INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL PARA LAS CIENCIAS E INGENIERIAS

PROYECTO FINAL

INTEGRANTE:

JUAN JOSÉ RAMÍREZ CUERVO

DOCUMENTO:

1000869826

PROFESOR:

RAÚL RAMOS POLLÁN

UNIVERSIDAD DE ANTIOQUIA

PROGRAMA DE BIOINGENIERÍA

FECHA:

12 DE NOVIEMBRE DE 2022

INTRODUCCIÓN

La constante de acoplamiento escalar, medida directamente por espectroscopia de resonancia magnética nuclear, es un parámetro clave para el análisis de estructuras moleculares y se usa ampliamente para predecir estructuras moleculares desconocidas. Estas son las interacciones magnéticas entre un par de átomos. La fuerza de esta interacción magnética depende de los electrones intermedios y los enlaces químicos que forman la estructura tridimensional de una molécula.

Debido al alto costo del tiempo y espacio computacional sustancial de los experimentos de RMN, es increíblemente desafiante medir la constante de acoplamiento escalar de moléculas desconocidas a gran escala.

Las redes neuronales gráficas de inteligencia artificial tienen un gran potencial en la construcción de modelos de topología de tipo molecular, lo que les otorga la capacidad de predecir rápidamente la constante de acoplamiento escalar a través de métodos de aprendizaje automático basados en datos y evitar cálculos químicos cuánticos que consumen mucho tiempo. Es un método rápido y fiable para predecir estas interacciones permitirá a los químicos médicos obtener conocimientos estructurales de forma más rápida y económica, lo que permitirá a los científicos comprender cómo la estructura química 3D de una molécula afecta a sus propiedades y comportamiento. En última instancia, estas herramientas permitirán a los investigadores avanzar en una variedad de problemas importantes, como el diseño de moléculas para llevar a cabo tareas celulares específicas o el diseño de mejores moléculas de fármacos para combatir enfermedades.

En este proyecto se quiere a cabo la predicción de la constante de acoplamiento escalar en una serie de moléculas mediante inteligencia artificial y Machine Learning dados los datos de resonancia magnética nuclear de una competición en Kaggle [1]. Este experimento es útil para comprender mejor la estructura de las moléculas en áreas como la ciencia ambiental, la ciencia farmacéutica y la ciencia de materiales.

Los datos se evalúan con el método Mean Absolute Error, se calculan para cada tipo de acoplamiento escalar y luego se promedian entre tipos, de modo que una disminución del 1 % en MAE para un tipo proporciona la misma mejora en la puntuación que una disminución del 1 % para otro tipo.

El criterio sobre cuál sería el desempeño deseable en producción para esta métrica, ya que el MAE para cualquier grupo tiene un piso de 1e-9, la puntuación mínima (mejor) posible para predicciones perfectas es de aproximadamente -20,7232.

DATASET

Este dataset tiene más de 130000 muestras de las interacciones entre los átomos, 47 columnas de las cuales algunas son cualitativas al indicar el índice del átomo repartidos entre 9 archivos csv.

Las columnas utilizadas para reprocesar los dataframes fueron:

- molecule_name: nombre de la molécula, al final no se incluyó
- atom_index_0: índice del primer átomo en cada interacción
- atom_index_1: índice del segundo átodo en cada interacción
- type: tipo de molécula de cada interacción
- X: momento dipolar en el eje X de la interacción
- Y: momento dipolar en el eje Y de la interacción
- Z: momento dipolar en el eje Z de la interacción
- potential_energy: energía potencial de la interacción
- 'fc', 'sd', 'pso', 'dso' son datos que aportan al resultado final de scalar_coupling_constant
- scalar_coupling_constant: constante de acoplamiento escalar

Estos datos son correspondientes a cada una de la interacciones entre los átomos, por eso se considera que son las más importantes y por ende las que se eligen para reorganizar los dataframes en uno solo.

En el caso de la columna ‘type’, se reorganiza para mostrarse en forma de one-hot en el dataframe, por lo que el número de columnas pasa de ser 1 a ser 8, por los 8 tipos diferentes de moléculas.

Dado que se tienen muchos datos, ya que entre las filas de train y de test se tienen más de 7 millones de filas, se reduce el número de estas eligiendo las moléculas desde la 1 hasta la 30000. De esta manera el número de filas es de aproximadamente 1 millón y de esta forma es posible que el colab lo ejecute

	id	molecule_name	atom_index_0	atom_index_1	type	scalar_coupling_constant
0	0	dsgdb9nsd_000001	1	0	1JHC	84.80760
1	1	dsgdb9nsd_000001	1	2	2JHH	-11.25700
2	2	dsgdb9nsd_000001	1	3	2JHH	-11.25480
3	3	dsgdb9nsd_000001	1	4	2JHH	-11.25430
4	4	dsgdb9nsd_000001	2	0	1JHC	84.80740
...	...	...	...	...	...	...
861374	861374	dsgdb9nsd_029999	16	4	2JHC	2.20243
861375	861375	dsgdb9nsd_029999	16	5	3JHC	8.13388
861376	861376	dsgdb9nsd_029999	16	6	3JHN	3.09138
861377	861377	dsgdb9nsd_029999	16	7	2JHC	7.16273
861378	861378	dsgdb9nsd_029999	16	8	1JHC	133.75200

861379 rows × 6 columns

Figura 1. Dataframe de entrenamiento hasta 30000 moléculas

	id	molecule_name	atom_index_0	atom_index_1	type
0	4659076	dsgdb9nsd_000004	2	0	2JHC
1	4659077	dsgdb9nsd_000004	2	1	1JHC
2	4659078	dsgdb9nsd_000004	2	3	3JHH
3	4659079	dsgdb9nsd_000004	3	0	1JHC
4	4659080	dsgdb9nsd_000004	3	1	2JHC
...	...	...	...	...	...
460447	5119523	dsgdb9nsd_030000	15	4	3JHC
460448	5119524	dsgdb9nsd_030000	15	5	3JHC
460449	5119525	dsgdb9nsd_030000	15	6	2JHN
460450	5119526	dsgdb9nsd_030000	15	7	1JHC
460451	5119527	dsgdb9nsd_030000	15	8	2JHN

460452 rows × 5 columns

Figura 2. Dataframe de test hasta 30000 moléculas

En este dataframe no hay ningún dato faltante, sin embargo en la tarea se pide que por lo menos el 5% deben ser faltantes, por lo que se eligen las tres columnas de momento dipolar en X, Y y Z para eliminar los datos iguales a 0.

Estos datos posteriormente deben ser llenados con algún valor para poder hacer el fit con los modelos. Se elige llenar los valores faltantes de cada columna con el promedio correspondiente a cada columna.

id	atom_index_0	atom_index_1	type_1JHC	type_1JHN	type_2JHC	type_2JHH	type_2JHN	type_3JHC	type_3JHH	type_3JHN
0	1	0	1	0	0	0	0	0	0	0
1	1	2	0	0	0	1	0	0	0	0
2	1	3	0	0	0	1	0	0	0	0
3	1	4	0	0	0	1	0	0	0	0
4	2	0	1	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...
861374	16	4	0	0	1	0	0	0	0	0
861375	16	5	0	0	0	0	0	1	0	0
861376	16	6	0	0	0	0	0	0	0	1
861377	16	7	0	0	1	0	0	0	0	0
861378	16	8	1	0	0	0	0	0	0	0

861379 rows × 18 columns

Figura 3. Dataset de entrenamiento final

id	atom_index_0	atom_index_1	type_1JHC	type_1JHN	type_2JHC	type_2JHH	type_2JHN	type_3JHC	type_3JHH	type_3JHN
4659076	2	0	0	0	1	0	0	0	0	0
4659077	2	1	1	0	0	0	0	0	0	0
4659078	2	3	0	0	0	0	0	0	1	0
4659079	3	0	1	0	0	0	0	0	0	0
4659080	3	1	0	0	1	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...
5119523	15	4	0	0	0	0	0	1	0	0
5119524	15	5	0	0	0	0	0	1	0	0
5119525	15	6	0	0	0	0	1	0	0	0
5119526	15	7	1	0	0	0	0	0	0	0
5119527	15	8	0	0	0	0	1	0	0	0

460452 rows × 18 columns

Figura 4. Dataset de test final

REFERENCIAS

[1] https://www.kaggle.com/competitions/champs-scalar-coupling/data