Cluster Analysis

张振
Id: 1024040913

***Abstract*—In the era of big data, cluster analysis is of great importance for dealing with complex data. The purpose of this study is to deeply understand and compare the partitioning-based, hierarchical-based and density-based cluster analysis methods and their classical algorithms. Through experiments on the dataset obtained from GitHub, the distances between each data point are calculated, and a distance matrix is generated to determine the parameters of the DBSCAN algorithm. The Silhouette Coefficient, Calinski-Harabasz Index and Davies-Bouldin Index are used as evaluation metrics to compare the performance of the three algorithms: K-Means, Hierarchical Clustering and DBSCAN. The results show that the performance of the K-Means and Hierarchical Clustering algorithms is similar on this dataset, while the DBSCAN algorithm has a poorer effect. It is speculated that the reason is that the distribution of the dataset is the spherical distribution that the former two are good at, while the DBSCAN is suitable for datasets with complex shapes. This study provides a valuable reference for the selection and application of cluster analysis algorithms.**

***Index Terms*—Clustering, Similarity measures, Data mining.**

I. Introduction

IN different fields such as engineering, science and technology, humanities, medical science, and our daily life, objects need to be grouped for various purposes. For example, people suffering from a particular disease have some common symptoms and will be placed in a group tagged with a certain label, usually the name of the disease. Obviously, those who do not have these symptoms (and thus do not have the disease) will not be placed in that group. The patients grouped for that disease will be treated accordingly, while those not belonging to that group should be handled differently. Whenever we find a labeled object, we will put it into the group with the same label. Since the labels are given in advance, this is a rather simple task. However, in many cases, such labeling information is not provided in advance, and we can only group objects based on certain similarities. Both of these situations represent a wide variety of problems that occur in data analysis. Generally speaking, these cases are dealt with within the scope of classification. Precisely, the first case, where the class (label) of an object is given in advance, is called supervised classification; while the other case, where the class label is not tagged to an object in advance, is called unsupervised classification[1].

In the current era of big data, a large amount of data can be obtained and used for analysis to obtain useful information. However, the data is complex and disorderly, and researchers need to divide it to some extent and analyze similar data together. This is what is called clustering.

Clustering is to divide a data set into different classes according to a specific criterion (such as the distance between data), so that the data objects within the same class are as similar as possible, and at the same time, the data objects not in the same class are as different as possible. The commonly used basic distance calculation method is the Euclidean distance. Specifically, we can understand that after clustering, the data of the same class are gathered together as much as possible, and the data of different classes are separated as much as possible. Clustering technology is booming, and the fields it is applied to include data mining, statistics, machine learning, spatial database technology, biology, and marketing[2]. Various clustering methods are constantly being proposed and improved, and different methods are suitable for different types of data. Therefore, the comparison of various clustering methods and clustering effects is a problem worthy of research.

Currently, there are a large number of clustering algorithms[3]. For specific applications, the choice of clustering algorithm depends on the type of data and the purpose of clustering. If clustering analysis is used as a descriptive or exploratory tool, multiple algorithms can be tried on the same data to discover the possible results that the data may reveal. The main clustering algorithms can be divided into the following categories: partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods.

The research on clustering problems is not limited to the above hard clustering, that is, each data can only be classified into one class. Fuzzy clustering is also a widely studied branch in cluster analysis. Fuzzy clustering determines the degree to which each data belongs to each class through the membership function, instead of rigidly classifying a data object into a certain class. At present, many algorithms about fuzzy clustering have been proposed, such as the famous FCM algorithm.

II. Object

The purpose of this experiment is to further understand the principles of partitioning-based, hierarchical-based, density-based clustering analysis methods and classical clustering analysis algorithms, and to implement specific algorithms through programming to conduct experiments on the dataset and compare the advantages and disadvantages among them.

III. Experiment

A. Clustering Algorithm

• ***K-Means***[4]: K-means is one of the more classic clustering algorithms in the partitioning method. Due to the high efficiency of this algorithm, it is widely used when clustering large-scale data. Currently, many algorithms are extended and

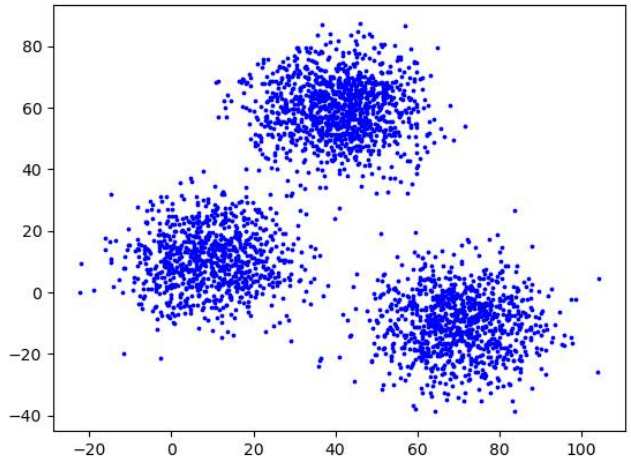


Fig. 1: Data Distribution.

improved around this algorithm. The goal of the K-means algorithm is, with k as a parameter, to divide n objects into k clusters so that there is a high similarity within the cluster and a low similarity between the clusters.

The processing process of the K-means algorithm is as follows: First, randomly select k objects, and each object initially represents the average or center of a cluster; for each of the remaining objects, assign it to the nearest cluster according to its distance from each cluster center; then recalculate the average of each cluster. This process is repeated continuously until the criterion function converges.

The algorithm flow of the K-means clustering algorithm is as follows:

Input: A database containing n objects and the number k of clusters;

Output: k clusters to minimize the sum of squared errors criterion.

Steps:

(1) Arbitrarily select k objects as the initial cluster centers;

(2) repeat;

(3) Based on the average value of the objects in the cluster, assign each object (again) to the most similar cluster;

(4) Update the average value of the cluster, that is, calculate the average value of the objects in each cluster;

(5) until no further changes occur.

• ***Hierarchical Clustering***[5]: According to whether the order of hierarchical decomposition is bottom-up or top-down, hierarchical clustering algorithms are divided into agglomerative hierarchical clustering algorithms and divisive hierarchical clustering algorithms.

The strategy of agglomerative hierarchical clustering is to first consider each object as a cluster, and then merge these atomic clusters into increasingly larger clusters until all objects are in one cluster or a certain termination condition is met. The vast majority of hierarchical clustering belongs to agglomerative hierarchical clustering, and they only differ in the definition of the similarity between clusters. Four widely used methods for measuring the distance between clusters are: minimum distance, maximum distance, average distance, and mean distance.

Here, take the process of the agglomerative hierarchical clustering algorithm with the minimum distance as an example:

(1) Consider each object as a class and calculate the minimum distance between each pair.

(2) Merge the two classes with the minimum distance into a new class.

(3) Recalculate the distances between the new class and all other classes.

(4) Repeat steps (2) and (3) until all classes are finally merged into one class.

• ***DBSCAN***[6]: This algorithm is based on the density of data points. If the density of data points in a certain area exceeds a certain threshold, these points will be grouped into a cluster. Data points that are density-connected form a cluster, and data points in low-density areas are regarded as noise points.

The implementation steps of the algorithm are as follows:

(1)Select an unvisited data point p as the starting point.

(2)Calculate the number of data points within the Eps neighborhood of point p. If the number is greater than or equal to minsample, mark p as a core point and create a new cluster.

(3)Add point p to the current cluster and add all unvisited data points within the Eps neighborhood of p to the current cluster.

(4)For each data point q in the current cluster, perform the following operations:

a) If q is a core point, add all unvisited data points within the Eps neighborhood of q to the current cluster.

b) If q is not a core point but is within the Eps neighborhood of other clusters, mark q as a boundary point and add it to the current cluster.

(5)When no more data points can be added to the current cluster, the current cluster is considered a complete cluster.

(6)Select the next unvisited data point as the starting point and repeat steps 2 to 5 until all data points have been visited.

(7)Mark the remaining unassigned data points as noise points.

B. Evaluation Metrics

Evaluating the effectiveness of clustering results, namely clustering evaluation or validation, is crucial for the success of clustering applications. It can ensure that the clustering algorithm identifies meaningful clusters in the data and can also be used to determine which clustering algorithm is the most suitable for a specific dataset and task, as well as to tune the hyperparameters of these algorithms (for example, the number of clusters in k-means, or the density parameter in DBSCAN). Although supervised learning techniques have clear performance metrics such as accuracy, precision, and recall, evaluating clustering algorithms is more challenging: clustering is an unsupervised learning method, so there are no ground truth labels against which the clustering results can be compared.

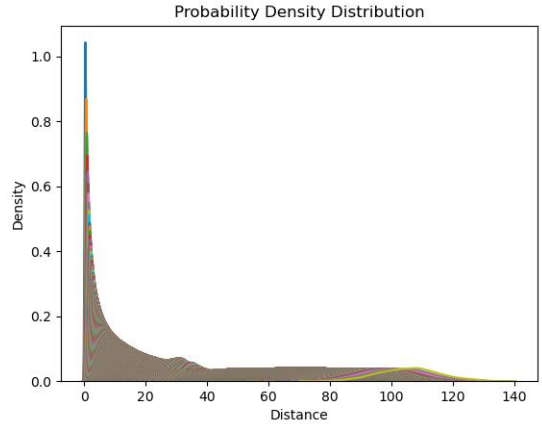


Fig. 2: Probability density curve.

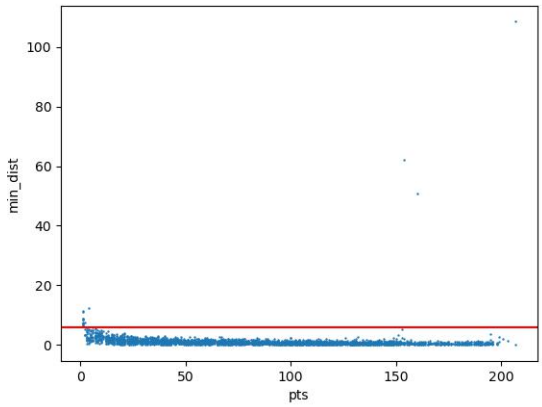


Fig. 3: The functional relationship between ρi and δi of each point.

Generally speaking, we use two types of clustering evaluation metrics (or measures):

Internal: No ground truth is required to evaluate the quality of clusters. They are entirely based on the data and the clustering results.

External: The clustering results are compared with the ground truth labels. (Since the ground truth labels are not available in the data, they need to be introduced from the outside.)

Since the dataset used in this paper does not have ground truth labels for comparison, the following three internal metric methods are adopted as the evaluation indicators for the clustering algorithm:

• ***Silhouette Coefficient(SC)***[7]: The silhouette coefficient (or score) measures the degree of separation between clusters by comparing the similarity of each object to its own cluster with the similarity to the objects in other clusters.

The value range of the Silhouette Coefficient is [-1, 1]: Close to 1: The data points are closely clustered within their own cluster and well separated from the nearest neighboring cluster. Close to 0: The data points are located on the boundary between two clusters, and the clustering effect is average. Close to -1: The data points may be wrongly assigned to the wrong cluster.

• ***Calinski-Harabasz Index(CHI)***[8]: Since the essence of the Calinski-Harabasz index is the ratio of the inter-cluster distance to the intra-cluster distance, and the overall calculation process is similar to the way of variance calculation, it is also called the variance ratio criterion.The advantage of the Calinski-Harabasz Index(CHI) is that it has a fast computing speed and is suitable for large-scale datasets. However, when the value of the number of clusters k is relatively large, the CH Index may lose its significance because at this time the within-cluster distances may become very small, resulting in a relatively large CH value.

• ***Davies-Bouldin Index(DBI)***[9]: The Davies-Bouldin Index (DBI) is an internal evaluation metric based on intra-cluster distances and inter-cluster distances, which is used to measure the compactness and separation of clusters. The calculation principle of the DB value is to calculate the sum of the average intra-cluster distances between any two clusters divided by the distance between the centers of these two clusters, and then find the maximum value. The advantage of the DB index is that it can handle clusters of different sizes and densities well and has strong robustness against noise and outliers. However, when the value of the number of clusters k is relatively large, the computational complexity of the DB index may be relatively high.

C. Dataset

Data sets for clustering can be downloaded from GitHub, address to https://github.com/mubaris/friendly-fortnight/blob/master/xclara.csv.The data distribution is shown in Fig.1.

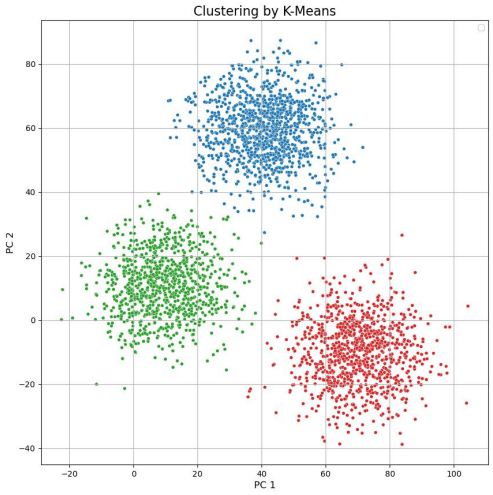


Fig. 4: K-Means clustering results .

IV. EVALUATE

It can be seen from the distribution of the dataset that the data can be roughly divided into three categories. Therefore,

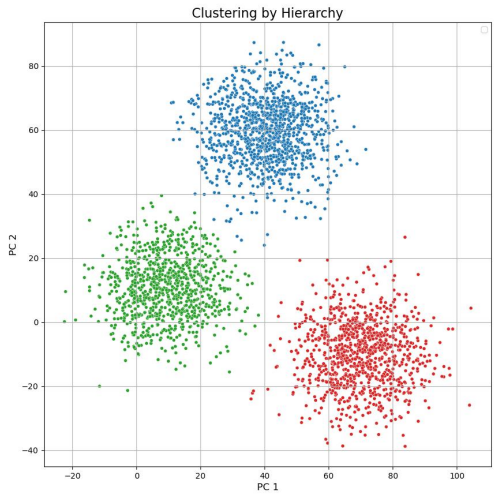
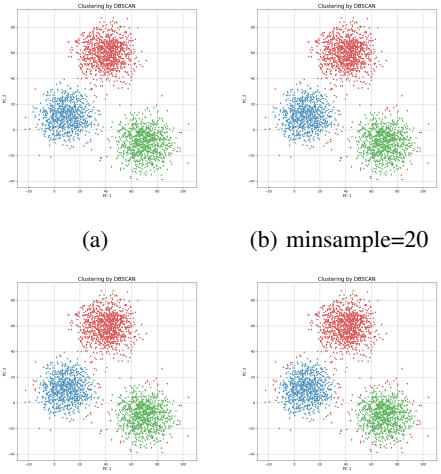


Fig. 5: Hierarchical Clustering clustering results.



(a) (b) minsample=20

(c) minsample=30 (d) minsample=40

Fig. 6: DBSCAN clustering results.

in this paper, the category parameter of the K-Means and hierarchical classification algorithms is set to 3.

The following is the parameter analysis of the DBSCAN algorithm

•Determine the value of the eps radius: Calculate the distances of each data point, generate a distance matrix, and then sort the distance matrix in ascending order row by row. In this way, each row of the matrix represents the sorted distances from the corresponding data point to all other points, and each column is the set of the i-th closest distance values to each data point. Then, plot the probability density distribution curve of the distance values column by column.

The result is shown in the figure. As can be seen from the Fig.2, the density at a distance of 6 has become very small. Those points that account for a very small proportion and are significantly farther away from other points are very likely to be noise. Therefore, Eps can be selected as 6.

•Determine the number of points within the minsample neighborhood: Calculate the local density value of each data point (that is, the number of points within the neighborhood range where Eps is 6) ρi, and then calculate the nearest distance of each point to the points with higher density. If it is a point with the highest density, take the distance to the farthest point in the dataset as δi. Then plot the graph with the local density value as the abscissa and the nearest distance to the points with higher density as the ordinate.

The result is shown in the Fig.3. The data points with δi < 6 (that is, below the red line in the figure) are all data points whose distances to the points with higher density are less than Eps. Select the minsample value according to the figure. The data points with ρi $>=$ minsample are all core points. The data points with ρi $<$ minsample and δi $<= 6$ may be boundary points or noise points, and the data points with ρi $<$ minsample and δi > 6 are all noise points. Therefore, the range of minsample taken in this paper is [10 - 40].

TABLE I: Clustering results

	SC	CHI	DBI
K-Means	0.6946	10826.6005	0.4205
Hierarchical Clustering	0.6942	10793.0639	0.4196
DBSCAN	0.6587	6757.8543	1.5089

Fig.4, Fig.5 and Fig.6 respectively show the clustering situations of the dataset by three algorithms, namely K-Means, Hierarchical Clustering and DBSCAN. In the DBSCAN algorithm, the parameter "eps" is set to 6, and the parameter "minsample" is set to 10, 20, 30 and 40 respectively. It can be seen that the clustering effect is the best when "minsample" is set to 10.

It can be seen from the table I, for the three indicators, the performance of the K-Means algorithm and the Hierarchical Clustering algorithm is almost the same. However, the classification results of the density-based DBSCAN algorithm are significantly inferior to the above two algorithms. This may be because the distribution of the dataset used in this paper is the spherical distribution that the above two algorithms are good at, while the DBSCAN algorithm can be applied to datasets with various complex shapes.

V. Conclusion

Big data can be said to be a very hot research topic nowadays and is a relatively core project in many research laboratories and technology companies. Especially in the era of rapid development of the Internet today, information means resources, and if you can master the information, you can seize the opportunity. As a key part of big data analysis, clustering provides analysts with extraordinary data preprocessing, enabling us to discover the logical relationships and situations hidden beneath the data.

This research focuses on cluster analysis methods and deeply explores the principles of partitioning-based, hierarchical-based and density-based clustering algorithms. It also conducts experimental analysis on a specific dataset

through programming. In the experiment, the dataset is pre-processed, the distance matrix is calculated to determine the parameters of the DBSCAN algorithm, and multiple internal evaluation metrics are used to comprehensively evaluate the performance of the algorithms. The results show that the performance of the K-Means and Hierarchical Clustering algorithms is similar on this dataset, while the DBSCAN algorithm has a poorer effect, which may be related to the spherical distribution characteristics of the dataset. Although the DBSCAN algorithm is suitable for datasets with complex shapes, it does not show its advantages here.

The results of this research provide practical references for the understanding, selection and application of clustering algorithms, which is helpful for selecting more appropriate clustering methods according to the characteristics of data in big data processing. Future research can further explore the impact of different types of datasets on the performance of algorithms and optimize the setting of algorithm parameters to improve the clustering effect.

REFERENCES

[1] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, and C.-T. Lin, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, 2017.

[2] G. J. Oyewole and G. A. Thopil, "Data clustering: application and trends," *Artificial Intelligence Review*, vol. 56, no. 7, pp. 6439–6475, 2023.

[3] L. Rokach and O. Maimon, "Clustering methods," *Data mining and knowledge discovery handbook*, pp. 321–352, 2005.

[4] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, Oakland, CA, USA, 1967.

[5] F. Murtagh, "A survey of recent advances in hierarchical clustering algorithms," *The computer journal*, vol. 26, no. 4, pp. 354–359, 1983.

[6] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *kdd*, vol. 96, pp. 226–231, 1996.

[7] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[8] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979.

[9] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of internal clustering validation measures," in *2010 IEEE international conference on data mining*, pp. 911–916, Ieee, 2010.