara R, Rey AM, Panuccio G, Corchado JM. Bidirectional-Pass Algorithm for Interictal Event Detection. In Practical Applications of Computational Biology & Bioinformatics, 14th International Conference (PACBB 2020). PACBB 2020. Advances in Intelligent Systems and Computing, vol 1240. Springer, Cham. https://doi.org/10.1007/978-3-030-54568-0_20

Castillo Ossa LF, Chamoso P, Arango-López J, Pinto-Santos F, Isaza GA, Santa-Cruz-González C, Ceballos-Marquez A, Hernández G, Corchado JM. A Hybrid Model for

COVID-19 Monitoring and Prediction. Electronics. 2021; 10(7):799.
https://doi.org/10.3390/electronics10070799

Intelligent Platform Based on Smart PPE for Safety in Workplaces. Márquez-Sánchez S, Campero-Jurado I, Herrera-Santos J, Rodríguez S, Corchado JM. Sensors (Basel). 2021 Jul 7;21(14):4652
https://doi.org/10.3390/s21144652

A. Canal-Alonso, R. Casado-Vara and J. Manuel Corchado, "An affordable implantable VNS for use in animal research," 2020 27th IEEE International Conference on Electronics, Circuits and Systems (ICECS), 2020, pp. 1-4,
doi: 10.1109/ICECS49266.2020.9294958

An Agent-Based Clustering Approach for Gene Selection in Gene Expression Microarray. Ramos J, Castellanos-Garzón JA, González-Briones A, de Paz JF, Corchado JM. Interdiscip Sci. 2017 Mar;9(1):1-13
DOI 10.1007/s12539-017-0219-6

Agradecimientos

El presente estudio ha sido financiado por el proyecto AIR Genomics (con número de expediente CCTT3/20/SA/0003), mediante la convocatoria 2020 PROYECTOS I+D ORIENTADOS A LA EXCELENCIA Y MEJORA COMPETITIVA DE LOS CCTT por el Instituto de Competitividad Empresarial de Castilla y León y fondos FEDER