Técnicas de Aprendizado de Máquina Aplicadas aos Dicalcogenetos de Metais de Transição

Guilherme S. Marcon[1], Naidel A. M. S. Caturello[2], Marcos G. Quiles[3], Juarez L. F. Da Silva[2*]

[1]Instituto de Ciências Matemáticas e de Computação de São Carlos; [2]Instituto de Química de São Carlos; [3]Universidade Federal de São Paulo.

*juarez_dasilva@iqsc.usp.br

Objetivos

Dicalcogenetos de metais de transição bidimensionais (DMTs 2D) possuem a fórmula MQ_2, onde *M* é um metal de transição e *Q* = S, Se, Te, com diversas propriedades e aplicações.[1] O tempo de cálculo para propriedades moleculares utilizando teoria do funcional da densidade (DFT) destes compostos é um entrave para estudos de DMTs 2D contendo centenas de átomos. Com o objetivo de diminuir o tempo de obtenção de propriedades moleculares de DMTs 2D, mantendo acurácia comparável à dos cálculos de DFT, objetiva-se no presente projeto a utilização de técnicas de aprendizado de máquina e aplicá-las na previsão dessas propriedades.

Métodos e Procedimentos

Os conjuntos de dados são provenientes do trabalho de Caturello et al.,[2] com $(MoQ_2)_n$, n = 1 – 16 e Q = S, Se, Te. Representam-se as moléculas como vetor de característica, através da Matriz de Coulomb e suas variações.[3] Os experimentos foram conduzidos utilizando a validação cruzada estratificada[4]. Os seguintes métodos de regressão foram consideradados: Linear, Kernel Ridge Regression e Redes Neurais MLP,[4,5] visando a predição das energias totais (E_{tot}). Por fim, cada conjuntos de dados foram expandidos de uma média de 600 moléculas para 7000. Diferente da literatura, que reporta os treinos pelo erro médio absoluto (MAE), optou-se por reportar a porcentagem do erro médio absoluto (MAPE), pelas grandes variações de E_{tot} com o tamanho das moléculas.

Resultados

Os melhores resultados foram utilizando: os autovalores da Matriz de Coulomb como vetor de característica, cujo tamanho é igual à quantidade máxima de átomos nas moléculas do conjunto, além disso, os autovalores não variam se a matriz possuir troca de linhas ou colunas, gerando representações mais únicas e menores; a validação cruzada estratificada para separação de treino e teste, com os datasets expandidos. Com essas configurações, os melhores modelos foram o Linear e o Kernel Ridge com kernel linear, seus MAPEs foram de $2{\cdot}10^{-5}$%. As Redes Neurais logo em seguida, com um erro de $2{\cdot}10^{-4}$%.

Conclusões

A quantidade de dados é essencial para a qualidade da previsão. Os autovalores da Matriz de Coulomb representam com grande acurácia moléculas com n <= 16. Ao contrário da matriz completa, que se obtém um vetor de característica polinomialmente maior, o que dificulta a regressão. Utilizar a validação cruzada estratificada também se mostrou melhor, já que ela também garante uma melhor distribuição das moléculas nas divisões. Futuramente, realizar-se-á os mesmos testes com moléculas de até 315 átomos.

Referências Bibliográficas

[1] Chen, W; et al. ACS Nano **2018**, *12*, 308–316.

[2] Caturello, N. A. M. S.; et al. J. Phys. Chem. C **2018**, *122*, 27059–27069.

[3] Montavon, G; et al. *Adv. Neur. In.* **2012**, *25*, 440-448.

[4] Friedman, J.; et al. *The elements of statistical learning*; Springer, NY, 2001; Vol. *1(10)*.

[5] Haykin, S. S. *Neural Networks and Learning Machines*; Prentice Hall, NY, 2009; Vol. *3*.