Benchmark of Clustering Techniques and Their Potential Applications to Polymer Material Science

Rachel Drake and Dr. Medina

LIBERTY UNIVERSITY

TRACER Lab

Abstract and/or Background

This benchmark investigates graph theory-based clustering techniques and their potential connections in materials modeling. Fields in, drug, perfume, and synthetic food development have been using graph-based research to refine their algorithms' predictive capabilities and create new chemical properties. These clustering methods can potentially be used to predict polymer properties and assist in accelerating the design of advanced multi-functional materials. Clustering is a method that groups together complex networks such that data points within a group lie within similar, specified parameters. Clustering computationally investigates patterns that help predict properties and abstract data. The strengths and weaknesses of various clustering methods are based on three points of criteria. First, the clustering algorithms are compared in terms of their mathematical and computational implementation in terms of complexity and usability. Then, their outputs are compared in terms of reliability, usefulness, and reproducibility. Finally, the clustering algorithms' time complexities are compared as the data sets grow larger. The following clustering techniques that are compared include commonly used methods for biological data such as k-means clustering and hierarchical clustering. Other algorithms included in this study are spectral clustering, motif-aware clustering, and graph-based Multiview clustering. These techniques will aid in advancing the investigation on how to embed polymers into graphs and find the best methods to get motifs and patterns that can represent complex polymeric structures and create better opportunities for making multi-functional polymeric membranes.

Introduction

Clustering algorithms divide complex neural networks into groups such that data points within that group are similar. Researchers use these methods to abstract and understand patterns within complex networks (Figure 1). Currently, clustering algorithms assist in the prediction of molecular and chemical properties. Some of the most effective clustering algorithms include k-means clustering, hierarchical clustering, spectral clustering, motif-aware clustering, and graph-based multiview clustering. Using modern clustering techniques on polymers represented as graphs could assist in the creation of multi-functional polymeric membranes. A comparison of these different techniques could assist in refining what clustering algorithms will be most beneficial in this field and further push research to improve these algorithms. A review of these algorithms' implementation, usefulness, and time complexity is a start to revealing the strengths and weaknesses of each of these algorithms.

Methods

Based on literature reviews, clustering algorithms were evaluated based on implementation, usability, and time complexity.

1. For implementation, the written algorithm, as well as the basic code for the algorithm on Python, with mathematical explanations were reviewed.
2. Usability was determined by assessing the most current algorithms used in the real world, along with literature reviews that compared these methods. These reviews determined if the algorithm was efficient in molecular simulations and development. If it was not a common method due to it being a more novel design, articles were reviewed to show the potential for these algorithms in polymer research.
3. Finally, time complexity was determined by comparing multiple variations of the algorithms. Ideas were also compared on how these algorithms may be improved and run faster in the future.

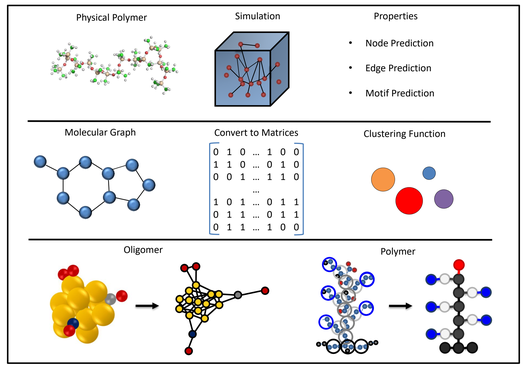

Figure 1: A representation of the basic steps in polymer simulations. Beginning with a physical polymer, simulations are run that provide methods to predict properties. This is accomplished by developing models and abstracting the polymer data into molecular graphs which are embedded into matrices and clustered based on desired conditions (Gartner, et. al., 2019). This process embeds models of oligomers, atoms to nodes, and polymers, groups of atoms to nodes (Deshpande, et. al., 2020).

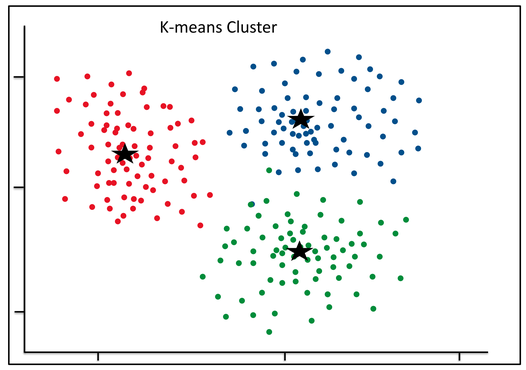

Figure 2: K-means creates k-clusters, each data point is associated with a given cluster in relation to its distance to the centroids, the stars. The algorithm computes the distance between each input to the centroid, and then reassigns it to minimize distance from the centroid and maximize distance from each other. (Rodriquez, et. al., 2019)

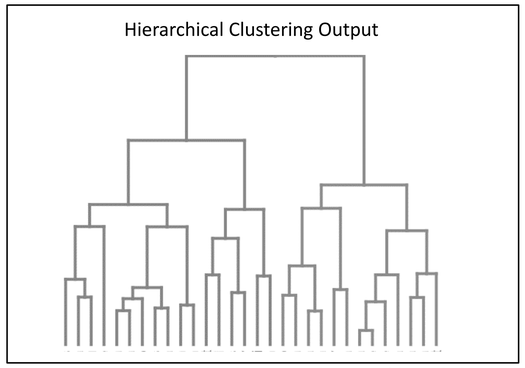

Figure 3: Data is divided or combined into a tree. The output of this method is a dendrogram, represented above, where the increasing dissimilarities of the data are represented by taller heights. (Caesar, et. al., 2018).

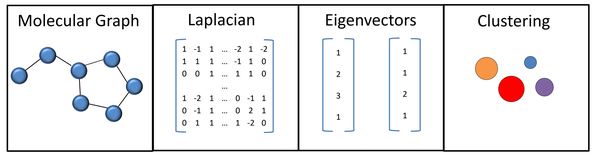

Figure 4: A given molecular graph is represented as a graph Laplacian matrix which derives the strongly connected nodes in the graph. Then the computed eigenvectors of this matrix are used to cluster the data points (Wiskott & Schonfeld, 2019).

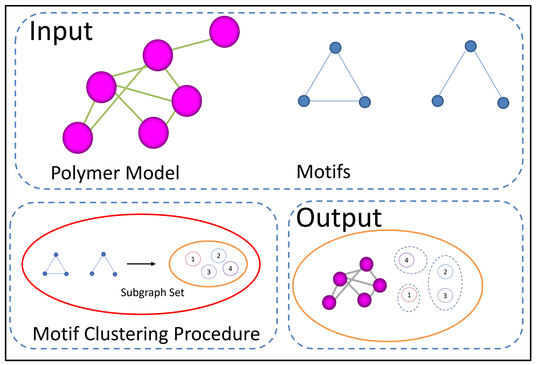

Figure 5: Given a polymer model, data is abstracted as a graph. Then the model is analyzed for a given motif to create subgraphs based on the connectedness of the set and partitioned by specified conditions to create clusters (Feng, et. al., 2019).

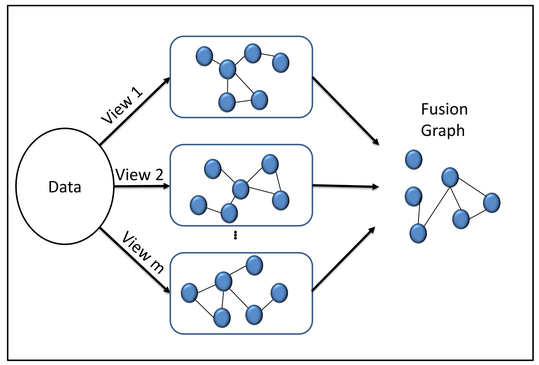

Figure 6: Based on given data, the graph-based multiview clustering algorithm first represents data from many different views, and the fuses these graphs together for clustering (Wang, et. al., 2020).

Results and/or Conclusion

K-Means Clustering

K-means clustering is an unsupervised machine learning algorithm, that partitions a group into k-clusters. It minimizes the distance within the clusters around a centroid, while maximizing its centroids distances from other centroids (Rodriquez, et. al., 2019) (Figure 2). This is a popular method of clustering due to its ease of implementation and readable results. One of the major weaknesses involves complications in finding optimal values for k, which will directly affect the outcome (Yuan & Yang, 2019). This clustering method was used to understand mechanical and physical properties of carbon-fiber reinforced polymers (Kurita, et. al., 2022). This algorithm is slower for large data sets and generally has a time complexity of $O(n^2)$ (Yuan & Yang, 2019).

Hierarchical Clustering

Hierarchical clustering is a greedy algorithm, that combines or divides data into clusters, creating a tree graph (Rani & Rohil, 2013) (Figure 3). This method is easy to use, versatile and well-defines clusters similarities (Partheniadis, et. al., 2020). However, this algorithm is inflexible and does not always represent the number of clusters accurately (Rani & Rohil, 2013). This method compared polymer functionalities and patterns to select exchangeable polymers (Partheniadis, et. al., 2020). This algorithm, like k-means, has a time complexity of $O(n^2)$ (Caesar, et. al., 2018).

Spectral Clustering

Spectral clustering provides quality representations of important features in graph neural networks (Zhiheng, et. al., 2020). This method transforms a given graph into a Laplacian matrix, that accurately determines strongly connected nodes within the graph (Wiskott & Schonfeld, 2019) (Figure 4). Spectral clustering is a well-used and current method that can accurately cluster molecular features derived from graphs (Glielmo, et. al., 2021). However, this method has a worse time complexity of $O(n^3)$ due to it requiring many matrices operations and needing another clustering algorithm, like k-means, to derive results. (Yan, Huang, & Jordan, 2009).

Motif Clustering

Motif clustering determines clusters by finding dense motif patterns to determine interconnectedness, and then form clusters based on desired, conditioned similarities (Feng, et. al., 2019) (Figure 5). This method was used to analyze bio-polymers and recognized patterns and the frequencies of these patterns (Helfrecht, et. al., 2019). Improvements in terms of time complexity are needed since it is currently $O(rn^3)$, where r is the number of subgraphs formed by a given motif (Feng, et. al., 2019).

Graph-based multiview Clustering

Graph-based multiview clustering takes many different views of a given graph(s), and then fuses them to create a unified representation of the data (Wang & Hang, 2020) (Figure 6). This is a novel method that has a lot of potential in providing accurate descriptions of data, and better predictions (Wang & Hang, 2020). Predictions on protein folds were accurately represented with this method using multiple graph inputs based on various data sources (Yan, et. al.,2021). This method has the worst time complexity that is dependent on the views and data representations, but has potential for improvement (Wang & Hang, 2020).

Conclusion

An overall direct comparison of these clustering methods against each other proves difficult, since they have varying capabilities in terms of outputs, performance on the desired data, and time complexity capabilities dependent on the input. While k-means, hierarchical, and spectral clustering have been more heavily researched, there is potential in using motif and multiview clustering in polymer analysis. Throughout this study it was found that many of these algorithms are often used together for improved clustering outputs, and their integration increases the instant understanding and prediction capabilities in analyzing drug, food, perfume, and polymer data properties. Future studies on optimizing these methods with one another will assist in designing the most beneficial clustering algorithms to understand complex graph neural networks and create polymeric multi-functional membranes.

References

Caesar L., Kvalheim, O., & Nadja C. (2018). Hierarchical cluster analysis of technical replicates to identify interferents in untargeted mass spectrometry metabolomics, Analytica Chimica Acta, Volume 1021, 2018, Pages 69-77, ISSN 0003-2670. https://doi.org/10.1016/j.aca.2018.03.013

Deshpande, S., Maxson, T. & Greeley, J. (2020). Graph theory approach to determine configurations of multidentate and high coverage adsorbates for heterogeneous catalysis. npj Comput Mater 6, 79. https://doi.org/10.1038/s41524-020-0345-2

Feng Y, et. al., (2019). COMICS: a community property-based triangle motif clustering scheme. PeerJ Comput Sci. doi: 10.7717/peerj-cs.180. PMID: 33816833; PMCID: PMC7924480

Gartner, T. & Jayaraman, A. (2019). Modeling and Simulations of Polymers: A Roadmap. Macromolecules. 52. 10.1021/acs.macromol.8b01836.

Glielmo, A., et. al. (2021). Unsupervised Learning Methods for Molecular Simulation Data. Chem Rev. 121(16), 9722-9758. https://doi.org/10.1021/acs.chemrev.0c01195

Helfrecht B.A., Gasparotto P, Giberti F, & Ceriotti M. (2019) Atomic Motif Recognition in (Bio)Polymers: Benchmarks From the Protein Data Bank. Front Mol Biosci. 6:24. doi: 10.3389/fmolb.2019.00024. PMID: 31058166; PMCID: PMC6482324

Kurita H., et. al. (2022). k-Means Clustering for Prediction of Tensile Properties in Carbon Fiber-Reinforced Polymer Composites. Advanced Engineering Materials. https://doi.org/10.1002/adem.202101072

Partheniadis, I., Toskas, M., Filippos-Michail S., Menexes, G., & Nikolakakis, I. (2020). Impact of Hot-Melt-Extrusion on Solid-State Properties of Pharmaceutical Polymers and Classification Using Hierarchical Cluster Analysis. Processes, 8(10), 1208. http://dx.doi.org/10.3390/pr8101208

Rani, Y., & Rohil, D.H. (2013). A Study of Hierarchical Clustering Algorithm. International Journal of Information and Computation Technology. ISSN 0974-2239 Volume 3, Number 10 (2013), pp. 1115-1122

Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. d. F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. PloS One, 14(1), e0210236-e0210236. https://doi.org/10.1371/journal.pone.0210236

Yan D., Huang L., and Jordan M. (2009). Fast approximate spectral clustering. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. Association for Computing Machinery, New York, NY, USA, 907–916.

Yan K., Wen J., Y. Xu and B. Liu. (2021). "Protein Fold Recognition Based on Auto-Weighted Multi-View Graph Embedding Learning Model," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 18, no. 6, pp. 2682-2691. doi: 10.1109/TCBB.2020.2991268

Yuan, C., & Yang, H. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. J, 2(2), 226–235. MDPI AG. Retrieved from http://dx.doi.org/10.3390/j2020016

Wang H., Yang Y. and Liu B.. (2020) "GMC: Graph-Based Multi-View Clustering," in IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 6, pp. 1116-1129. doi: 10.1109/TKDE.2019.2903810

Wang W., Wang H., Zhou J., Fan H., & Liu W. (2020). Machine learning prediction of mechanical properties of braided-textile reinforced tubular structures, Materials & Design, Volume 212, 2020, 110181, ISSN 0264-1275, https://doi.org/10.1016/j.matdes.2021.110181

Wiskott, L., & Schönfeld, F. (2019). Laplacian Matrix for Dimensionality Reduction and Clustering. ArXiv, abs/1909.08381

Zhiheng L., Wellawatte G. P., Chakraborty, M., Gandhi, H. A., Xu, C., & White, A. D. (2020). Graph neural network based coarse-grained mapping prediction. Chemical Science (Cambridge), 11(35), 9524-9531. https://doi.org/10.1039/d0sc02458a