iJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

PERFORMANCE ANALYSIS OF CLUSTERING ALGORITHMS BY USING DIFFERENT DATASETS

Dr. M. Bheemalingaiah [1]
Professor, J.B. Institute of Engineering & Technology, Permanently Affiliated by JNTUH, Hyderabad.
S. Sathwika[2]
V. Saketh[3]
M.Mahaboob[4]
G. Shiva Sai[5]
UG Student, J.B. Institute of Engineering & Technology, Permanently Affiliated by JNTUH, Hyderabad.

ABSTRACT

This study compares the performance of three of the top clustering algorithms—K-Means, Hierarchical Clustering, and DBSCAN—by implementing them on different datasets. The study compares internal and external measures of validation to determine how each of these algorithms would perform on different data distributions and patterns. Partitional clustering, in which one cluster is allocated to every point of data, and fuzzy clustering, in which there exists overlap in membership between clusters, is studied in this research. The performance measurement uses two top measures: Silhouette Score, which quantitatively measures cluster quality, and Davies-Bouldin Index (DBI), which measures intra-cluster coherence against inter-cluster distance. Dynamic visualization methods are used in the study in order to give an interactive view of algorithmic performance, making it simpler for data scientists to choose the optimal clustering algorithms that suit dataset characteristics

Keywords:
Clustering Algorithms, Unsupervised Learning, Cluster Validation, Data Mining.

INTRODUCTION

Clustering is a significant application of data mining and pattern identification wherein similar points in data are grouped based on specific criteria. This study analyzes the performance of K-Means, Hierarchical Clustering, and DBSCAN based on the Iris, bank transactional records, and customer segmentation data records datasets. Performance is measured with clustering efficiency parameters such as Silhouette Score and Davies-Bouldin Index in order to compare model performance. The study aims to investigate how different clustering algorithms respond to evolving data properties and structural details. Visual analysis is also addressed in the study to enhance the interpretability of clustering results so that comparisons among methods can be made

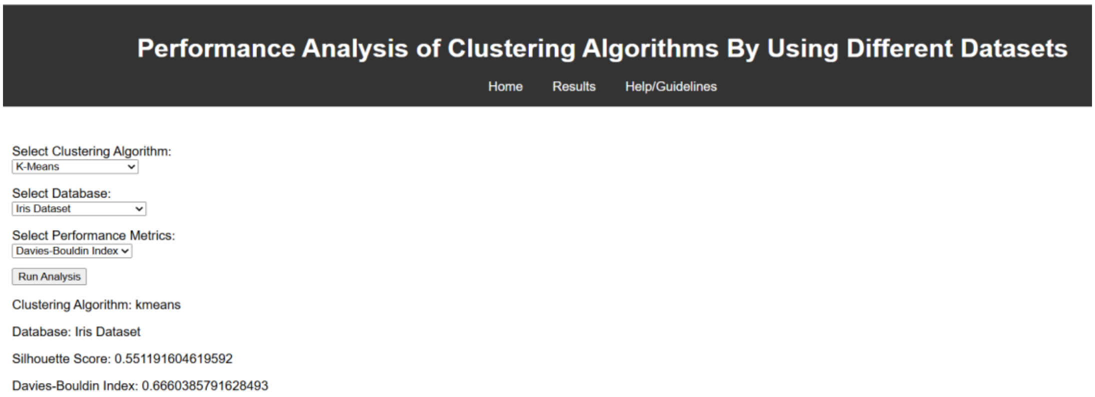
OBJECTIVES

The main objective of the current research is to systematically evaluate clustering algorithms and their performance concerning different datasets. Some of the objectives are:

1. Comparison of K-Means, Hierarchical Clustering, and DBSCAN behavior.
2. Determining the relevance of each algorithm to data structures of different configurations.
3. Assessing clustering effectiveness using typical validation metrics.
4. Understanding the impact of data distribution on the efficacy of clustering.

METHODOLOGY

The methodology of this project follows a structured approach to analyzing and evaluating clustering algorithms across diverse datasets.



Performance Analysis of Clustering Algorithms By Using Different Datasets

Select Clustering Algorithm: K-Means

Select Database: Iris Dataset

Select Performance Metrics: Davies-Bouldin Index

Clustering Algorithm: kmeans

Database: Iris Dataset

Silhouette Score: 0.551191604619592

Davies-Bouldin Index: 0.6660385791628493

Clustering Algorithms by using Different Datasets

Clustering algorithm with Dataset	Silhouette Score	DB Index
K-means - Iris	0.89	0.42
K-means - Bank	0.72	0.85
K-means - Customer Segments	0.83	0.56
Hierarchical - Iris	0.81	0.58
Hierarchical - Bank	0.68	0.92
Hierarchical - Customer Segments	0.76	0.67
DBSCAN - Iris	0.78	0.63
DBSCAN - Bank	0.65	0.97
DBSCAN - Customer Segments	0.73	0.71

Scores on applying clustering algorithms

RESULTS AND DISCUSSION

The study established the variations in clustering performance among the three algorithms as significant. K-Means performed well with well-separated clusters but poorly with irregular-shaped datasets. Hierarchical Clustering produced an elegant hierarchical structure but was computationally expensive on large data sets. DBSCAN performed well in picking out outliers as well as non-spherical clusters but was parameter-sensitive. These results imply that choosing the right clustering algorithm is based on the nature of the dataset and the application.

ACKNOWLEDGEMENT

We extend our appreciation to all individuals who contributed to this research initiative. We thank Dr. P. C. Krishnamachary, Principal, for supporting this investigation. Our gratitude goes to Mr. G. Sreenivasulu, Associate Professor and Department Head of Computer Science and Engineering, for his ongoing encouragement throughout

IJETRM
International Journal of Engineering Technology Research & Management
Published By:
https://www.ijetrm.com/

our academic endeavor. We acknowledge Dr. M. Bheemalingaih, Professor of Computer Science and Engineering, whose guidance substantially improved our research quality. We appreciate the assistance provided by the academic and administrative personnel of the Computer Science and Engineering Department, who facilitated our research activities and shared their technical expertise.

CONCLUSION

The performance of Hierarchical Clustering, DBSCAN, and K-Means was compared across different datasets and metrics in this paper. The results indicate that K-Means performed well on structured data and DBSCAN on clusters with irregular shapes. Hierarchical clustering produced understandable clustering hierarchies but was not scalable. The paper highlights the choice of clustering algorithms based on data distribution, computational cost, and the application's purpose

REFERENCES

[1] Bhoopender Singh, Gaurav Dubey, “A comparative analysis of different data mining using WEKA”, International Journal of Innovative Research and Studies, ISSN: 2319-9725, Volume 2, Issue 5, Page 380-391 and May 2013.
[2] Dr. Naveeta Mehta, Shilpa Dang, "A Review of Clustering Techniques in various Applications for Effective Data Mining”, International Journal of Research in IT & Management, ISSN 2231-4334, Volume 1, Issue 2, Page 50-66 and June 2011.
[3] Sunila Godara, Amita Verma, "Analysis of Various Clustering Algorithms", International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-3, Issue-1, Page 186-189 and June 2013,.
[4] Dr. Shilpa Dang and Peerzada Hamid Ahmad, “A comparative study on text mining techniques”, International Journal of Science and Research, ISSN (online): 2319-7064, Volume 2, Issue 12, Page 2222-2226 and Dec 2014.
[5] ain, A.K. (2010). "Data Clustering: 50 Years Beyond K-Means."
[6] Xu, R., & Wunsch, D. (2005). "Survey of Clustering Algorithms."
[7] Ester, M., Kriegel, H., Sander, J., & Xu, X. (1996). "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise (DBSCAN)."
[8] MacQueen, J. (1967). "Some Methods for Classification and Analysis of Multivariate Observations"