GDD-K-Means Text Clustering Algorithm Based on Grid Filtering Distance and Density of Outliers

Yao Wang, Bin Wang*, Xiuwen Qi
School of Mathematics and Data Science, Changji College, Changji, Xinjiang, China
**Corresponding Author.*

Abstract: In the era of big data, fully mining and utilizing the value of big data in line with the requirements of big data strategy plays a significant role in social development. Clustering algorithm can effectively partition unlabeled data sets through unsupervised learning process, and traditional K-Means algorithm is still the most widely used algorithm at present. By studying and learning various improved algorithms of traditional K-Means clustering algorithm, this paper has optimized the problems such as unsatisfactory clustering results caused by outliers and disadvantages of initial center point affecting initial partitioning. Good results have been obtained. Firstly, the grid filtering and LOF detection method of weighing distance and density are used to remove outliers. Then, the randomness of initial center selection is better eliminated by combining the "max-min principle" with the strategy of maximum weight, and the number of clusters is determined according to the BWP index. Experimental results have shown that compared with the currently popular clustering algorithms, the proposed GDD-K-Means clustering algorithm has achieved better results in different data sets, and the accuracy and F-number and other evaluation indexes are improved to a certain extent, and the calculation time complexity is effectively reduced.

Keywords: Data Mining; K-Means Algorithm; Grid Filtering Outlier; Number of Class Centers

1. Introduction

With the rapid development of Internet technology, the world has entered the era of big data. It has become an indisputable fact that the amount of data on the Internet is growing at an alarming rate, the complexity and diversity of big data and the particularity of data mining technology application have also brought severe challenges to data mining technology. In such an environment, it becomes particularly important to realize the effective mining and application of massive data and grasp the mystery and mystery behind the data. In real life, there is a large amount of unlabeled data, and how to deeply mine unsupervised data through big data strategy to obtain great value has become an important topic [1].

The clustering algorithm deals with the unsupervised data set well and mining the data effectively, which makes the discrete information accumulate into valuable information clusters. The clustering generated by cluster analysis is the process of grouping data objects, where objects within the same cluster are similar to each other but significantly different from objects in other clusters. These clusters are collections of data objects that have high intrinsic consistency in features, but exhibit significant differences between different clusters. [2]. As a classical clustering algorithm, the K-Means algorithm is widely popular for its concise ideas and easy implementation, a wide range of application scenarios, easy to operate, and strong compatibility. It is still a valuable research direction for big data analysis and processing. However, K-Means algorithm also has limitations, such as the algorithm needs to set the number of categories in advance, and it is difficult for users to give appropriate values when they do not know enough about the data. Another limitation is that the randomness of The initial center point of the algorithm makes the clustering result easily fall into the local optimal solution, and the clustering result is unstable. Asha A considered the distribution of all data samples from a global perspective and calculated the arithmetic average of points among all samples, thus determining the initial center point of clustering [3]. However, the determination of this point consumes a lot of operation time and is easily affected by noise, so the data quality is required to be high. Researchers also failed to screen the noise effectively, resulting in the noise point being taken as the initial center, which may lead to inaccurate clustering and long calculation time [4]. Researchers considered the influence of density and combined with the "maximum-minimum principle", thus effectively avoiding local optimality, but data in low-density regions may be mistaken for outliers [5]. Ahmad W, et al. combined canopy algorithm with density, although it could effectively process low-density region data and automatically determine the number of class centers, it would only stop after traversing all data points, failing to consider the clustering effect and the accuracy of the number of centers, and lacking in the processing of noise points and outliers, making it easy to miss key information [6]. Depth-based methods can solve this problem by mapping data points into space, assuming that the data points will be wrapped layer by layer from the inside out, and the more data points in the outer layer will be defined as more abnormal, but the operation is not practical in high-dimensional data. Researchers have proposed many detection methods based on density-clustering, such as LOF, INFLO, INS, etc., which have high effectiveness and strong simplification and will be widely used [7]. This algorithm identifies outlier subsets by evaluating the degree of abnormality of each data point in the dataset, they also have some shortcomings that cannot be ignored. This algorithm identifies and determines a subset of outliers by evaluating the degree of abnormality of each data point in the dataset, and usually selects several data points with large outlier value as outlier points. This method, which uses outlier factors to determine outlier subsets, has high detection efficiency in small-scale data sets where the number of outliers is known. However, most outlier detection algorithms do not pre-set the specific number of outliers during execution, which gives them some flexibility in determining outliers, but also brings challenges, and the data scale is large. Obviously, preliminary filtering of data will reduce the interference to normal data.

Based on the research and learning of various improved algorithms of traditional K-means clustering algorithm, this paper has optimized the problems such as the unsatisfactory clustering results caused by outliers and the disadvantages of the initial center point affecting the initial division, and achieved good results. By comparing with three classical algorithms in four data sets, this paper verifies that the proposed algorithm significantly improves the clustering quality and accuracy.

2. Algorithm Introduction

2.1 Introduction to K-Means ++ Algorithm

In the classics k-means algorithm, the sample set D is divided into several disjoint subsets through iteration technology, and the cluster class is divided according to the distance between the sample and the cluster center point, and the iteration until convergence. The final goal is to minimize the square error, and the objective function is:

$$\mathrm{E} = \sum_{i=1}^{K} \sum_{x=C_i} \|x - \mu_i\|^2 \tag{1}$$

Traditional K-Means clustering algorithm has defects. Cluster k needs to be set in advance, and its value is difficult to estimate. In many cases, a given data set cannot be determined to cluster into several classes in advance. The K-Means ++ algorithm is optimized accordingly. Maximizing the distance between cluster centers, although intuitive and simple, is extremely effective. By increasing the spacing between cluster centers, the clustering effect can be improved, ensuring the tightness within each cluster and the discrimination between different clusters, and it well improves the randomness of the initial center point selection of K-Means algorithm. After determining the initial point, the rest parts are consistent with the original clustering algorithm. Researchers although it is easy to implement for K-Means++, it is not easy to use it on large data, and it requires traversal of all data, so its applicability to large data is limited [8]. Noise points in low-density regions are more likely to be selected as clustering centers, so that the data belonging to such centers is too small, and the possibility of change in the subsequent K-Means algorithm iteration process is very small, resulting in the failure to achieve the classification effect due to K clustering centers. Through the introduction function, it is found that the noise point in the low-density region is more likely to be selected as the clustering center, which reduces the amount of data belonging to this type of center, and the possibility of change in the subsequent iteration process is small, and it is difficult to achieve good clustering effect. For a data set composed of two clusters, one of which contains a noise point, the K-Means++ algorithm will divide the data set into two cases. In the first case, the noise points will be classified into a single class, and the remaining points will be classified into a class. In the second case, the noise points will be classified into a class with the surrounding data points, resulting in the misclassification of the original data points. The correct clustering result is not obtained.

2.2 Canopy + K-Means Algorithm Introduction

Canopy algorithm pre-classifies data on the basis of K-Means algorithm, and can be approximated by the number of large circles generated by canopy when the number of cluster centers cannot be determined artificially [9]. Canopy processes data sets through two artificially determined thresholds t1 and t2(e.g., Figure 1. Effect of canopy classification), which can sort chaotic data into several data piles with certain rules. The algorithm flow is as follows:
(1) Determine the two thresholds t1 and t2 (t1>t2). (2) Select a data at random from the data set and calculate the distance between this data and canopy (if there is no canopy at present, this point is directly used as the canopy center point). (3) If this distance is less than t1, this data is marked with a weak label and t1 is added to this canopy (at the same time, this data can be used as a new canopy to calculate the distance of other data to this point). (4) If this distance is less than t2, this data will be strongly marked and the data set in it will be deleted. At this time, it is considered that this data point is close enough to the canopy and no new canopy needs to be formed. (5) Repeat the process 2-4 until there is no data in the data set.

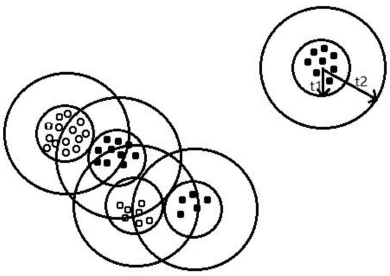

Figure 1. Effect of Canopy Classification

To solve the problem that the number of classification is difficult to determine, Canopy+ K-Means algorithm can approximate the number of cluster centers by the number of large circles generated by canopy. However, the selection of initial values such as the initial Canopy center point and Canopy area size has a great influence on the clustering quality when the algorithm is applied in practice [10].

2.3 Local Anomaly Detection Method of LOF Based on Density

Many researchers have different definitions of outliers according to different detection methods. Researchers gave the most classic definition of outliers: an outlier represents a data point that deviates so badly from other data points that it is suspected to be generated by different mechanisms. The density-based LOF anomaly detection algorithm calculates the degree of dispersion obtained by comparing the local density of an object P with the surrounding density. [11]. The following definitions are involved:
(1) Reach distance
The KTH reachable distance from point O to point P is defined as follows:

$$RD_K(p,o) = max\{k - distance(o), d(p,o)\} \tag{2}$$

Where: d (p, o) represents the distance between P and O. distance indicates the distance from the KTH point of point P, excluding point P; K-distance represents the k th distance, and the K th distance of P is also the distance from the K th point of P, excluding the point P. By definition, the KTH reachable distance from point O to point P is at least the KTH distance from O, or the true distance between OP's.
(2) Local reachability density
The locally accessible density of point P is represented as follows:

$$\mathrm{LRD}_k(p) = \frac{1}{\left(\frac{\sum_{o\in N_K(P)} reach - dist_k(p,o)}{|N_K(P)|}\right)} \tag{3}$$

Where: Reach-distance (p) represents the relative distance between point P and K field; K-distance neighborhood of p represents k field distance NK(P) of point P, and indicates the number of K field points of P; local reachability density to represent the local relative density of a data object.
(3)Local outlier factor
Expressed as follows:

$$LOF_K(P) = \frac{\sum_{O\in N_K(P)} \frac{lrd(o)}{lrd(p)}}{|N_K(P)|} = \frac{\sum_{O\in N_K(P)} lrd_k(o)}{|N_K(P)|} / lrd_k(p) \tag{4}$$

LOF represents the local outlier factor. The value of $LOF_K(P)$ approaches 1, indicating that point P is close to its domain density value and belongs to a cluster. The more its ratio is less than 1, the higher the density of the point P is than its domain density, and the point P is defined as a dense point. The greater the ratio is than 1, the lower the density of point P is compared to its domain density, and point P is defined as a dispersion point [12].

3. GDD-K-Means Clustering Algorithm Improved Based on Traditional Algorithm

3.1 Outlier Removal

Firstly, a grid-based outlier filtering method is used to initially screen the candidate outlier subset through grid filtering. The algorithm focuses on considering the density threshold of the grid distribution of data points in the global range to determine whether there is an outlier. The density threshold is taken as a filter, and the data set with a density less than the threshold is taken as the candidate outlier subset. This stage can effectively reduce the amount of computation to a certain extent. Then the density outlier detection method is used to determine more accurate abnormal data points. The performance of the algorithm is improved effectively, and the time complexity of the algorithm is reduced. In the grid filtering stage, the data set is scanned and each data point is mapped to the corresponding grid cell to complete the mapping task.

(1)Grid step parameter setting

Divide the data set into each grid, and set the number of data sets as $\mathrm{N} = \{n1, n2, n3, \dots, n_n\}$,The number of grids is m * m, and the number of grids and the number of datasets are mutually dependent. There is a functional relationship between the size of the data set and the size of the grid, and the mesh partitioning function can be defined as:

$$\mathrm{m} = \left[|N|^{1/3} + |N|^{1/4}\right] \tag{5}$$

Where: N is the size of the data set and m is the number of rows in the grid.

(2) Set the density threshold of the grid

The size of the data set is closely related to the number of grids, which varies with the size of the data set. We can determine the density threshold by the size of the data set. When the data set is larger, the data distribution in the grid will be more dispersed, the step size will increase, and the density threshold will be correspondingly larger. Otherwise, it is impossible to ensure that the non-dense grid data will be integrated as candidate outliers, resulting in the loss of filtering effect of grid filtering. Then the density threshold of the grid will have a functional relationship with the size of the data set, and the density threshold function is defined as:

$$\beta = \left[\frac{|N|^{1/3} + |N|^{1/4}}{3}\right] \tag{6}$$

Where: β is the threshold of mesh density and N is the size of the data set. The mesh step size and the mesh density threshold are calculated reasonably by the mesh division function and the mesh density threshold function, and the mesh density threshold can be used to judge whether the mesh is dense.

In the technique of grid filtering outliers, a preliminary subset of candidate outliers has been obtained, and then a density-based detection method is used to detect outliers more accurately. In the algorithm of grid-based outlier filtering, the density is represented by the number of points per unit area. The more the number of points per unit area, the greater the density and the greater the probability of becoming a normal point. On the contrary, the more likely it is to become a noise point (e.g., Figure 2. canopy classification rendering).

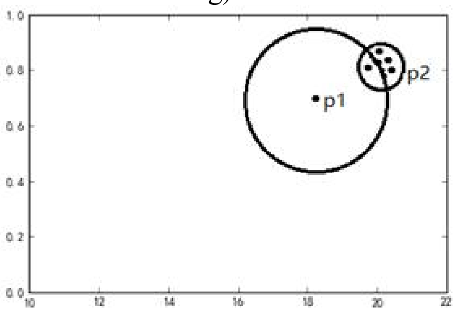

Figure 2. Canopy Classification Rendering

As shown in Figure 2, the area of the unit circle formed by the field of PI points is much larger than that formed by the field of P2 points. Taking the ratio of the number of fields to the area formed by the unit circle as the basis for judging the density of data points greatly reduces the time required by the detection algorithm and improves the efficiency of the algorithm. Based on the idea that the distribution of anomalous data is usually much more sparse than that of normal clusters, we will compare the number of data points in the k-distance field with the area of the circle formed by the K-distance field to the judgment basis of the density between data points.

3.2 Center Point Selection Method

In the traditional k-Means algorithm, an initial cluster center is randomly selected to determine an initial partition, and then the cluster is performed by iterative technology and the initial partition is continuously optimized. However, the selection of the initial center has a great influence on the result, and the effective selection of the initial value plays a crucial role in the clustering result. Therefore, we will select the center points successively according to the "maximum-minimum principle", which is shown as follows:

$$\mathrm{d}_l = \max_i \left[\min\left[d_{i1}, d_{i2}\right]\right] > \theta * \|z_1 - z_2\| \tag{7}$$

Where: θ is the selected scale coefficient, passed by $\mathrm{d_i} = \min[\mathrm{d_{i1}}, \mathrm{d_{i2}}], \mathrm{i} = 1,2, \ldots, \mathrm{N}$ The minimum value between samples is obtained, where the distance between samples is expressed by $\mathrm{d_{ij}} = \left\|\mathrm{x_i} - \mathrm{z_j}\right\|, \mathrm{j} = 1,2.$

The data point with the largest index value of the central point is used as the clustering center of the experimental data for pre-classification. The calculation methods of the distance in this algorithm all use Euclidean distance, and the radius of the data point is continuously and adaptively calculated according to the greedy strategy. The calculation formula is as follows:

$$\varepsilon = \frac{1}{K}\sum_{P_{I \in C}} d(P_i, P_{K-nearrst(i)}) \tag{8}$$

Where: $P_{K-nearrst(i)}$ represents the K points nearest to the point; In general, the value of K is 4 in the two-dimensional spatial cluster, and in other cases the value of $\lfloor n/25 \rfloor$ in the data set is taken. Where (n is the total number of data samples and ⌊ ⌋ is rounded down). The weight of the object is calculated according to the distance between the object and the data object q in the ε− domain, and the weight is processed to get the central point index of each data, the calculation formula is as follows:

$$C_P = W_P * \theta_P \tag{9}$$

Where: θ_P represents the distance between the data point θ and the center point i closest to itself, the calculation formula is as follows:

$$\theta_P = \min_{1 \le i \le k} d(i, p) \tag{10}$$

Is the weight of the data point, which represents the domain density of the point; The calculation formula is as follows:

$$W_P = \sum_{q=1}^{m} \frac{range - d(p,q)}{range} \tag{11}$$

Where: p indicates the number of current center points; m represents the number of objects of the data object in the field of ε− of the data point p.

It reflects the dimension size of the dataset in the vector space. Adopts the Euclidean distance calculation method, the calculation formula is as follows:

$$\mathrm{range} = \sqrt{\sum_{z=0}^{x} \left\|\max_z - \min_z\right\|^2} \tag{12}$$

Where: x represents the dimension of the data; Range represents the modulus of the whole dimension range of the dataset. The contribution degree of each data point in the ε− field of the point is greater the closer the point is, and the contribution degree range is [0,1]. In summary, the larger the weight value, the more data around the object point, the more dense. The larger the value, the farther away from the generated cluster center, the closer the cluster clustering. The larger the central index obtained by multiplication, the higher the degree of difference between the two clusters, the better the clustering effect. The time consumption is mainly determined by the number of iterations, and the number of iterations of K-means algorithm can be effectively reduced and the time performance of the algorithm can be improved by effectively selecting the clustering center of the center point index.

There is no cluster center point at the beginning of clustering, and the center point index of data cannot be calculated due to the lack of θ parameter. The more times a data point appears within a given range, the denser the data point, which is more conducive to the convergence of the objective function as the cluster center point. Therefore, selecting the point with the largest weight as the initial center point is also conducive to improving the tightness within the cluster, and conforms to the idea of "high density around the cluster center point". The initial center point obtained by combining the "max-min principle" with the strategy with the maximum weight can better eliminate the randomness of the initial center selection.

The method of selecting cluster center points is based on the idea of "there is a distance between cluster centers and the density of points around cluster centers", which is similar to the actual clustering effect. According to the central point index, it can be seen that the greater the density around the cluster central point, the greater the weight of the data point. If the "max-min principle" is applied to the central point index, the larger the weight of a data point and the greater the distance from other central points, the greater the possibility of this point becoming a new cluster central point, the greater the clustering between the cluster centers, the better the clustering result.

3.3 Selection of K Value of Cluster Number

After the data is pre-classified by the central point index, that is, all data points will be classified into the category of the nearest central point. The K-means algorithm belongs to the unsupervised clustering method, and the cluster number k needs to be set manually, but it is difficult to estimate. Therefore, it is an urgent problem to divide a given data set into several categories. This paper will refer to a new clustering index BWP proposed in the literature to automatically select K clustering centers. Through the introduction of BWP index and the combination of canopy algorithm, the algorithm can automatically determine the number of classes. The average value of BWP index is calculated as follows:

$$\overline{BWP}(j,i) = \frac{1}{n}\sum_{i=1,j\in j}^{n} \frac{b(j,i)-w(j,i)}{b(j,i)+w(j,i)} \quad (13)$$

Where: n represents the size of the dataset, $b(j,i)$; $w(j,i)$ will be defined as follows: There exists a dataset S with n data objects. Suppose that n data objects are divided into k classes, define the interclass distance of object i of class j, $b(j,i)$, to be the minimum value of the average value of samples from this sample to every other class, and define the intra-class distance of object i of class j, to be the average value of the distance between this data object and other data objects of class j. The calculation formula is as follows:

$$b(j,i) = \min_{1\le c\le k, c\ne j} \left(\frac{1}{n}\right)^{n} \sum_{p=1}^{c} \left\| x_p{}^{(c)} - x_i{}^{(j)} \right\|^2 \quad (14)$$

$$w(j,i) = \left(\frac{1}{n_j - 1}\right)^2 \sum_{p=1,p\ne i}^{j} \left\| x_p{}^{(j)} - x_i{}^{(j)} \right\|^2 \quad (15)$$

The change of BWP index determines whether to select the next cluster center point. According to the canopy clustering algorithm idea, all data points in the ε-field of the new cluster center will not participate in the subsequent center point selection. Compare the change of the mean value of BWP index of data points before and after the pre-classification. If the mean value of BWP index increases, this point will be used as a new clustering center, and the generation of clustering center will cause changes of data points, which will eventually be divided into the category of the center point closest to itself.

Therefore, the center point index needs to be updated every time a new center point is generated. If the BWP indicator becomes smaller or no data points exist, the selection of the center point will stop.

3.4 Basic Ideas of GDD-K-Means Clustering Algorithm

Firstly, the grid filtering method is used for preliminary screening of the data set, and the data whose density is less than a specific threshold is divided into candidate subsets. Then, the classical density-based LOF algorithm is used to accurately remove the outliers. When the outliers are removed from the data set, the point with the largest weight is selected as the initial center point, and then n center points are selected successively according to the central point index. When n+2 center points are generated, the average value of BWP index decreases, then the selection is stopped and n+1 center points are obtained. Finally, the generated center point is used as the initial clustering center to execute the k-means clustering algorithm, and the final clustering result is obtained to end the operation.

4. Experimental Results and Analysis

4.1 Experimental Environment

The experimental environment building platform of K-means, k-means ++, canopy + K-means, GDD-K-Means algorithm is: Core i3720M (1.80GHz) processor, The algorithms are developed using python3.6 language and Anaconda3&Spyder3 as development tools.

4.2 Experimental Data

In order to verify the effectiveness of the GDD-K-Means clustering algorithm proposed in this paper, test the time complexity of the algorithm and the performance of the clustering algorithm, verify the effectiveness and availability of the algorithm through the operational structure of the actual data set, and compare the detection performance of the algorithm in different data scales and different data structures. The data will be verified by selecting a synthetic UCI data set. It is a database proposed by the University of California for machine learning. The experimental data sets have clear classifications, so the quality of clustering can be directly observed. Experimental data will select Seeds, Wine two data sets of different data size detection, Dataset1, Dataset2 are used to represent the four data sets. Table 1 describes the basic characteristics of relevant experimental data sets.

Table 1. Experimental Data Description

Datasets	Samples	Attributes	Categories
Seeds	210	7	3
Wine	178	13	4

4.3 Data Processing

Different feature values in a data set often have different dimensions. When different features are listed together, small data in absolute values will be ignored by big data in data mining processing due to the different expression ways of the features themselves. In order to ensure that each feature is at the same level, and ensure that each feature reasonably participates in the execution of the algorithm, the impact of dimension is eliminated. The data were normalized using the Min-Max Scaling method.

For the data set with a feature, by traversing each data in the feature vector, Max and Min are recorded, and max-min is used as the base (that is, Min=0, Max=1) for data normalization processing. The calculation formula is as follows:

$$x_{normalization} = \frac{x_i - \text{Min}\,(xi)}{\text{Max}\,(xi) - \text{Min}\,(xi)} \quad (16)$$

4.4 Experimental Evaluation

In order to perform effective cluster evaluation on clustering results, this paper adopts the most commonly used evaluation criteria in data mining field: BWP index, Rand index, contour coefficient, recall rate, accuracy rate and F-measure as performance indicators.

4.5 Comparison Experiment of Clustering Results

In the test on UCI data sets, grid filtering is used to preliminary screen the data, filter out the densest data clusters, and minimize the data set to be detected as much as possible. The grid filtering algorithm screens the preliminary candidate subset of the data set in Table 2, and calculates the number of grid divisions and the threshold of grid density. The results are as follows:

Table 2. Experimental Data Statistics

Datasets	Number of meshing	Grid density threshold	The threshold is partially reached
Seeds	9*9	3	5
Wine	9*9	3	5

If the density coefficient is lower than the density threshold, we define it as a candidate anomaly subset. If the density coefficient is higher than the density threshold, we define it as a dense subset, which does not fit into the candidate subset. Through grid filtering, the data set is initially screened to effectively remove the most dense part of the data set and reduce the data set to be used as much as possible. Then, through density-based detection method, noise points are more accurately screened, which improves the accuracy of the algorithm, greatly reduces the running time and improves the efficiency of the algorithm.

In the second stage, based on the density-based outlier detection method, we finally calculate the abnormal data set from the candidate subset and carry out removal processing. The grid filtering method effectively reduces the number of data sets. Because the local density of data points in the sparse grid is small, which contradicts the feature of the cluster center with the large local density, the selected points will not become the cluster center, and the removal will not affect the selection of the cluster center. Without affecting the selection of clustering, successfully shortened the execution time of the algorithm, thereby significantly improving its running speed. (e.g., Figure 3. Cluster evaluation index results).

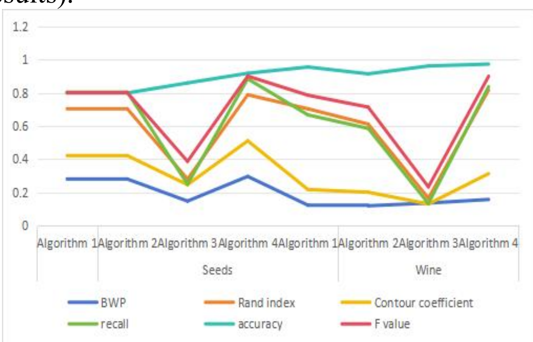

Figure 3. Cluster Evaluation Index Results

This paper selects the point with the largest weight as the initial central point strategy through the central point index C, accurately clusters the initial central point of the cluster, improves the tightness within the cluster, eliminates the randomness of the initial center selection, and makes the cluster can be divided in one stage. Figure 3 shows the experimental results of the improved clustering algorithm and the performance indicators of the other three algorithms, and the clustering effect can be evaluated.

The experimental results clearly indicate that, in Figure 3, the BWP index, contour coefficient, Rand index, accuracy rate and recall rate of clustering results obtained by the GPD-K-means algorithm proposed in this paper are significantly better than K-Means algorithm, K-Means ++ algorithm and Canopy + means algorithm. The reason is that the traditional algorithm randomly selects the initial center point. Although K-Means ++ incorporates the factor of distance in the selection of the center point, the generated data with far distance from the center point is more likely to be selected as the next clustering center, and the selection is still random. In the four data sets, the proposed algorithm has the best performance in the evaluation index, and the BWP index is obviously higher than other clustering algorithms, which indicates that the intra-class precision is stronger and the inter-class separation is better. Through performance indicators, it is found that accuracy, recall rate and F value are greatly improved, which indicates that the effective improvement of clustering results depends on the effective selection of initial clustering center by central indicators, reasonable selection of cluster number and removal of the influence of noise points. It indicates that Canopy + K-Means clustering algorithm can obtain better clustering effect. Secondly, the data is de-noised in the early stage, and the initial center is selected in the calculation process, which consumes a certain amount of time, but the determination of the initial center provides a strong support for the classification of clusters, reducing the number of iterations and the time complexity of the algorithm operation. In the comparison of four clustering algorithms, the advantages of the modified algorithm in running time and clustering effect are shown directly. Therefore, the improved clustering algorithm not only improves the accuracy and stability of clustering, but also improves the operation efficiency.

Through the operation on the four data sets, the clustering results of the GDD-K-Means clustering algorithm and the three clustering algorithms introduced previously are compared, as shown in Figure 4 and Figure 5.

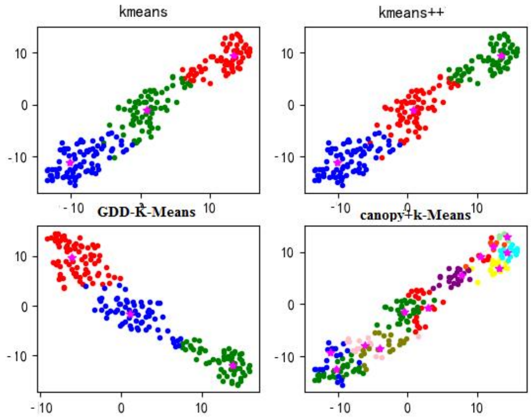

Figure 4. Seeds Data Clustering Class Result Visualization

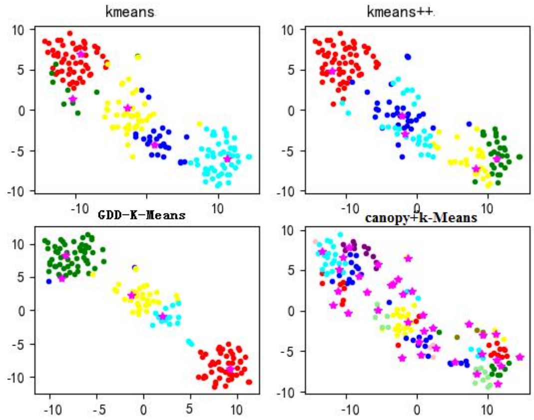

Figure 5. Wine Data Clustering Class Result Visualization

From the observation of data distribution in Figure 4, we can see that the data set has poor focusing and small span, and the clustering effect in this paper is good. Although there are still a few singular values, the cluster center quality is high, and canopy algorithm is obviously not applicable to this data set. By observing Figure 5, we can clearly see that, the linear distribution of data integration and the small span of coordinate interval can be considered as concentrated distribution. For linear distribution data, through the above three algorithms in this paper can support linear distribution data, while canopy has weak support for linear distribution data. Based on the above observations, it can be concluded that the algorithm proposed in this paper is effective in dealing with irregular data sets, linear distributed data sets with strong focus and dense data clustering, and the clustering center quality is higher than k-means, K-Means ++ and canopy.

By examining the visualization chart of the above clustering results, it can be clearly observed that the improved GDD-K-Means clustering algorithm proposed in this paper performs well in data clustering and has significant clustering effects.

5. Closing Remarks

With the rapid increase of Internet data, the noise points of data will also increase, which will affect the clustering effect of data. In this paper, the method based on grid filtering outliers not only ensures the purity of sample data, but also provides a guarantee for the accuracy and timeliness of k-means clustering algorithm in selecting the initialization center. Through dynamic outlier detection, the computational complexity is reduced and the computational efficiency is improved to some extent. The GPD-K-Means algorithm weighs the relationship between distance and density through the weight and the central point index, and solves the problem that randomly select the number of initial central points due to the influence of the initial center selection, so that the improved clustering effect is better. Through the results of different evaluation indicators, it can be found that the proposed algorithm achieves better results than the traditional algorithm in the processing of different data amounts, but there is still room for further improvement and improvement in the screening of noisy data. The next step of this paper will be to remove outliers more accurately and apply it to the processing and analysis of text data.

Acknowledgments

This paper is supported by Changji College 2023 Campus level Research Project (No. KYLK030).

References

[1] Juwaied A, Strumillo J L. Improving Performance of Cluster Heads Selection in DEC Protocol Using K-Means Algorithm for WSN. Sensors, 2024, 24(19): 6303-6303.

[2] Shi J. Optimization of frozen goods distribution logistics network based on k-means algorithm and priority classification. Scientific reports, 2024, 14(1): 22477.

[3] Asha A, Rajesh A, Lenin M K. Adaptive fuzzy-based node communication performance prediction with hybrid heuristic Cluster Head selection framework in WSN using enhanced K-means clustering mechanism. Journal of Ambient Intelligence and Smart Environments, 2024, 16(3): 309-335.

[4] Sabbagh A A, Hamze K, Khan S, et al. An Enhanced K-Means Clustering Algorithm for Phishing Attack Detections. Electronics, 2024, 13(18): 3677-3677.

[5] Klen M A, Bonduà S, Kasmaeeyazdi S, et al. A fuzzy K-Means algorithm based on Fisher distribution for the identification of rock discontinuity sets. International Journal of Rock Mechanics and Mining Sciences, 2024, 182105879-105879.

[6] Ahmad W, Singh A, Kumar S, et al. Optimizing Energy Efficiency in Wireless Sensor Networks using Enhanced K-Means Cluster Head Selection. International Journal of Communication Networks and Information Security, 2024, 16(3): 565-573.

[7] Zeng B, Li S, Gao X. Threshold-driven K-means sector clustering algorithm for wireless sensor networks. EURASIP Journal on Wireless Communications and Networking, 2024, 2024(1): 68-68.

[8] Kaizheng W, Yitong F, Shunzhen Z, et al. Cloud detection from Himawari-8 spectral images using K-means ++ clustering with the convolutional module. International Journal of Remote Sensing, 2024, 45(3): 930-953.

[9] Preciado J L A, Aké C S, Martínez V F. Identification of Patterns in CO2 Emissions among 208 Countries: K-Means Clustering Combined with PCA and Non-Linear t-SNE Visualization. Mathematics, 2024, 12(16): 2591-2591.

[10] Jahandoost A, Torghabeh A F, Hosseini A S, et al. Crude oil price forecasting using K-means clustering and LSTM model enhanced by dense-sparse-dense strategy. Journal of Big Data, 2024, 11(1): 117-117.

[11] Nowak A B, Czesław H. Outliers in Covid 19 data based on Rule representation - the analysis of LOF algorithm. Procedia Computer Science, 2021, 1923010-3019.

[12] Alok M, Bradford T. Applied Unsupervised Learning with R: Uncover hidden relationships and patterns with k-means clustering, hierarchical clustering, and PCA. Packt Publishing Limited: 2019-03-27.