An Evaluation of K-means Clustering Algorithm for Pattern Recognition

H. Vasanth Kumar

PG Scholar, Dept. of Computer Science Sri Venkateswara University, Tirupati

***Abstract*— *Pattern recognition plays a crucial role in various domains, including biology, medicine, and data analysis. Clustering algorithms, such as k-means, are commonly used for pattern recognition tasks. In this research paper, we evaluate the effectiveness of the k-means clustering algorithm on the well-known Iris dataset for identifying distinct patterns and grouping similar instances together. It highlights the evaluation of internal validation metrics, the comparison of cluster assignments with true labels, and the characteristics of the formed clusters.*

I. INTRODUCTION

Machine Learning algorithms are iterative processes or sets of methods that assist a model in adapting to data with a specific objective. Clustering is a machine learning technique that involves grouping similar data points into clusters or subgroups based on the similarity of their features [1][2]. The goal of clustering is to identify natural patterns or structures within the data, without any prior knowledge of the underlying categories or labels.

Clustering algorithms usually work by defining a distance metric or similarity measure between the data points and then grouping them into clusters based on their proximity to each other in the feature space. Grouping can be utilized for different applications, like client division, irregularity recognition, and picture division [3[4]. It is a valuable device for exploratory information investigation and can give experiences into the basic examples and designs inside the information.

II. CLUSTERING

Clustering is an unsupervised machine learning technique that aims to group similar data points together based on their characteristics or proximity. It helps identify patterns and structures within a dataset without prior knowledge of the class labels [5]. In other words, your dataset lacks a label, a target variable that would be connected to the patterns discovered by the explanatory variables. Clustering is an AI method that includes gathering comparable data of interest into groups or subgroups in light of the comparability of their elements. The objective of bunching is to distinguish normal examples or designs inside the information, with next to no earlier information on the basic classes or names.

Bunching calculations typically work by characterizing a distance metric or comparability measure between the data of interest and afterward gathering them into groups in light of their nearness to one another in the element space [6][8] There are different sorts of bunching calculations, including progressive grouping, k-implies bunching, and thickness based bunching, each with its own assets and shortcomings.

III. METHODOLOGY

There are various types of clustering algorithms, including hierarchical clustering, k-means clustering, and density-based clustering, each with its own strengths and weaknesses. In this paper, we used K-means clustering algorithm for Iris data.

3.1 K-means Clustering Algorithm

The K-means algorithm is a widely used method that is applicable for clustering data based on quantitative variables. The method is based on an iterative algorithm in which the process is initiated by providing a fixed set of centroids [7]. Each data point to be clustered is then assigned to its closest centroid using a squared Euclidian distance measure. To assign a point to a cluster, the goal is to minimize the sum of average pair-wise distance within-cluster dissimilarity. The centroids are then updated by computing the average of all the points assigned to each cluster. These steps are iterated until the assignment of the data points to each centroid does not change significantly. This method is efficient to analyze large datasets however, its application is limited to clustering based on the quantitative variable as it utilizes the Euclidian distance as the dissimilarity matrix[6] .

Centroid: In k-means clustering, a centroid represents the center of a cluster. It is a multidimensional point that is determined by taking the mean of all the data points within a cluster. The centroid serves as a representative point for the cluster.

Euclidean Distance: Euclidean distance is a measure of the straight-line distance between two points in a multidimensional space. It is commonly used to calculate the similarity or dissimilarity between data points in k-means clustering. The closer the points, the smaller the distance.

3.2 Algorithm

1. Initialization: Initialization is the process of selecting the initial positions of the centroids before starting the k-means algorithm. Different initialization methods can be used, such as randomly selecting data points as centroids or using more sophisticated techniques like k-means++.

2. Assignment Step: In the assignment step of the k-means algorithm, each data point is assigned to the nearest centroid based on the Euclidean distance. The goal is to minimize the distance between each data point and its assigned centroid.

3. Update Step: In the update step, the centroids are recalculated based on the current cluster assignments. The new centroids are computed by taking the mean of all the data points within each cluster.

4. Convergence: The k-means algorithm iteratively performs the assignment and update steps until convergence is achieved. Convergence occurs when the centroids no longer change significantly between iterations or when a predefined number of iterations is reached.

IV. EXPERIMENTAL RESULTS

The study was conducted using the Python programming language, leveraging the Scikit-learn library for data representation, manipulation, and visualization. We apply the k-means algorithm to the Iris dataset was obtained from the UCI ML repository with varying values of k, the number of clusters [9]. The Iris dataset consists of measurements of four features—sepal length, sepal width, petal length, and petal width—from three different species of Iris flowers: Setosa, Versicolor, and Virginica. It has been widely used as a benchmark dataset for evaluating clustering algorithms. The dataset contains 150 instances, with 50 instances per species. The experimental results are shown in the figure-1.

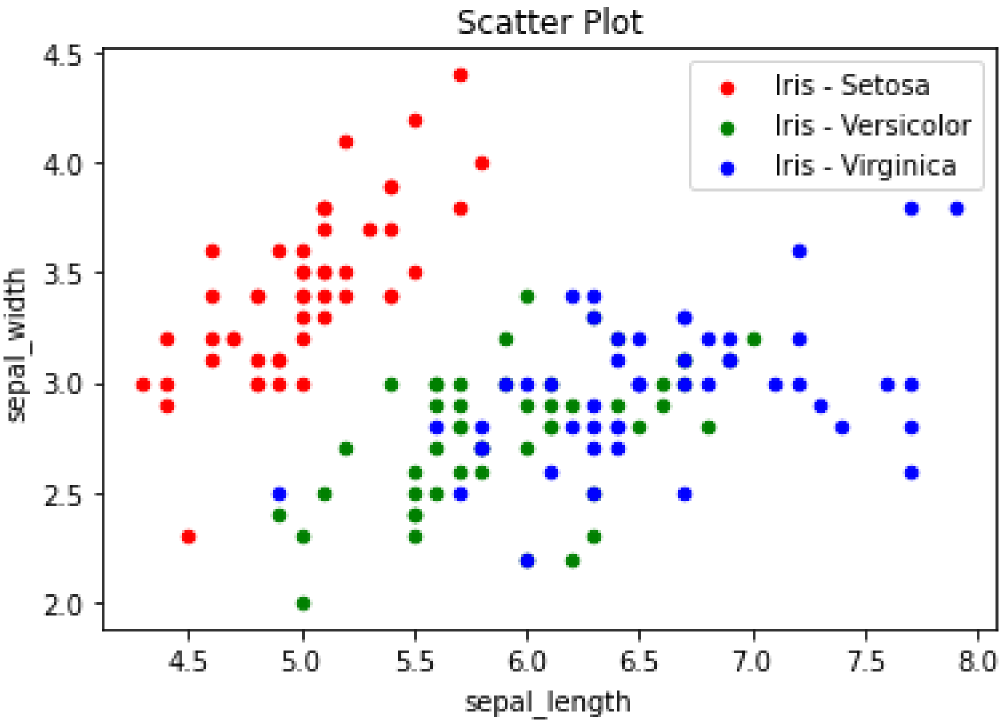

Figure-1: K-means Results

Our experimental results show that k-means clustering can effectively group Iris instances into distinct clusters. By analyzing the internal validation metrics, we identify the optimal value of k that produces the most compact and well-separated clusters. Furthermore, the external evaluation metrics indicate that the clusters generated by k-means align well with the true labels of the Iris dataset.

V. CLUSTER CHARACTERISTICS

We analyzed the characteristics of each cluster formed by k-means.

Cluster 1 predominantly consisted of Iris Setosa instances, demonstrating the distinct separation of this species from the others.

Cluster 2 primarily included Iris Versicolor instances, indicating the effective grouping of this species.

Cluster 3 was predominantly composed of Iris Virginica instances, further highlighting the successful separation of this species.

These findings align well with the ground truth labels, affirming the accuracy of the clustering results.

VI. CONCLUSION

This research paper presents an evaluation of the k-means clustering algorithm on the Iris dataset for pattern recognition. The results demonstrate the effectiveness of k-means in identifying distinct patterns and grouping similar instances together. This work contributes to the field of pattern recognition and provides valuable insights for future research and applications in various domains. We discuss the implications of our findings and highlight the strengths and limitations of the k-means algorithm for pattern recognition tasks using the Iris dataset. The insights gained from this research can guide the selection of suitable parameters and provide a benchmark for evaluating other clustering algorithms on the Iris dataset.

REFERENCES

[1] D. Hand, H. Mannila, P. Smyth.: Principles of Data Mining. The MIT Press. (2001)

[2] G. Ravi Kumar, S. Rahamat Basha, Surya Bhupal Rao, “A Summarization on Text Mining Techniques for Information Extracting from Applications and Issues”, Journal of Mechanics of Continua and Mathematical Sciences, Special Issue, No.-5, January (2020) pp 324-332, ISSN:0973-8975.

[3] G. Ravi Kumar, K. Tirupathaiah and B. Krishna Reddy, “Client Churn Prediction of Banking and fund industry utilizing Machine Learning Techniques”, International Journal of Computer Sciences and Engineering, Volume-7, Issue-6, e-ISSN: 2347 — 2693, PP: 871-875, June 2019

[4] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques.2nd ed. San Francisco: Morgan Kaufmann, 2005.

[5] J. Han and M. Kamber,” Data Mining concepts and Techniques”, the Morgan Kaufmann series in Data Management Systems, 2nd ed. San Mateo, CA; Morgan Kaufmann, 2006.

[6] N. Michael, “Artificial Intelligence - A Guide to Intelligent Systems”,2nd edition, Addison Wesley, 2005.

[7] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Reading, MA: Addison-Wesley, 2005.

[8] Surya Bhupal Rao, S.Rahamat Basha, G. Ravi Kumar, “A Comparative approach of Text Mining: Classification, Clustering and Extraction Techniques”, Journal of Mechanics of Continua and Mathematical Sciences, Special Issue, No.-5, January (2020) pp 120-131, ISSN: 0973-8975.

[9] UCI machine learning repository. http://archive.ics.uci. edu/ml/.