Journal of Computers Vol. 34 No. 1, February 2023, pp. 75-86
doi: 10.53106/199115992023023401006

Canopy-MMD Text Clustering Algorithm Based on Simulated Annealing and Canopy Optimization

Jun-Wu Zhai, Yu-Chen Tian, Wen-Tao Li, Kun Liang*

College of artificial intelligence, Tianjin University of science and technology, Tianjin, 300457, China
{junwu_zhai, liwentao}@mail.tust.edu.cn, liangkun@tust.edu.cn

Received 12 April 2022; Revised 4 August 2022; Accepted 10 August 2022

Abstract. Aiming at the problems that traditional K-means text clustering cannot automatically determine the number of clusters and is sensitive to initial cluster centers, this paper proposes a Canopy-MMD text clustering algorithm based on simulated annealing and silhouette coefficient optimization. The algorithm uses the simulated annealing algorithm combined with the silhouette coefficient to optimize the Canopy algorithm to find the optimal number of clusters, and uses the optimal number of clusters to determine the scale coefficient in the MMD algorithm, and finally achieves a better text clustering effect. The Sohu News dataset of Sogou Lab is experimentally analyzed and compared with the clustering results obtained by traditional K-means and algorithms in the literature. The experimental results show that the clustering performance of the algorithm is better than the traditional K-means algorithm and the algorithm in the literature, and the accuracy, precision, recall and F value are improved by 8.02%, 8.91%, 8.02%, 9.51% compared with the traditional K-means algorithm, which can be widely used in fields such as text mining, knowledge graph and natural language processing.

Keywords: simulated annealing, canopy, K-means, silhouette coefficient, text clustering

1 Introduction

With the rapid development of Internet technology, a variety of data has exploded, including a large number of text information. Therefore, it is very important to use text classification technology to organize and manage massive data scientifically. Machine learning techniques do not require human intervention for text classification, so they are widely used [1]. Clustering is the process of dividing a group of physical or abstract objects into categories according to certain criteria. After clustering, objects in each cluster are as similar as possible and objects in different clusters are as different as possible [2].

As the most classical clustering algorithm, K-means algorithm has been widely used in text clustering. However, K-means algorithm also has limitations, such as high requirements on the initial clustering center, easy convergence of the algorithm to the local minimum, etc., and unreliable text clustering results [3-6]. In the traditional text clustering method, aiming at the two shortcomings of K-means algorithm, different scholars have improved it from different angles. For example, in literature [7], the combination of particle swarm optimization algorithm and K-means algorithm for document clustering analysis improves the defect that K-means algorithm is easy to fall into local optimum. Literature [8] uses kernel function to execute K-means algorithm. And the improved algorithm is used for text clustering, which improves some defects of traditional K-means algorithm. Reference [9] modified the similarity calculation between words to improve the performance of K-means clustering. Referring to [10], the K-means algorithm is optimized by using density peak, and the local density and relative minimum distance of samples are introduced to optimize the initial clustering center. However, the problem of falling into local optimum is not fundamentally solved, and the reliability of text clustering is reduced.

Aiming at the problem that the number of clusters cannot be determined for a non-given corpus, the Canopy algorithm can roughly divide the data into several overlapping subsets, treat each subset as a cluster, and use a similarity measure method with low computational cost. You can roughly determine the number of clusters in a short time. It is mainly used for high-dimensional data clustering and is usually used as the initialization operation of other clustering methods, but it is sensitive to the initial threshold T1 and T2 and needs to be set manually. Some researchers also combined Canopy algorithm with K-means algorithm for cluster analysis. In reference to

* Corresponding Author

[11], the Canopy algorithm is combined with K-means algorithm for Chinese text clustering, which improves the clustering performance compared with K-means algorithm. Reference [12] combined Canopy algorithm with K-means algorithm to increase parallelization. The research reduces the calculation of redundant distance and speeds up the convergence speed of the algorithm.

To solve the problem that K-means is sensitive to the initial cluster center, MMD algorithm adopts the max-min distance principle and selects K data objects with the farthest distance from each other as the initial cluster center. Referring to [13], an improved algorithm based on density is proposed, which uses the K points with the highest density as the initial clustering center. Reference [14] combined MMD algorithm with K-means algorithm to perform cluster analysis on college scores. The selection of initial cluster center is optimized. These algorithms improve the accuracy of clustering to a certain extent, but they are sensitive to the initial scale coefficient, and do not fundamentally solve the problem that the algorithm is easy to fall into local optimal solutions, and they have not been tried in the field of text clustering.

Based on the above problems, the main contributions of this paper can be summarized as the following three points:

(1) Aiming at the problem that the number of clusters is uncertain and the Canopy algorithm is sensitive to the initial threshold in a non-given corpus, an improved Canopy algorithm based on simulated annealing is proposed, which mainly realizes the functions of finding the optimal initial threshold and automatically calculating the number of clusters.

(2) Aiming at the problem that the traditional text clustering algorithm K-means is sensitive to the selection of initial clustering centers and easy to fall into local optimum, this paper studies and designs MMD algorithm based on contour coefficient optimization, which mainly realizes the automatic selection of initial clustering centers and the re-optimization of clustering results.

(3) Finally, Canopy MMD algorithm is used to cluster the text, and cosine distance is used to calculate the similarity between the text samples, as well as the proposed accuracy, accuracy, recall and F1 value.

In the second part of this paper, we will introduce the improved Canopy algorithm based on simulated annealing. The third part introduces MMD algorithm based on contour coefficient. The fourth part describes the improvement of the core clustering algorithm: Canopy-MMD text clustering algorithm based on performance evaluation and Canopy optimization. The fifth part introduces the experiment. Finally, the sixth part summarizes the experimental results.

2 Improved Canopy Algorithm Based on Simulated Annealing

2.1 Traditional Canopy Algorithm

Canopy algorithms are generally “coarse” clustering before K-means algorithms. The selection of K value of the K-means algorithm has a great influence on the clustering result, and the size of K value needs to be set in advance when it is used, which has poor anti-interference ability against noise. Therefore, Canopy algorithm is used to determine the clustering number K of the data set with unknown sample distribution. Canopy algorithm is shown in Fig. 1, and the main steps are summarized as follows:

Step 1: Randomly arrange the original sample set into the sample list $L=[x_1, x_2, \ldots, x_m]$ (no change will be made after the arrangement), set the initial distance threshold T_1 and T_2 according to prior knowledge or cross-verification parameters, and $T_1>T_2$.

Step 2: Take any sample P from the list as a Canopy center and remove P from the list.

Step 3: Take any sample Q from the list, calculate the distance between Q and all Canopy centers, and take the minimum distance D: if $D\leq T_2$, then Q is strongly correlated with the Canopy, and Q belongs to the Canopy. Update Canopy’s centroid coordinates to the position of the center of all strongly correlated samples, and remove Q from the list. If $D\leq T_1$, Q is weakly correlated with the Canopy, Q belongs to the Canopy, and Q is added to it. If $D>T_1$, then Q forms a new Canopy center, removing Q from the list.

Step 4: Repeat Step 3 until the number of elements in the list is zero.

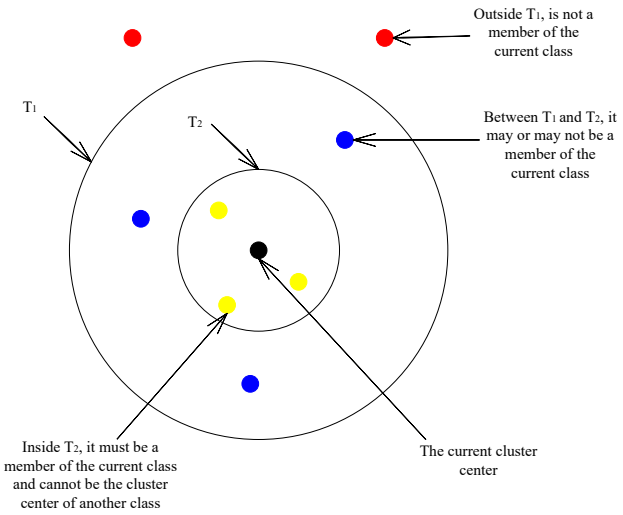

Fig. 1. Canopy algorithm schematic diagram

2.2 Simulated Annealing Algorithm

Simulated annealing [15] algorithm is an algorithm proposed by Metropolis on the basis of studying the annealing process of solid substances. It is widely used in the field of combinatorial optimization, such as machine learning and other scientific fields. The algorithm is an approximate method for solving optimization problems based on Monte Carlo design. The algorithm has the advantages of simple structure, easy implementation, few initial conditions, flexible use, high efficiency and strong ability to combine with other algorithms, so it has a broad application prospect. The core of simulated annealing algorithm is based on thermodynamic criterion of solid cooling. It can not only accept the optimal value, but also accept the bad value according to the probability of temperature variable, which increases the probability of the algorithm jumping out of the local optimal solution in the optimization process. The steps of the simulated annealing algorithm are summarized as follows:

Step 1: Set the initial temperature, randomly initialize an initial solution, and calculate the value of the objective function under the initial solution.

Step 2: Randomly take a new solution and calculate the objective function value of the new solution.

Step 3: According to the Metropolis criterion, judge whether to accept the new solution according to the objective function value $E(X_{new})$ of the new solution, the objective function value $E(X_{old})$ of the old solution and the current temperature T. If the new solution is accepted, assign the value of X_{new} to X_{old} instead of the old solution. Where the Metropolis criterion for acceptance probability is:

$$p = \begin{cases} 1 & , E(X_{new}) \le E(X_{old}) \\ e^{-\frac{E(X_{new})-E(X_{old})}{T}} & , E(X_{new}) > E(X_{old}) \end{cases} . \tag{1}$$

Step 4: Drop the temperature.

Step 5: Check whether the temperature is too high. Otherwise, go to Step 2.

2.3 Improved Canopy Algorithm Based on Simulated Annealing

Since the selection of initial threshold T_1 and T_2 in Canopy algorithm has a great impact on the clustering results, and given the sensitivity of Canopy algorithm to the initial threshold, this paper proposes an improved Canopy algorithm based on simulated annealing to measure the clustering quality by calculating the sum of squares of total errors. The definition formula of the sum of squared errors is shown in Equation (2).

$$SSE = \sum_{i=1}^{k} \sum_{x \in S_i} (d(c_i, x))^2 \quad . \tag{2}$$

Where d is the cosine distance between two points, k is the number of clusters, c_i is the center of the ith cluster, and S_i is the set of points of the ith cluster.

Generally, the relationship between the number of clustering K and SSE can be drawn as a line chart, and the approximate range of the optimal value K can be determined according to the elbow method [16]. The principle is as follows: In the clustering process of data sets, with the continuous increase of K, the data is segmented more finely, the clustering center gradually increases, and SSE gradually decreases. When K is less than the real number of clusters, the value of K becomes larger and larger, and SSE changes more. The two-dimensional image shows that the line between two points is steeper and the absolute value of the slope is larger. When K is equal to the real number of clusters, the value of K increases, SSE changes little, and there are parallel lines between two points in the two-dimensional image. As shown in Fig. 2(a), the range of "elbow" is the best range of K value. However, under certain circumstances, SSE changes without obvious inflection points may also occur, as shown in Fig. 2(b). You can set the threshold manually. The optimal number of clusters is determined according to the decrease of SSE variation to a certain threshold.

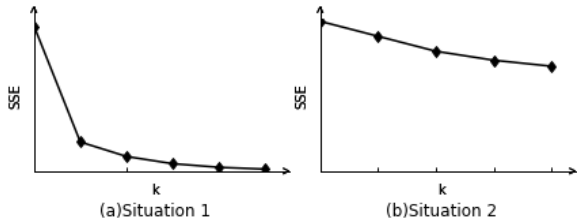

Fig. 2. Relationship between SSE and K values

Data Set Sampling. Because the data set may be too large and the number of labels of each class is unbalanced, the computation time of SSE clustering is high and the computation cost is too high. In order to determine the number of clusters, the subsets extracted from the original data set can be used to perform proportional extraction of each category of data.

Given a data set $X=\{X_1, X_2, \ldots, X_n\}$, each sample in X has m features, a smaller data set X' is randomly selected from the original data set, make $|X'|= s \times |X|$. Among them, $|X|$ and $|X'|$ respectively represent the number of samples contained in the data set, and s is the sampling ratio, which is a hyperparameter and needs to be determined according to the density of data distribution in the data set.

Algorithm Principle. Elbow method of uncertain factor is only determine the scope of the K value, to determine the maximum variations in SSE how much is the accurate values of K, the relationship between K and SSE can be the first point in the line chart and the last point connected to form a straight line $y = ax + b$, called the line, in line with the increase of x, y is a linear gradient, The SSE variation increases first and then decreases with the increase of K. Therefore, the SSE variation reaches the peak at a certain K value, and the K value at this time is the K value with the largest SSE variation, which is also the optimal number of clusters.

Literature [17] proposed the search range of the optimal cluster number in fuzzy clustering method, and theoretically proved that it is reasonable for the maximum cluster number not to exceed $\sqrt{n}$. In addition, literature [18] uses the distance cost function as the validity test function of spatial clustering, calculates and determines the upper bound of the optimal solution according to the empirical rule, gives the condition of the upper bound of K value, and theoretically proves the rationality that the maximum clustering number of the empirical rule does not exceed $\sqrt{n}$. So you can set the maximum number of clustering for the square root of $\sqrt{n}$, most part of class number is set to 2 to determine the baseline equation, with the help of searching ability and the Canopy of the simulated annealing algorithm is combined with, in terms of baseline and clustering results as the objective function value of the absolute value of the difference of SSE can find a set of initial threshold T_1, T_2, makes the biggest change range of SSE. The algorithm flow chart is shown in Fig. 3.

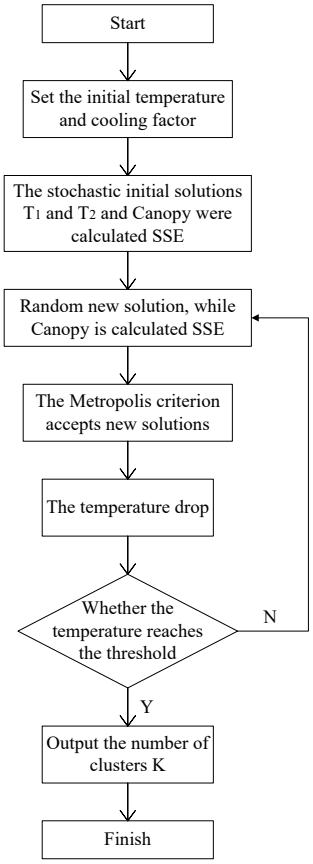

Fig. 3. Flow chart of simulated annealing improved Canopy algorithm

3 MMD Algorithm Based on Contour Coefficient Optimization

3.1 MMD Clustering Algorithm

MMD algorithm, also known as max-min distance clustering algorithm, is a heuristic clustering algorithm in pattern recognition. Based on Euclidean distance, samples were selected as clustering centers as possible. This algorithm can avoid the initial cluster center of K-means algorithm being too close to each other, but the scaling coefficient θ needs to be set in advance. The main idea is to determine the cluster center according to the determined distance threshold, and then assign the samples to the corresponding category of each cluster center according to the nearest neighbor principle. The main steps of the algorithm are summarized as follows:

Step 1: Set the sample list to $x_1, x_2, \cdots, x_N$, set the scaling coefficient θ.

Step 2: Select a sample as the first cluster center z_1.

Step 3: Select the sample farthest from z_1 from the list as the second cluster center z_2.

Step 4: Calculate the distance between x_i and z_1 and z_2 of the remaining samples, and work out their minimum value, that is:

$$d_{ij} = \| x_i - z_j \|, j = 1,2$$

$$d_i = \min(d_{i1}, d_{i2}), i = 1,2,\cdots,N$$

Step 5: If $d_l = \max_i[\min(d_{i1}, d_{i2})] > \theta \| z_1 - z_2 \|$, the corresponding sample x_l will be used as the third clustering center $z_3 = x_l$ and go to Step 6; otherwise, go to Step 7.

Step 6: Assume that there are k clustering centers, calculate the distance d_{ij} from each sample that is not used as the clustering center to each cluster center, and calculate $d_l = \max_i[\min(d_{i1}, d_{i2}, \cdots, d_{ik})] > \theta \| z_1 - z_2 \|$, then go

to Step 6 for $z_{k+1} = x_l$; otherwise, go to Step 7.

Step 7: When it is determined that there are no new clustering centers, the samples are divided into various categories according to the principle of minimum distance, namely, calculation:

$d_{ij} = \| x_i - z_j \| (j = 1, 2 \cdots, N)$

If the $d_{il} = \min_j [d_{ij}]$, then judge $x_i \in \omega_l$.

3.2 MMD Algorithm Based on Contour Coefficient Optimization

Since the cluster center selected by the MMD algorithm is fixed when it is run once, and the cluster center is not updated iteratively, it may lead to poor clustering performance when the samples are assigned to each cluster center according to the nearest neighbor principle. To solve this problem, this paper proposes an MMD algorithm based on contour coefficient iterative optimization. The contour coefficient of the internal evaluation index is introduced to evaluate the clustering results of each sample of MMD algorithm. The calculation formula of the contour coefficient is shown in Equation (3).

$$SC_i = \frac{b_i - a_i}{\max(b_i, a_i)} \quad . \tag{3}$$

Where SC_i is the sample contour coefficient, b_i is the minimum of the average distance between the *i*th sample and the samples outside the cluster, a_i is the average distance between the *i*th sample and the samples inside the cluster.

It can be seen from the calculation formula of the contour coefficient that the value range of the contour coefficient is [-1, 1], which can measure the similarity degree of a sample in a class and the separation degree from other classes. A positive value of the contour coefficient indicates that the sample is assigned to the correct class cluster, while a negative value indicates that the sample is assigned to the wrong class cluster. The contour coefficient of each sample was calculated, and the positive and negative values of each contour coefficient were judged. A negative value indicates that there is room for further optimization of cluster performance.

In order to further optimize the clustering results, the fault tolerance factor α and interference factor β are introduced, in which the fault tolerance factor α is used to measure the proportion of the number of samples with negative contour coefficient to the total number of samples, and is used as the end condition of the iterative optimization process. The calculation formula of α is shown in Equation (4).

$$\alpha = \frac{1}{N} \sum_i s(SC_i) \quad . \tag{4}$$

The *N* is the total number of samples, $s(x)$ is a deformation of the symbolic function, and the function expression is shown in Equation (5).

$$s(x) \begin{cases} 1 & x < 0 \\ 0 & x \geq 0 \end{cases} \quad . \tag{5}$$

After allocating the samples to the nearest cluster according to the nearest neighbor principle, the traditional MMD algorithm adds a step of iteratively updating the cluster center. After the each iteration, the profile coefficient of each sample is calculated. For samples with a silhouette coefficient less than 0, "redistribution" is performed until the ratio of the number of samples with a negative silhouette coefficient to the total number of samples is less than the fault tolerance factor α.

The repeated optimization based on the profile coefficients mentioned above is sensitive to the initial arrangement order of the samples, so the disturbance factor β is introduced to perturb the initial samples to a certain extent. The perturbation operation is one of the important factors to determine whether the algorithm can find the

global optimal solution. The non-convergence problem is caused by iterative optimization process. The perturbation operation is the process of exchanging initial samples randomly, and also the process of exchanging initial samples continuously with certain probability. The random exchange operation refers to the exchange of the positions of two samples x_i and x_j randomly selected for the initial sample $X = \{x_1, x_2, \cdots, x_N\}$. The perturbation operands follow the power-law distribution of the perturbation factor β, which is a hyperparameter of the algorithm. The schematic diagram of disturbance operation is shown in Fig. 4, and the flow chart of MMD algorithm based on contour coefficient optimization is shown in Fig. 5.

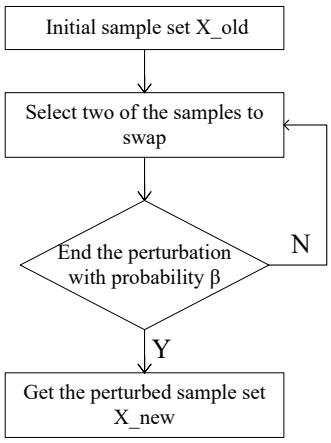

Fig. 4. Schematic diagram of disturbance operation

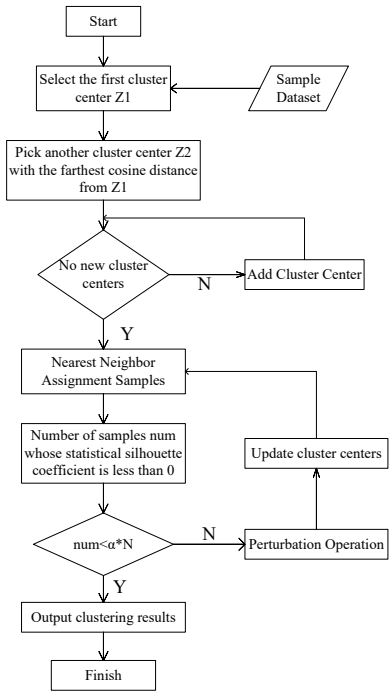

Fig. 5. Flow chart of MMD algorithm based on contour coefficient optimization

4 Canopy-MMD Text Clustering Algorithm Based on Simulated Degradation and Canopy Optimization

4.1 Data Preprocessing and Similarity Measurement

Belongs to the text data of unstructured data, so need to be done before a text document text clustering of pretreatment, the text data type conversion for numerical data can be used in the input text clustering algorithm, including the text pretreatment of the basic steps for: text extraction, unless in simplified Chinese, traditional Chinese, remove stop words, Chinese word segmentation and part-of-speech tagging, and text vector said.

This paper uses Jieba word segmentation in the Python language to perform word segmentation and partial phonetic annotation on text documents. The commonly used text representation models include Boolean space model (BM), suffix tree model (STM), vector space model (VSM) and probabilistic retrieval model (PM). In this paper, text vectorization using classical vector space model (VSM), and the vector representation is adopted for document D. The calculation formula of TF-IDF is shown in Equation (6):

$$TF - IDF_{i,j} = TF_{i,j} \times IDF_i \quad . \tag{6}$$

Where, $TF_{i,j}$ represents the frequency of occurrence of the word *i* in document *j*, and IDF_i is the anti-document frequency of the word *i*.

The similarity between text samples can be expressed by calculating the distance between two sample text vectors. In the text clustering analysis, using Euclidean distance to measure the similarity between texts will produce large errors. Therefore, this paper adopts the classical cosine similarity to measure the text similarity, as shown in Equation (7).

$$\cos(x_i, x_j) = \frac{x_i \cdot x_j}{\| x_i \| \cdot \| x_j \|} \quad . \tag{7}$$

As the cosine similarity between two documents is higher, the probability of two documents belonging to the same class cluster is higher. According to literature [19], the distance definition is modified into Formula (8).

$$d(x_i, x_j) = 1 - \cos(x_i, x_j) \quad . \tag{8}$$

Among them, the distance between two documents is negatively correlated with the cosine similarity value. When the text similarity is the highest, the cosine similarity value is 1, while the distance between two documents is the smallest, which is 0. When the text similarity is the lowest, the cosine similarity is 0, while the distance between two documents is the largest, with a value of 1.

4.2 Algorithm Principle

In this paper, the improved Canopy and MMD algorithm are used to cluster the text, as follows.

Firstly, the Canopy algorithm based on simulated annealing was used for “rough” clustering. The goal of “rough” clustering is to quickly obtain the number of clusters *K* of a dataset with unknown sample distribution. In the process of “rough” clustering, the optimization ability of simulated annealing is used to find the *K* value that maximizes the objective function by taking the difference between baseline and SSE as the objective function. The time complexity of simulated annealing is $O(T)$, and that of Canopy algorithm is $O(N * K)$, where *T* is the number of iterations, *N* is the total number of samples, and *K* is the number of clusters.

Secondly, the proportion coefficient in MMD algorithm can be known from the prior knowledge that the larger θ is, the fewer clusters will be generated by the algorithm, and the dichotomous condition will be satisfied. The *K* value determined by the Canopy algorithm was used to adjust the parameters automatically. The time complexity of the bisection method is $O(log2\varepsilon)$, where ε is the search accuracy.

Finally, the MMD algorithm of contour coefficient optimization is run to calculate the contour coefficient of each sample, perturb the sample set, update the clustering center iteratively, and reclassify the samples with con-

tour coefficient less than 0 until the fault tolerance evaluation condition is met, and the clustering is finished, and the text document classification is completed. The time complexity is $O(T*N*K)$. The total time complexity of this algorithm is $O(log2\varepsilon*T*N*K)$. In this paper, the sum of intra-class distances between text documents is selected as the clustering objective function, as shown in Equation (9).

$$J = \sum_{j=1}^{K} \sum_{\forall x_i \in c_j} d(x_i, c_j)^2 \quad . \tag{9}$$

Where, $d(x_i, c_j)$ can be calculated from formula (8), and c_j represents the cluster center.

5 Experiment

5.1 Experimental Setup

The experimental data set in this paper comes from Sohu news corpus provided by Sogou Lab, which includes news data from 18 domestic, international, sports, society and entertainment channels of Sohu news website from June to July 2012. In this paper, 1000 news articles in five categories of finance and economics, health, IT, education and sports are selected from the data set, and 200 news articles in each category are selected as the experimental data set. Each piece of news should focus on about 500 words.

The experimental configuration is 64-bit Win10 operating system, running environment is Intel(R) Core(TM) I7-8565U CPU, 1.80ghz1.99ghz, 8.00GB. In order to compare the performance of each algorithm, this paper runs each algorithm repeatedly for 10 times in the experiment, and takes the average value of 10 clustering evaluation indexes as the performance evaluation of the algorithm.

5.2 Experimental Results and Analysis

To test the performance of Canopy-MMD algorithm, four commonly used evaluation indexes in the field of text clustering are used in the experiment, namely Accuracy, Precision, Recall and F-measure. The text clustering experiment is compared with the traditional K-means algorithm, literature [7] algorithm, literature [10] algorithm and literature [11] algorithm. The parameter Settings of each algorithm are shown in Table 1. The SSE and K-plots and the objective function convergence of the Canopy algorithm based on simulated annealing are shown in Fig. 6 and Fig. 7. The results of the experimental dataset are recorded in Table 2.

Table 1. Parameter settings of each algorithm

Algorithm	Parameter settings
K-means	The number of clustering K=10
Canopy-KMeans	The threshold value $T_2 = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} d(x_i, x_j)}{c_n^2}$; $T_1 = T2 + 1e-3$
MMD	Scale coefficient Θ=0.5
Canopy-MMD	Initial temperature T_0= 1e-5, termination temperature T_n= 1e-14, S =0.1, *alpha*=0.16, *beta*=0.5

Table 2. Experimental results

Algorithm	Accuracy	Precision	Recall	F1
Canopy-MMD	**0.849**	**0.880**	**0.849**	**0.864**
K-means	0.786	0.808	0.786	0.789
MPK-Clusters	0.786	0.830	0.786	0.807
DCC-K-means	0.823	0.833	0.823	0.830
Canopy+K-means	0.740	0.731	0.740	0.736

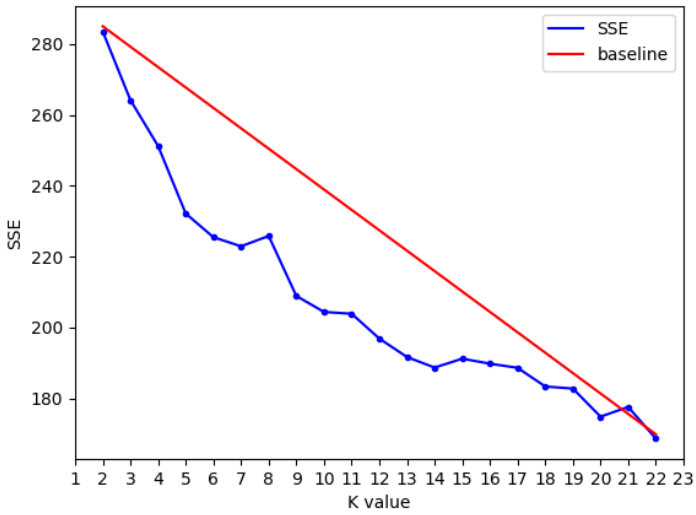

Fig. 6. Relation diagram of SSE and K

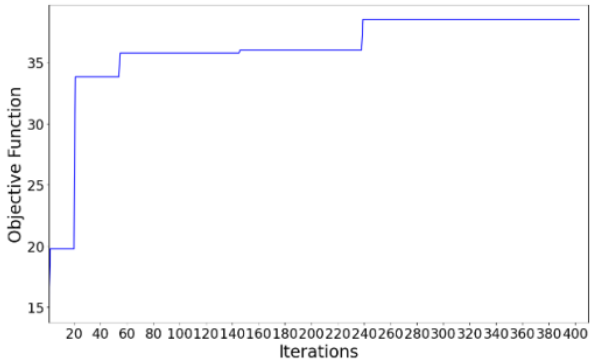

Fig. 7. Convergence of objective function

From Fig. 7, you can see that the proposed based on improved canopy of simulated annealing algorithm has better convergence and optimization ability, in the process of former 200 iterations are close to the global optimal solution, in the later iterations, in the near global optimal solution range of mutation is concentrated, to find a better solution, find more close at the optimal solution, has a good optimization ability. The comparison of each effect of the algorithm is shown in Fig. 8.

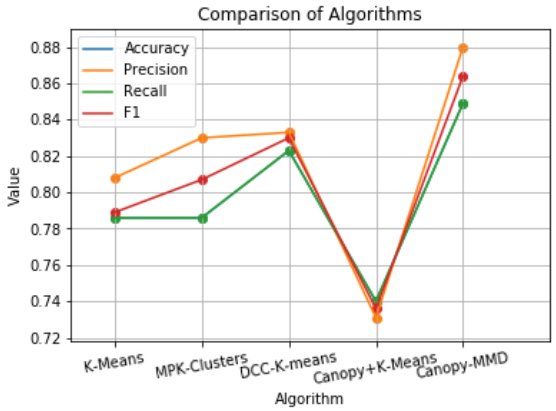

Fig. 8. Algorithm comparison

It can be seen from Table 2 and Fig. 8 that the Canopy-MMD algorithm proposed in this paper has higher accuracy, recall and F-value than K-means algorithm, MPK-clusters algorithm, DCC-K-means algorithm and Canopy+ K-means algorithm, and the text clustering performance is better. Compared with the traditional K-means algorithm, the accuracy, recall and F-value are improved by 8.02%, 8.91%, 8.02% and 9.51%, respectively. The reason is that the traditional K-means algorithm is easy to converge to the local extremum in high-dimensional sparse data, and convergence stagnation occurs in the process of text clustering, which is in the local optimum. Therefore, the reliability of text clustering results is not high. The Canopy-MMD algorithm optimizes sample clustering by finding the optimal number of clusters and contour coefficient through simulated annealing, avoiding the algorithm from falling into local extreme values, and thus improving the reliability of text clustering.

6 Experiment

In this paper, the improved canopy algorithm based on simulated annealing and MMD algorithm based on contour coefficient optimization are combined to solve the problem that the traditional K-means algorithm is easy to fall into local extreme values in the process of high-dimensional sparse text vector data clustering, which leads to the unreliable text clustering results and is sensitive to the number of clusters and the initial cluster center. The idea of iteratively updating the clustering center of K-means algorithm is added to avoid the problems of K-means algorithm. Through the text data clustering experiment, the results show that Canopy MMD algorithm has fast convergence speed and excellent optimization ability to find the optimal number of clusters when clustering five types of text documents, which significantly improves the accuracy, accuracy, recall and F-value of text data clustering. Therefore, the clustering results of Canopy-MMD text clustering algorithm are more reliable.

The current problem in this experiment is that the results of text word segmentation are not particularly ideal, and the text data cannot be separated according to the semantics. The next work is to improve the word segmentation optimization of text data and further improve the accuracy of text clustering.

7 Acknowledgement

This work is supported by the Scientific research project of Tianjin Education Commission (No. 2019KJ235), and the National Natural Science Foundation of China (No. 61807024, No. 61702367).

References

[1] L. Zhang, Y. Jiang, L. Sun, An Improved TF-IDF Text Clustering Method, Journal of Jilin University (Science Edition) 59(5)(2021) 1199-1204.
[2] Y.F. Yang, G.X. He, Y.D. Li, K-means algorithm for optimizing initial clustering center selection, Computer knowledge and technology 17(5)(2021) 252-255.
[3] C.S. Pan, B. Zhang, Y.N. Lvu, X.L. Du, S.M. Qiu, K-means Text Clustering based on Improved Gray Wolf Optimization Algorithm, Computer Engineering and Applications 57(1)(2021) 188-193.
[4] J.C. Yang, C. Zhao, Survey on K-Means Clustering Algorithm, Computer engineering and applications 55(23)(2019) 7-14+63.
[5] J.H. Qin, W.M. Fu, H.J. Gao, W.X. Zheng, Distributed K-Means Algorithm and Fuzzy C-Means Algorithm for Sensor Networks Based on Multiagent Consensus Theory, IEEE Transactions on Cybernetics (47)(3)(2017) 772-783.
[6] S. Khanmohammadi, N. Adibeig, S. Shanehbandy, An improved overlapping K-Means clustering method for medical applications, Expert Systems with Applications (67)(2017) 12-18.
[7] Y.L. Niu, B. Wu, Research on text clustering algorithm based on improved particle swarm optimization and K-means, Journal of Lanzhou University of Arts and Science (Natural Science Edition) 33(4)(2019) 44-47.
[8] G.F. Zhang, G.W. Wu, Improved k-means text clustering based on kernel function, Computer applications and software 36(9)2019) 281-284+301.
[9] J.F. Wang, X.X. Jia, Z.Q. Li, Research and implementation of short text clustering based on improved k-means algorithm, Information technology 43(12)(2019) 76-80.
[10] S.X. Tian, L.X. Ding, J.Q. Zheng, K-means text clustering algorithm based on density peak optimization, Computer engineering and design 38(4)(2017) 1019-1023.
[11] L. Zhang, X.W. Mou, Canopy+K-means clustering algorithm for Chinese text, Library Tribune 38(6)(2018) 113-119.
[12] L. Wang, J.C. Jia, Research on Parallelization based on improved Canopy-kmeans algorithm, Computer Measurement and Control, 29(2)(2021) 176-179+186.
[13] K.Q. Ma, Y.J. Yang, H.W. Qin, L. Geng, P.D. Wang, K-means clustering algorithm combining max-min distance and

weighted density, Computer engineering and applications 56(16)(2020) 50-54.
[14] H.B. Gu, W.P. Zhao, Cluestering algorithm based on Max- min Distance for students' score analysis in universities and applications, Journal of Hebei University of Engineering (Natural Science Edition) 27(1)(2010) 96-98 + 108.
[15] K.S. Chen, S.D. Xian, P. Guo, Adaptive Temperature Rising Simulated Annealing Algorithm for Traveling Salesman Problem, Control Theory & Applications 38(2)(2021) 245-254.
[16] A. Ng, Clustering with the K-Means Algorithm, the k-means learning section of Stanford University's Machine Learning course, 2012.
[17] J. Yu, G.S. Cheng, Search range of Optimal Cluster Number in Fuzzy Clustering Method, Science in China Series E: Technical Science 32(2)(2002) 274-280.
[18] S.L. Yang, Y.S. Li, X.X. Hu, R.Y. Pan, Optimization Study on k Value of K-means Algorithm, Systems Engineering-Theory & Practice 2006(2) 97-101.
[19] D. Yang, S.L. Zhu, Z.Y. Bian, Application of improved k-means algorithm in text mining, Computer technology and development 29(4)(2019) 68-71.