Centro de Información de COVID (CIC): Charlas científicas de relámpago

Transcripción de una presentación de Ho-Joon Lee (Universidad de Yale), febrero de 2022

Título: Un paisaje de interacciones proteína-proteína virus-huésped en la infección por SARS-CoV-2 en humanos por aprendizaje automático

Financiado por la Oficina de Infraestructura Cibernética Avanzada de la NSF, Dirección de Informática e Información e Ingeniería (OAC/CISE) a través del Programa del Fondo de Semillas del Northeast Big Data Innovation Hub.

Grabación de YouTube con diapositivas

Información del seminario web del CIC de febrero 2022

Editora de la Transcripción: Saanya Subasinghe

Editora de la Translación: Isabella Graham Martínez

Transcripción

Ho-Joon Lee:

Diapositiva 1

Hola, todo el mundo mi nombre es Ho-Joon Lee de la Escuela de Medicina de Yale y voy a hablar de *Un paisaje interactome de SARS-CoV-2 virus-humano proteína-proteína interacciones por aprendizaje automático.*

Diapositiva 2

Se trata de dos objetivos. La primera es desarrollar la secuencia de proteínas basada en el aprendizaje automático de clases múltiples o clasificadores de aprendizaje profundo para la evidencia o la predicción del nivel de confianza utilizando la base de datos Viruses.STRING. La segunda es - usando esos clasificadores queremos crear un proyecto de paisaje interactivo de las interacciones proteína-proteína humana del virus citoesquelético.

Diapositiva 3

Así que aquí hay una visión general de nuestro flujo de trabajo de aprendizaje automático y aprendizaje profundo. Así que usamos la base de datos Viruses.STRING, que no incluía el SARS-CoV-2 en el momento

del análisis. Esta es la red de PPI virus-humanos PPIs que contienen más de 80.000 interacciones entre alrededor de 1.200 proteínas de virus de 102 especies de virus y alrededor de 8.500 proteínas humanas. Y cada interacción tiene una puntuación combinada que va desde cero hasta mil, que convertimos en cinco clases de pruebas. Y esta es la distribución del número de IBP para las clases de evidencia. Y vamos a centrarnos en los IBP experimentales que pertenecen a la clase de evidencia 3 o 2 basados en el índice cero aquí. Y basado en los datos, primero extraemos las características de los nodos, otras características de las proteínas que son composiciones fraccionarias de 20 aminoácidos. Y en este punto estamos desarrollando dos modelos diferentes - uno es más canónico [inaudible] modelos como Random Forests y XGBoost, en este caso. Y otro se basa en el aprendizaje profundo. Utilizamos específicamente redes neuronales gráficas como GraphSAGE o versión datalizada de HinSAGE. Para el aprendizaje automático conectado, también extraemos las características de Edge que son 72 medidas de distancia o similitud entre los perfiles de composición de aminoácidos entre las proteínas del virus y las proteínas humanas. Y sobre la base de las características, desarrollamos los bosques aleatorios y XGBoost. Para Random Forests, optimizamos 36 modelos mediante investigación con regulación de solicitud temporal y 432 modelos para espacio ejecutivo que tiene el mismo trasplante contemporáneo. Y, en resumen, obtenemos hasta un 67% de precisión y un 37% de precisión para los casos de Random Forests y un 74% de precisión y un 67% de precisión para los casos de XGBoost. Y este trabajo, esta parte, ha sido publicado como preprint recientemente. Así que puede consultar el artículo en detalle [https://www.biorxiv.org/content/10.1101/2021.11.07.467640v2]. Y para GraphSAGE aquí presente, todavía en lectura avanzada y preparación, pero voy a mostrarles, brevemente, los resultados de GraphSAGE también. Porque esto muestra más del 70% de precisión, lo que también es muy prometedor.

Diapositiva 4

Y aquí voy a mostrarles un ejemplo de rendimiento para los mejores modelos para el 20% de [inaudible] con esta semilla aleatoria. Vemos, en este caso, cuando el bosque muestra un 60% de precisión, XBG fue un 67,7% de precisión. Y si nos fijamos en las métricas informáticas, de nuevo, voy a centrarme en este [EC3?] que implica principalmente PPIs expandidos. Y si nos fijamos en las clases individuales, centrándose en el f1-score, el refuerzo adicional muestra mejores resultados f1 en cuatro clases individuales.

Diapositiva 5

Basado en, basado en este modelo de impulso [inaudible]. Las características importantes se identificaron utilizando dos métodos alternativos aquí. Uno por el índice de Gini y el otro por el análisis de SHAP, que se basa en SHAP, llegó a los valores de SHAP. Y curiosamente, vemos que la cisteína y la histidina son la mayoría - dos características más importantes. Donde este menos [C_minus y H_minus] significa que la fracción de cisteína entre virus y humanos. Y la relación significa la relación entre las fracciones cisteína y reacciones histidinas entre virus y humanos.

Diapositiva 6

Un experimento de control que realizamos es comparar la predicción de IBP experimentales y - con una predicción de IBP de minería de texto en el virus [inaudible]. Debido a que el tamaño de los datos, la diferencia es bastante grande aquí, seis diferencia al cuadrado, pero lo que observamos aquí es que XGBoost, de hecho, muestra una mayor precisión. Con 94% de precisión en comparación con 90% de precisión para el caso de minería de texto. Así que a pesar de la diferencia de tamaño de los datos, XGBoost muestra un buen rendimiento de predicción. Y este es el acuerdo entre las actividades de fuerza aleatoria para ec3 y la prueba de unión como esperamos muestra principalmente ec1 o ec2.

Diapositiva 7

Así que basándonos en esos resultados alentadores, aplicamos esos clasificadores al SARS-CoV-2 para nuestro segundo objetivo de dos maneras. Primero, aplicamos eso a la base de datos IntAct, que es una colección de PPIs experimentales. Y aquí les muestro la red de XGBoost con [inaudible] evidencia predicha. Entonces EC3 para azul, EC4, rojo. Así que esto puede ser visto como priorizar una red. Así que aunque estos enlaces serían unos 2.000 enlaces de datos experimentales son igualmente significativos, también podemos priorizar los enlaces basados en esta clase [inaudible] predijo XGBoost en este caso. En segundo lugar, también aplicamos eso a la interacción a nivel de toda la proteína [inaudible] los viejos pares de más de medio millón entre 27 proteínas SARS-CoV-2 y aproximadamente más de 20.000 proteínas humanas. Y aquí les muestro el subconjunto de 22.000 PPIs con clase de evidencia de al menos 2. O bien uso XGBoost o Random Forest. Y este es el otro subconjunto - 140 PPIs con la clase de evidencia más alta, 5, por XGBoost. Y sobre la base de esta red de interacción observamos que muchas proteínas humanas están enriqueciendo la contracción del músculo liso vascular y los objetivos y también los componentes de H2A.

Diapositiva 8

Hay algunas aplicaciones más de este trabajo que se han encontrado en el último mes, en realidad. Así que Giuseppe Novelli, que es el renombrado genetista en Roma, en Italia, me contactó por correo electrónico y por sorpresa, el mes pasado. Había leído mi preprint contándome sobre su publicación terapéutica de calidad para ligasas HECT E3 y su idea de utilizar los resultados de este importante trabajo a través de esta investigación en curso. E inmediatamente nos dimos cuenta de que podemos ayudarnos unos a otros en base a mis resultados en los resultados de la red interactiva. Y encontramos que la proteína de dominio HECT tiende a interactuar con las proteínas SARS-CoV-2 con una clase de evidencia mayor que 2 con significación estadística. En otras palabras, las proteínas del dominio HECT son favorecidas por el SARS-CoV-2. Con base en esa observación usted se pregunta si hay otras familias de proteínas favorecidas por el SARS-CoV-2. Además, también podemos extender eso a otras especies de virus como el metapneumovirus humano, en el que también está trabajando el Dr. Novelli.

Diapositiva 9

Así que finalmente, voy a mostrarles brevemente sobre las redes neuronales gráficas que usan la arquitectura GraphSAGE y HinSAGE. A la izquierda están las precisiones de 15 modelos diferentes usando

tres pesos diferentes de Java en las columnas y cinco métodos diferentes de incrustación de bordes. Como ven, sin tasas de abandono, de hecho, vemos más del 70% de valores de precisión y precisiones que son muy prometedores. Esto se basa en Viruses.STRING y si aplicamos eso a la base de datos IntAct de SARS-CoV-2, verá que la predicción está enriquecida con evidencia de clase 2, o 3, de hecho, que son en su mayoría IBP [inaudibles]. Y este consenso es el número de acuerdos entre estos 15 modelos diferentes. Vemos más consenso de, como, 8-9 para contra dos en comparación con 6-7 contra 1, pero creo que esto también es muy importante porque - vamos a ver.

Diapositiva 10

De acuerdo, con eso me gustaría agradecer a mis colaboradores por discusiones muy útiles y comentarios y apoyo. Y el Centro de Yale para Computación de Investigación para recursos computacionales. Y la comunidad COVID HASTE de la Escuela de Ingeniería y Ciencias Aplicadas de Yale. Y finalmente el Fondo Semilla del Centro de Big Data del Noreste para apoyar este trabajo. Gracias.