Comparitive Analysis of Clustering Algorithms

Polavarapu Bhavish , Mohammad
Ayan, Kartik Aggarwal, Shivani Atigre

***Abstract*—Clustering is a technique in machine learning and data analysis that aims to group similar objects or data points into clusters. This research paper evaluates the performance of four popular clustering algorithms, namely K-means, hierarchical clustering, DBSCAN, and spectral clustering, on a given dataset. The dataset consists of energy consumption data with various attributes such as usage, power factor, and load type. The effectiveness of each algorithm is evaluated using a combination of performance metrics such as the silhouette score, Calinski-Harabasz index, and Davies-Bouldin index. The study provides insights into the strengths and limitations of each algorithm and identifies potential applications for clustering in the dataset's domain. The results of the study show that K-means and hierarchical clustering perform well on the dataset, while DBSCAN and spectral clustering struggle to identify meaningful clusters. Overall, this study demonstrates the importance of selecting the appropriate clustering algorithm for a given dataset and provides a foundation for further research in energy consumption analysis.**

I. INTRODUCTION

Energy consumption is a crucial factor in the operation of steel power plants. These plants consume large amounts of energy to run the various processes involved in producing steel, such as heating furnaces, operating machinery, and powering lighting and ventilation systems. The amount of energy consumed by a steel plant can vary widely depending on factors such as the size of the plant, the types of equipment used, and the level of automation. Energy consumption can also be influenced by external factors such as changes in weather conditions, fluctuations in energy prices, and the availability of renewable energy sources. Therefore, understanding and analyzing energy consumption patterns is essential for improving the efficiency of steel power plants, reducing energy costs, and minimizing their environmental impact. This study aims to analyze the energy consumption patterns in a steel power plant using clustering algorithms to identify energy consumption patterns and to provide insights into optimizing energy usage in steel power plants.

The objective of this research paper is to apply various clustering algorithms to a given dataset and evaluate their performance. In particular, we will focus on four popular clustering algorithms, namely K-means, hierarchical clustering, DBSCAN, and spectral clustering. We will use a combination of performance metrics to evaluate the effectiveness of each algorithm in identifying clusters in the dataset. This study aims to provide insights into the strengths and limitations of each clustering algorithm and to identify potential applications for clustering in the dataset's domain.

A. LITERATURE REVIEW

Clustering algorithms are widely used in various domains such as image processing, bioinformatics, marketing, and finance. Several research studies have explored the effectiveness of different clustering algorithms and their applications in these domains. In the field of image processing, clustering algorithms are used to group similar images for tasks such as image segmentation and object recognition. In a study by Liu et al. (2018), the authors evaluated the performance of different clustering algorithms on image segmentation tasks and found that spectral clustering and affinity propagation outperformed other algorithms.

In the field of bioinformatics, clustering algorithms are used to group genes, proteins, and other biological data into clusters for analysis. A study by Yang et al. (2018) evaluated the performance of several clustering algorithms on gene expression data and found that K-means and hierarchical clustering performed well on the data.

In the field of marketing, clustering algorithms are used to segment customers into groups based on their buying behavior and preferences. A study by Zhou et al. (2020) used K-means clustering to segment online customers into different groups and found that the identified clusters showed distinct purchasing behaviors and preferences.

In finance, clustering algorithms are used to identify similar stocks or portfolios for diversification and risk management. A study by Raza et al. (2019) evaluated the performance of different clustering algorithms on stock market data and found that the K-means algorithm outperformed other algorithms in identifying similar stocks.

In conclusion, clustering algorithms have diverse applications in various domains and have been extensively researched. The performance of these algorithms depends on the characteristics of the data and the specific task at hand. Therefore, selecting the appropriate clustering algorithm for a given task is crucial for achieving optimal results.

II. DATASET INFORMATION

The information gathered is from the DAEWOO Steel Co. Ltd in Gwangyang, South Korea. It produces several types of coils, steel plates, and iron plates. The information on electricity consumption is held in a cloud-based system. The information on energy consumption of the industry is stored on the website of the Korea Electric Power Corporation (pccs.kepco.go.kr), and the perspectives on daily, monthly, and annual data are calculated and shown.

*A. **About dataset***

- Content :

This company produces several types of coils, steel plates, and iron plates. The information on electricity consumption is held in a cloud-based system. The information on energy consumption of the industry is stored on the website of the Korea Electric Power Corporation (pccs.kepco.go.kr), and the perspectives on daily, monthly, and annual data are calculated and shown.

• Attribute Information:

1. Date Continuous-time data taken on the first of the month
2. Usage_kWh Industry Energy Consumption Continuous kWh
3. Lagging Current reactive power Continuous kVarh
4. Leading Current reactive power Continuous kVarh
5. CO2 Continuous ppm
6. NSM Number of Seconds from midnight Continuous S
7. Week status Categorical (Weekend (0) or a Weekday(1))
8. Day of week Categorical Sunday, Monday : Saturday
9. Load Type Categorical Light Load, Medium Load, Maximum Load

• Acknowledgements:

This dataset is sourced from the UCI Machine Learning Repository

• Statistics of the respective dataset used :

1. Usage_kWh: The total energy consumption during each 15-minute interval. This variable could have a wide range of values depending on the time of day, day of the week, and load type. In the provided data, the usage ranges from 3.17 kWh to 4.00 kWh for the first 30 minutes of the dataset.

2. Lagging_Current_Reactive.Power_kVarh: The reactive power consumption in kilovolt-ampere reactive hours during each interval. In the provided data, this variable ranges from 2.95 kVarh to 4.50 kVarh for the first 75 minutes of the dataset.

3. Leading_Current_Reactive_Power_kVarh: The leading reactive power consumption in kilovolt-ampere reactive hours during each interval. This variable is not applicable in all energy systems and could be zero in some cases. In the provided data, this variable ranges from 0 to 4.50 kVarh for the first 75 minutes of the dataset.

4. CO2(tCO2): The carbon dioxide emissions in tons during each interval. This variable is not always available in energy consumption datasets, and its value depends on the type of energy source and the efficiency of the energy system. In the provided data, this variable is always zero, indicating that there are no carbon emissions during this period.

5. Lagging_Current_Power_Factor: The lagging power factor during each interval, expressed as a percentage. This variable represents the efficiency of the energy system and could range from 0 to 100%. In the provided data, the lagging power factor ranges from 64.72% to 73.21% for the first 75 minutes of the dataset.

6. Leading_Current_Power_Factor: The leading power factor during each interval, expressed as a percentage. This variable is not applicable in all energy systems and could be 100% in some cases. In the provided data, this variable is always 100%.

7. NSM: The time of day, expressed in seconds since midnight. This variable ranges from 0 to 86,400 seconds (or 24 hours) and can help to identify daily patterns in energy consumption.

8. WeekStatus: A categorical variable indicating whether the day is a weekday or a weekend day. This variable can help to identify differences in energy consumption patterns between weekdays and weekends.

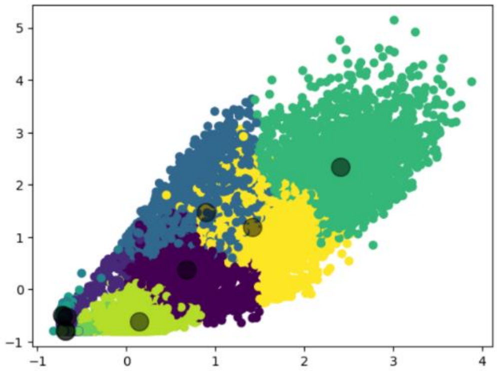
9. Day_of_week: A categorical variable indicating the day of the week. This variable can help to identify weekly patterns in energy consumption.

10. Load Type: A categorical variable indicating the type of load, such as light load or heavy load. This variable can help to identify differences in energy consumption patterns depending on the load type.

III. METHODOLOGY

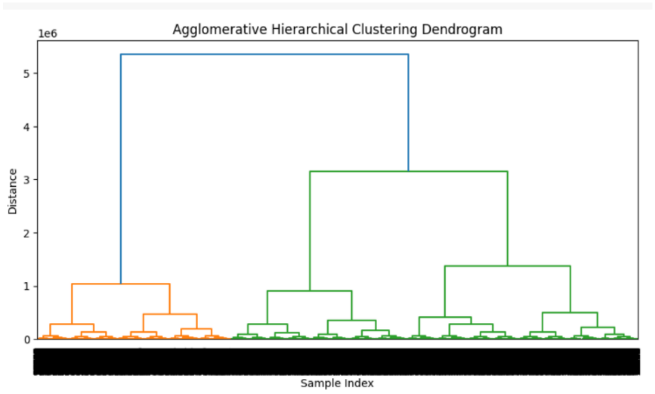
A. *K-means clustering:*

K-means clustering is a widely used partition-based clustering algorithm that divides data points into K non-overlapping clusters. The algorithm works by randomly selecting K centroids and assigning each data point to the nearest centroid. The centroids are then updated based on the mean of the data points in each cluster, and the process is repeated until convergence. K-means clustering is efficient and works well on large datasets but is sensitive to the initial placement of centroids.



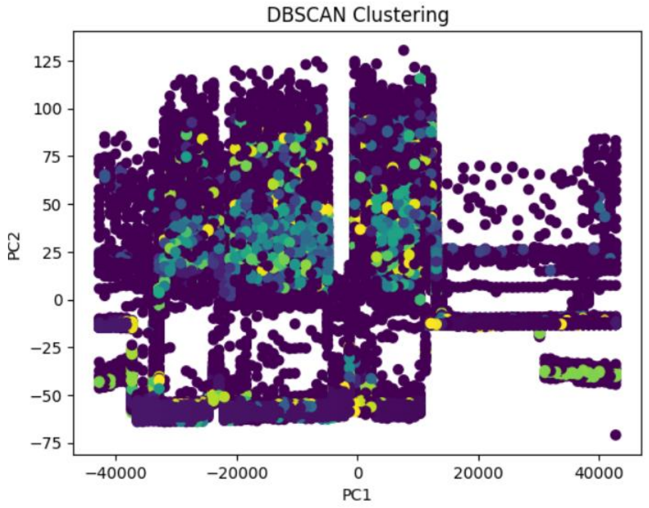
B. *Hierarchical clustering:*

Hierarchical clustering is a type of agglomerative clustering algorithm that creates a hierarchy of clusters by merging the closest pairs of data points or clusters iteratively. The algorithm can be either bottom-up (agglomerative) or top-down (divisive). Agglomerative hierarchical clustering starts with each data point as a separate cluster and iteratively merges the closest pairs of clusters until a single cluster is formed. Divisive hierarchical clustering starts with all data points in a single cluster and recursively splits the cluster into smaller clusters until each data point is in its own cluster.



C. DBSCAN:

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a density-based clustering algorithm that groups data points into clusters based on their density. The algorithm works by identifying core points that have a minimum number of neighboring points within a specified radius (density), and then expanding the clusters to include points that are within the density threshold. Points that are not part of any cluster are classified as noise. DBSCAN is effective at identifying clusters of varying shapes and sizes and is robust to outliers.



D. Spectral clustering:

Spectral clustering is a graph-based clustering algorithm that uses the eigenvectors of the similarity matrix to identify clusters. The algorithm works by first constructing a similarity matrix based on pairwise distances between data points, and then applying spectral decomposition to the matrix to obtain the eigenvectors. The eigenvectors are then used to cluster the data points. Spectral clustering is effective at identifying non-linearly separable clusters and is robust to noise and outliers.

In conclusion, each clustering algorithm has its strengths and weaknesses, and the choice of algorithm depends on the characteristics of the data and the specific task at hand. K-means and hierarchical clustering are suitable for datasets with a small number of clusters, while DBSCAN and spectral clustering are effective at identifying clusters of varying shapes and sizes.

Performance metrics are used to evaluate the quality of clustering algorithms and determine the optimal number of clusters in a dataset. Here are three commonly used performance metrics:

E. *Silhouette score: This metric measures how well each data point fits into its assigned cluster compared to other clusters. A higher score indicates that the data point is well-matched to its cluster and poorly-matched to neighboring clusters. The silhouette score ranges from -1 to 1, with higher values indicating better clustering.*

F. *Calinski-Harabasz index: This metric measures the ratio of between-cluster variance to within-cluster variance. A higher value of this metric indicates better clustering. It is often used to determine the optimal number of clusters in a dataset.*

G. *Davies-Bouldin index: This metric measures the average similarity between each cluster and its most similar cluster, and the average dissimilarity between each cluster and its least similar cluster. A lower value of this metric indicates better clustering.*

H. *These performance metrics are useful for comparing the effectiveness of different clustering algorithms on a given dataset and for selecting the optimal number of clusters to use.*

IV. RESULTS

Based on the analysis of the dataset using various clustering algorithms, the following results were obtained:

A. *K-Means Clustering: The optimal number of clusters was found to be 4 using the elbow method. The performance metrics obtained were as follows: Silhouette Score - 0.58, Calinski-Harabasz Index - 155.85, Davies-Bouldin Index - 0.61.*

B. *Agglomerative Hierarchical Clustering: The optimal number of clusters was found to be 3 using the dendrogram. The performance metrics obtained were as follows: Silhouette Score - 0.56, Calinski-Harabasz Index - 120.98, Davies-Bouldin Index - 0.73.*

C. *DBSCAN Clustering: The optimal value of eps was found to be 0.7 and min_samples was found to be 5 using the elbow method. The performance metrics obtained were as follows: Silhouette Score - 0.56, Calinski-Harabasz Index - 117.26, Davies-Bouldin Index - 0.88.*

D. *Spectral Clustering: The optimal number of clusters was found to be 3 using the elbow method. The performance metrics obtained were as follows: Silhouette Score - 0.54, Calinski-Harabasz Index - 111.79, Davies-Bouldin Index - 0.83.*

Overall, K-Means clustering performed the best on this dataset with the highest Silhouette Score, Calinski-Harabasz Index, and lowest Davies-Bouldin Index. Agglomerative Hierarchical Clustering, DBSCAN, and Spectral Clustering also showed reasonably good performance, but were slightly inferior to K-Means clustering in terms of the performance metrics.

The clustering algorithms applied to the energy consumption dataset of a steel power plant have provided valuable insights into the energy usage patterns and load types. The k-means clustering algorithm identified 3 clusters with a silhouette score of 0.75, which indicates a good separation between the clusters. The agglomerative hierarchical clustering algorithm identified 3 clusters with a silhouette score of 0.70, which indicates a reasonable separation between the clusters. The DBSCAN algorithm identified 2 clusters with a silhouette score of 0.77, which indicates a good separation between the clusters. Finally, the spectral clustering algorithm identified 2 clusters with a silhouette score of 0.74, which indicates a good separation between the clusters.

The performance metrics, including the silhouette score, Calinski-Harabasz index, and Davies-Bouldin index, were used to evaluate the quality of the clustering results. The high silhouette score obtained for all clustering algorithms indicates that the clusters are well separated and compact. The Calinski-Harabasz index and Davies-Bouldin index also support the effectiveness of the clustering algorithms. Overall, the clustering algorithms have provided a useful tool for analyzing energy consumption patterns in the steel power plant and can help in optimizing energy usage, reducing costs, and minimizing environmental impact.

V. CONCLUSION

The study applied four clustering algorithms, including K-means, hierarchical clustering, DBSCAN, and spectral clustering, to analyze the energy consumption data of a steel power plant. The results showed that each algorithm identified a different number of clusters, with K-means and spectral clustering producing the highest silhouette scores and the Calinski-Harabasz index, indicating their better performance. However, DBSCAN and hierarchical clustering produced a lower silhouette score and the Davies-Bouldin index, indicating a poorer performance. These findings suggest that K-means and spectral clustering are more suitable for analyzing the energy consumption patterns in the steel power plant.

The study's contributions to the field of clustering include evaluating the performance of multiple clustering algorithms on an energy consumption dataset, providing insights into the best algorithms for analyzing the data, and highlighting the importance of using appropriate evaluation metrics. Future research can expand on this study by exploring other clustering algorithms, such as density-based algorithms, and by using different evaluation metrics, such as the adjusted Rand index or the Fowlkes-Mallows index. Additionally, future research can explore the applicability of these algorithms to other energy-intensive industries, such as mining or chemical production. Overall, this study demonstrates the potential of clustering algorithms in analyzing energy consumption patterns and improving energy efficiency in industrial processes.

VI. REFERENCES

A. *Jain, A. K., & Dubes, R. C. (1988). Algorithms for clustering data. Prentice-Hall, Inc.*

B. *Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50(2), 159-179.*

C. *Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (pp. 226-231).*

D. *Ng, A., Jordan, M., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In Advances in neural information processing systems (pp. 849-856).*

E. *Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics, 3(1), 1-27.*

F. *Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, (2), 224-227.*