Explorando Algoritmos de Agrupamento Alternativos além do k-means: O Exemplo de Estruturas Moleculares Usadas na Produção de Hidrogênio

Marcos V. C. R. da Trindade[1], Marcos G. Quiles[2], Juarez L. F. Da Silva[3]

[1]Instituto de Ciências Matemáticas e de Computação, USP, São Carlos, SP, Brasil.
[2]Instituto de Ciência e Tecnologia, UNIFESP, São José dos Campos, SP, Brasil
[3]Instituto de Química de São Carlos, USP, São Carlos, SP, Brasil.

marcos_cota@usp.br

Objetivos

O principal objetivo deste trabalho é realizar um estudo aprofundado sobre o uso de diferentes algoritmos de clusterização em cenários de química computacional. Busca-se ir além das limitações do kmeans [1] e investigar como métodos mais robustos, como Hierarchical Density Based Spatial Clustering (HDBSCAN) [2], Gaussian Mixture Models (GMM) [3] e Spectral Clustering [4], se adaptam a dados de alta dimensionalidade e a padrões de agrupamento não esféricos ou irregulares. O foco é identificar a abordagem mais eficaz para o agrupamento de configurações moleculares com base nos autovalores das matrizes de Coulomb [5].

Um objetivo secundário, mas igualmente fundamental, é a otimização da ferramenta computacional existente, de modo a tornar viável o processamento de milhões de configurações moleculares. A remoção de gargalos computacionais, alcançada por meio de estratégias de paralelização e reorganização do fluxo de execução, permitirá a realização de testes empíricos em larga escala com maior agilidade. A disponibilização dessa ferramenta aprimorada para a comunidade científica é também uma meta importante, pois cria condições para ampliar a colaboração entre pesquisadores, simplificar análises exploratórias e acelerar descobertas em áreas como química quântica e energias sustentáveis [6].

Métodos e Procedimentos

O estudo foi conduzido com base em uma ferramenta computacional própria, desenvolvida para automatizar todas as etapas de geração, processamento e análise de estruturas moleculares. O funcionamento inicia-se com a criação das estruturas, que são definidas a partir de restrições geométricas cúbicas ou esféricas. Em seguida, podem ser realizadas permutações de átomos e adição de ligantes, o que aumenta a diversidade estrutural do conjunto. Após cada processo, calcula-se as matrizes de Coulomb de cada estrutura e extraem-se seus autovalores, que funcionam como representações numéricas e são utilizados como entrada para os algoritmos de clusterização [5].

Na fase de análise são aplicados métodos de agrupamento como kmeans, HDBSCAN, Gaussian Mixture Models e Spectral Clustering, com a finalidade de comparar o desempenho diante de diferentes características dos dados. Para tornar o processo eficiente em larga escala, foram implementadas otimizações como a manutenção de um dicionário atômico em memória, que reduziu de forma expressiva o tempo de leitura de arquivos, e a paralelização do cálculo das matrizes de Coulomb, que acelerou o processamento total. Também foram incorporadas projeções em PCA e métricas internas de qualidade que auxiliam a interpretação dos agrupamentos e

permitem comparações quantitativas entre os algoritmos.

Resultados

As melhorias implementadas resultaram em uma redução de aproximadamente 86% no tempo total de processamento, garantindo a viabilidade do projeto. A combinação de otimização e paralelização tornou a plataforma mais ágil e confiável, criando condições adequadas para o estudo comparativo entre os algoritmos. A análise de desempenho mostrou, contudo, que o cálculo dos autovalores ainda é a etapa mais custosa em termos de tempo, representando um ponto de atenção para trabalhos futuros. Com a conclusão desta primeira fase, a ferramenta encontra-se pronta para ser utilizada na investigação empírica. A Figura 1 ilustra um exemplo de visualização em PCA, na projeção PC1 contra PC2, após a aplicação de ligantes, destacando os clusters formados. Esse tipo de visualização é fundamental para compreender a organização dos grupos e avaliar a eficácia de cada algoritmo na representação do espaço amostral.

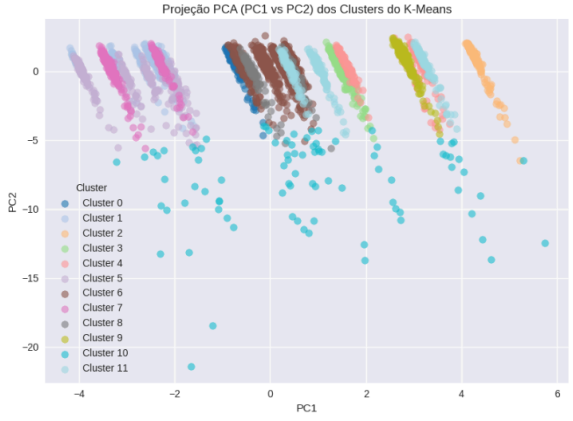

Figura 1: Exemplo de visualização de dados em PCA (PC1 × PC2), mostrando clusters formados após a aplicação de ligantes.

Conclusões

O projeto, em sua fase atual, já demonstra avanços significativos tanto em desempenho quanto em confiabilidade da ferramenta desenvolvida. A estratégia de otimização adotada mostrou-se bem-sucedida e assegurou as condições necessárias para o estudo comparativo entre diferentes algoritmos de clusterização. Os próximos resultados devem permitir a identificação da abordagem mais eficaz para o agrupamento de dados moleculares, o que terá impacto direto na forma como grandes conjuntos de dados são analisados em química computacional. A seleção de configurações moleculares representativas será assim acelerada, contribuindo para a inovação em pesquisas aplicadas e para a expansão do uso de métodos de aprendizado não supervisionado em química teórica e computacional [6].

Os autores declaram não haver conflito de interesses. Marcos V. C. R. da Trindade realizou as otimizações necessárias para o desenvolvimento do projeto. O colaborador Marcos G. Quiles e o orientador Juarez L. F. Da Silva planejaram, orientaram e revisaram o estudo, aprovando todos a versão final.

Referências

[1] A. K. Jain. Data clustering: 50 years beyond kmeans. Pattern Recognit. Lett., 31(8):651–666, 2010.

[2] H. P. Kriegel, P. Kröger, J. Sander, and A. Zimek. Density based clustering. WIREs Data Min. Knowl. Discov., 1(3):231–240, 2011.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. B, 39(1):1–38, 1977.

[4] D. Pandove, S. Goel, and R. Rani. Systematic review of clustering high dimensional and large datasets. ACM Trans. Knowl. Discov. Data, 12(2):21, 2018.

[5] J. Schrier. Can one hear the shape of a molecule (from its Coulomb matrix eigenvalues)? J. Chem. Inf. Model., 60(8):3804–3811, 2020.

[6] A. Glielmo, B. E. Husic, A. Rodriguez, C. Clementi, F. Noé, and A. Laio. Unsupervised learning methods for molecular simulation data. Chem. Rev., 121(16):9722–9758, 2021.