Improved Spectral Clustering Algorithm Based on Similarity Measure

Jun Yan[1], Debo Cheng[2,*], Ming Zong[2], and Zhenyun Deng[2]

[1] Geographic Center of Guangxi, Nanning, Guangxi, 530023, China
[2] Guangxi Normal University, Guilin, Guangxi, 541004, China
Cheng7294@foxmail.com

Abstract. Aimed at the Gaussian kernel parameter σ sensitive issue of the traditional spectral clustering algorithm, this paper proposed to utilize the similarity measure based on data density during creating the similarity matrix, inspired by density sensitive similarity measure. Making it increase the distance of the pairs of data in the high density areas, which are located in different spaces. And it can reduce the similarity degree among the pairs of data in the same density region, so as to find the spatial distribution characteristics complex data. According to this point, we designed two similarity measure methods, and both of them didn't introduce Gaussian kernel function parameter σ . The main difference between the two methods is that the first method introduces a shortest path, while the second method doesn't. The second method proved to have better comprehensive performance of similarity measure, experimental verification showed that it improved stability of the entire algorithm. In addition to matching spectral clustering algorithm, the final stage of the algorithm is to use the k-means (or other traditional clustering algorithms) for the selected feature vector to cluster, however the k-means algorithm is sensitive to the initial cluster centers. Therefore, we also designed a simple and effective method to optimize the initial cluster centers leads to improve the k-means algorithm, and applied the improved method to the proposed spectral clustering algorithm. Experimental results on UCI[1] datasets show that the improved k-means clustering algorithm can further make cluster more stable.

Keywords: spectral clustering, Gaussian kernel, density sensitive, similarity measure, shortest path, k-means.

1 Introduction

The clustering is that the sample divided into many different clusters, the obtained clusters after division should meet the samples in the same cluster has a relatively high similarity degree, the similarity of samples in different clusters vary greatly [8, 9, 17].

[*] Corresponding author

[1] University of California Irvine(UCI for short), and the website is http://archive.ics.uci.edu/ml/

X. Luo, J.X. Yu, and Z. Li (Eds.): ADMA 2014, LNAI 8933, pp. 641–654, 2014.

Cluster analysis technique can be used at all stages of data mining, such as data preprocessing stage, the demand for data, you can have a complex structure of multidimensional data clustering process by making complex structure data standardization, and thus provide data mining pretreatment for other methods of data mining. Cluster analysis can dig out hidden natural classification in the datasets, and the data is divided into different clusters [5,16,18]. Traditional clustering method usually does not require a priori information, it is unsupervised learning. Selecting a suitable method of similarity measure in cluster analysis is crucial, it is used as the basis for division. And a clustering method using different similarity measure is likely to produce different clustering results, even if the same method of similarity measure, when scale parameters are set differently, it would also lead large differences in the clustering results. Because data mining is widely used, its application has been optimistic about the prospects, clustering technology becomes a hot research topic. Therefore, people continue to strengthen the academic study of various clustering algorithms, their work has a very important significance.

Traditional clustering algorithms, such as k-means algorithm [6, 12], EM algorithm [2], which are based on the assumption that the sample space are convex and spherical. When the sample space is not convex, the algorithm is easy to fall into local optimum, so the application of these algorithms is limited. In recent years, spectral clustering algorithm as a new clustering technique attracts the researcher attentions and becomes a hot topic of machine learning, pattern recognition and other areas [19]. It is known that spectral clustering are based on dividing the spectrum theory. Compared with other traditional clustering technology, it can find clustering in the sample space with randomly distributed clustering structure, and eventually converges to the global optimal solution. According to the feature vector data, spectral clustering constructs a more simple data space during the implementation process, it not only reduces the dimension of the sample data, but also makes the distribution structure of the sample data become clearer in the subspace.

Spectral clustering method by solving the Laplacian matrix of feature vectors, and then execute the selected feature vectors division to identify the type of non-convex clusters [15, 20]. The implement of the algorithm is also simple. The efficiency of spectral clustering only concerns the number of data points, rather than the dimension of datum, so the algorithm can avoid the curse of dimensionality caused by high-dimensional. Spectral clustering demonstrated an excellent clustering effect in numerous experiments and applications, and its performance is better than those traditional clustering algorithms, moreover it can handle large data sets. Therefore, the current spectral clustering applications such as image and video segmentation, speech recognition, text mining and so on.

Traditional spectral clustering algorithms typically use Gaussian kernel function as a similarity measure, it needs to introduce a parameter σ . During creating a similarity matrix **W**, the raw spectral clustering is often based on Euclidean distance, but it is impossible to accurately reflect the complexity of the data distribution [22, 23, 24, 25, 26, 27, 28, 29]. Inspired by the algorithm proposed by Wang Lin [13, 14], this paper makes it to increase the distance of pairs of data in the high density areas which are located in different rooms, while can reduce the similarity degree between the pairs of data in the same density region, so can find the spatial distribution characteristics of complex data. Meanwhile, in order not be affected by parameter σ, the Gaussian kernel function is not introduced in the design of similar matrix, so do not need to set any parameters and it is not affected by parameter σ. Distance can be directly used to calculate the similarity between data points, but also achieve the effect of scaling the similarity, and clustering results obtained are more stable. In addition, we all know that the time complexity of shortest path algorithm is generally $\mathbf{O}(n^3)$ when solving figure issues, even though there are a lot improved fast algorithms, the time complexity is still very high.

When the data collection is increasing, the usual computer running the algorithm will not be executed, and data mining is often facing a huge data sets [30, 31, 32, 33]. Based on these reasons, this paper proposed another simple method of similarity measure, which does not need to calculate the shortest path, and does not introduce any parameter. Finally, the last step of traditional spectral clustering algorithm is completed by k-means algorithm. In the experiments, for a certain σ value to do many times experiments, the obtained clustering results (although the difference of clustering results is unapparent) is often different, that is caused by the k-means algorithm. Based on the study k-means algorithm and related improvements, this paper designs a simple and effective method of selecting initial cluster centers. The method is only to do a simple change based on the traditional k-means that is increasing random number (*e.g.* 50) in the step of random initializing the clustering center. For each randomly initialized k cluster centers, computing European cluster each other. Then save and choose the largest distance of k clustering centers as the initial clustering center. Applying the obtained clustering centers in the proposed clustering algorithm, it can make the clustering effect greatly improved, while maintaining the clustering results stable.

The remainder of the paper is arranged as follows. We review the previous work on the applications of clustering in Section 2. We give the detail of the proposed method in Section 3. Furthermore, we analyze the experimental results in Section 4, and give our conclusion in Section 5.

2 Related Work

Clustering analysis method is one of the main analytical methods in data mining, and it has been be studied by many researchers. We introduce the related research work as follows:

Guha proposed a new clustering algorithm called CURE that is more robust to outliers, CURE achieves this by representing each cluster by a certain fixed number of points [3]. Nearest neighbor consistency is a central concept in statistical pattern recognition, Ding and He extended this concept to data clustering, requiring that for any data point in a cluster, its k-nearest neighbors and mutual nearest neighbors should also be in the same cluster. And they proposed kNN and kMN consistency enforcing and improving algorithm that indicates the local consistency information helps the global cluster objective function optimization [1]. Gelbard compared different clustering methods using several datasets, and proposed a novel method called

Binary-Positive clustering, which is a kind of hierarchic cluster algorithm. It adjusted to use binary datasets and developed Binary-Positive similarity measures, and shows that converting raw data into Binary-Positive format will improve clustering accuracy and robustness, especially while using hierarchical algorithms [4]. Zhang Proposed an algorithm for predicting item's long-term popularity through influential users, whose opinions or preferences strongly affect that of the other users. They employed the k-means clustering algorithm and clustering smoothing model for recommender systems has shown extremely performance [21]. Michael present a tabu search based clustering algorithm, to extend the k-means paradigm to categorical domains, and domains with both numeric and categorical values [7]. As is the case with most data clustering algorithms, the algorithm requires a presetting or random selection of initial points (modes) of the clusters. The differences on the initial points often lead to considerable distinct cluster results. Sun present an experimental study on applying Bradley and Fayyad iterative initial-point refinement algorithm to the k-modes clustering to improve the accurate and repetitiveness of the clustering results [10].

3 Method

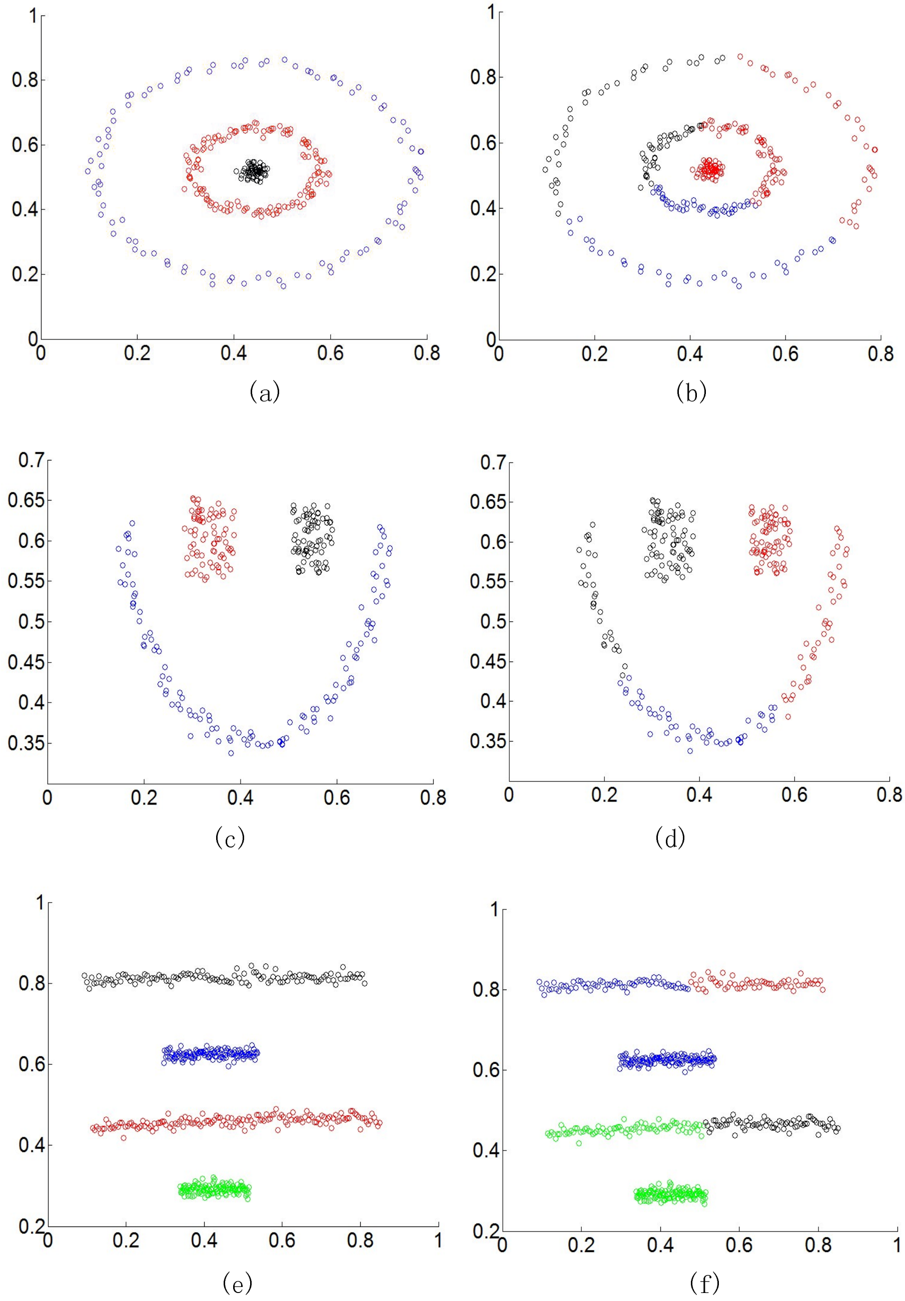
In this part, we analysis the spectral clustering from our sub experimental results and point out the existed problem, and then introduced the relevant knowledge involved in this paper, such as the similarity measure and k-means, lastly, the improved algorithm we design.

3.1 Spectral Clustering Method

The spectral clustering algorithm is a matching algorithm. The algorithm has three main processes:

The first step: calculate feature values and feature vectors of the sample similarity matrix; the second step: select the appropriate feature vector; the third step: use the k-means algorithm (or other traditional clustering algorithms) to cluster finally feature vector. Mainly, spectral clustering and the k-means clustering algorithm have a large difference is because that the ultimate objects of clustering are not different. Feature vector is selected clustering in spectral clustering, but k-means clustering is performed directly on the sample.

The following sub experiments is used to compare differences in the two algorithms on the clustering effect. The used data are three dedicated individuals working data sets (total six data sets) for studying the spectral clustering. In the Fig 1, (a), (c), (e) is the results of the spectral clustering algorithm, and (b), (d), (f) is the result of using the k-means clustering algorithm. It should be noted that the spectral clustering algorithm parameter σ is set to 0.01 (the value of σ is determined by many experiments). Through the left graphs, they show the perfect clustering results of spectral clustering. That the k-means algorithm have a limitation is because that it is based on a simple distance measure, which easy to fall into local optimal solution. So it is impossible to cluster these three artificial data sets correctly.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 1. The result of clustering on Spectral clustering and k-means algorithm

Through the above experimental results shows that, although the spectral clustering algorithm can find three dedicated individuals working data sets to cluster correctly, but the value of σ is needed for high demands. But the parameter σ usually needs to be set manually, and the different values of σ corresponding the clustering results will vary greatly, therefore, we need to make many times experiments to compare for selecting the corresponding σ value of the optimal clustering results. Spectral clustering find the right clustering results when σ is set of 0.01 in the Fig 1. Noted that the same values σ may belong to the different data sets that corresponding to the correct clustering is usually not the same, which related to the size of data attribute. Following by the experiments on UCI benchmark data sets shows these problems. This part of the experiment using the UCI data sets given in the Table 1.

The experiments using Normalized Mutual Information [11] (NMI) to evaluate the quality of the final clustering, and to achieve the level of quality to measure the accuracy of the matching index ACCURACY (ACC). The value of NMI (given in Eq. (4)) and ACC is bigger, the clustering quality is better. In our experiment, set σ =0.01, 0.05, 0.1, 0.5, 1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, and make a repeat experiment for the dataset (each dataset set corresponding the values of σ , and then run 50 times), finally output the average value.

From Fig. 2 and Fig. 3, we can see clustering efficiency on the four data sets relatively large difference by spectral clustering algorithm. And the clustering effect by the parameter of σ impact is large relatively on the same data set. So we know spectral clustering algorithm on the performance is not enough stable. In order to reduce the influence of this parameter, and get a more stable spectral clustering algorithm, this paper constructs a similarity matrix to improve the traditional spectral clustering.

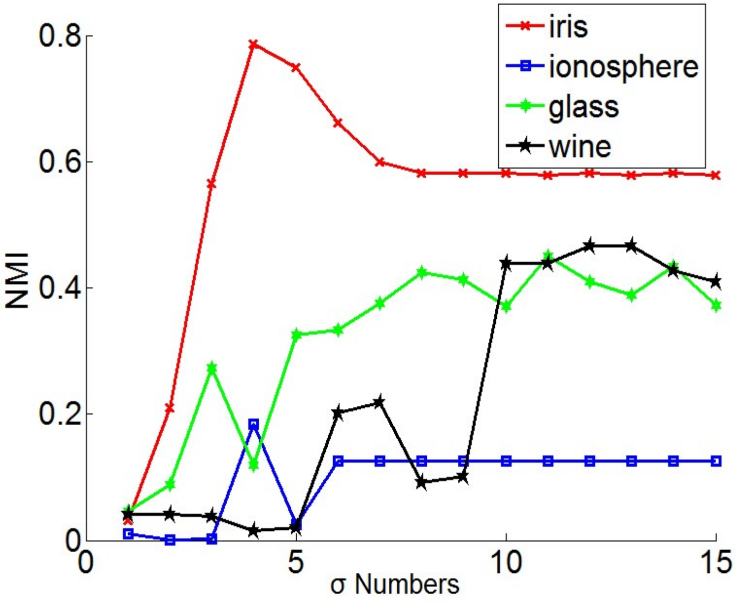


Fig. 2. Impact on the clustering performance parameters σ NMI

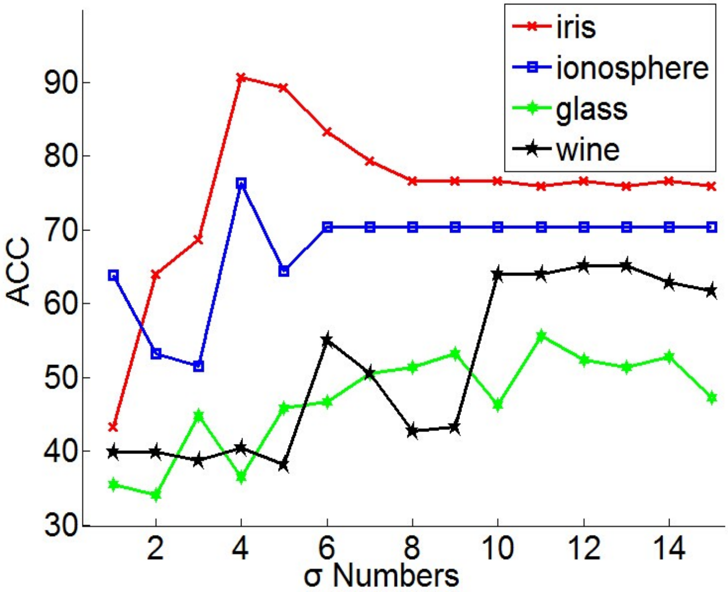


Fig. 3. Impact on the clustering performance parameters σ ACC

3.2 Similarity Measure

In the process of creating a similarity matrix **W**, the traditional spectral clustering usually calculated the similarity by Gaussian kernel based on Euclidean distance, but it can't reflect the complexity of the data distribution accurately. Therefore, this paper proposes that improving the similarity measure based on data density during creating the similarity matrix **W**. Making it to increase the distance of pairs of data in the high density areas which are located in different rooms, while can reduce the similarity degree between the pairs of data in the same density region, so can find the spatial distribution characteristics of complex data. The mainly step is to find the space distribution characteristics. Furthermore, in order to the clustering result is not affected by the parameter σ during design the similarity matrix, we don't introduce the Gaussian kernel function in the new algorithm, therefore, the clustering results are more stable and no affected by the parameter.

Firstly, we need to define the local line length:

$$\mathbf{L}(\mathbf{x}_i, \mathbf{x}_j) = e^{d^2(\mathbf{x}_i, \mathbf{x}_j)} - 1 . \tag{1}$$

Where $d(\mathbf{x}_i, \mathbf{x}_j)$ is the Euclidean distance between the data points $\mathbf{x}_i$ and $\mathbf{x}_j$.

We regard the point as a vertex **V** in weighted undirected graph $\mathbf{G}=\{\mathbf{V},\mathbf{E}\}$, the set of edge represents the connection weights between each pair of data points during the calculation. We regard $p \in V^l$ as the path of length $l=|p|-1$, which connected between the point p_1 and p_2. And p_{ij} represents $(p_k, p_{k+1}) \in \mathbf{E}$ the set of all path that connect the data points of $\{\mathbf{x}_1, \mathbf{x}_2\}$, where $(1 \le i, j < n)$. And the new Similarity distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ defined as follows:

$$\mathbf{W}_{ij} = \frac{1}{\text{In}(1+d_m(\mathbf{x_i},\mathbf{x_j}))+1}. \tag{2}$$

Where $d_m(\mathbf{x}_i,\mathbf{x}_j) = \min\limits_{p \subset p_{i,j}} \sum_{k=1}^{|p|-1} (e^{d^2(p_k, p_{k+1})} - 1)$, $d_m(\mathbf{x}_i,\mathbf{x}_j)$ is the shortest path distance between points $\mathbf{x}_i$ and $\mathbf{x}_j$. And $d\ (p_k, p_{k+1})$ is the Euclidean distance $\mathbf{x}_i$ and $\mathbf{x}_j$ in graph the shortest path between any two adjacent points.

Traditional spectral clustering algorithms regard Gaussian kernel function as a typically similarity measure, but it needs to introduce a parameter σ. Similarity measure mentioned without introducing the kernel functions in this paper, so there is no need to set any parameters. And not only directly used the distance to calculate the similarity between the data points, but also reached a scaled similarity effect.

It is generally known that the time complexity of the shortest path algorithm to solve the issue of graph is $\mathbf{O}(n^3)$, and the time complexity is very high. When the data collection is increasing, the algorithm will not be executed in PC. Moreover, data mining usually need to face large data sets, so this paper proposed another simple method of similarity measure, without calculate the shortest path.

In Similarity measure, the similarity distance is defined as follows:

$$\mathbf{W}_{ij} = \frac{1}{\text{In}(1+d(\mathbf{x}_i,\mathbf{x}_j))+1}. \tag{3}$$

Where $d(\mathbf{x}_i,\mathbf{x}_j)$ is the Euclidean distance $\mathbf{x}_i$ to $\mathbf{x}_j$.

3.3 k-Means Method and the Optimization Clustering Method

From the foregoing, the final step of the traditional spectral clustering algorithm is completed by the k-means algorithm to cluster. Do many times experiments of spectral clustering for a certain σ value, they often get different clustering results. Due to the k-means algorithm is dependent on initial cluster centers seriously, however, the initial cluster centers usually randomly generated, so each cluster center is not the same at the start of the basic algorithm, and which may produce the unstable clustering results.

The pseudo of the traditional k-means algorithm as follows:

```
Algorithm 1: k-means method
Input: Contains n data members to be clustering of the dataset X and the number of clusters k.
output: k clusters.

1: Preprocessing the dataset X,
2: Choose from the dataset X random produce k data members as the initial cluster centers,
3: Calculate the Euclidean distance all data objects in the dataset X to each cluster center,
   the data object is divided into the cluster with a minimum Euclidean distance,
4: Separately calculate the average of in each class all the data objects, we regard these
   average values as the new cluster center class,
5: Until the criterion function is convergent or the cluster centers is no changed.
```

The pseudo of improved spectral clustering as follows:

```
Algorithm 2: The improved spectral clustering
Input: $n$ data points$\{ x_i \}_{i=1}^{n}$ .
Output: The divided of data points $C_1, C_2, \cdots C_k$ 。

1: Construct the matrix $\mathbf{W} \in R^{m \times n}$ by the similarity measure based on data density
   similarity matrix w. In matrix, any element of can be represent $\mathbf{W}_{ij} = \frac{1}{\text{In}(1 + d_m(x_i, x_j)) + 1}$
   or $\mathbf{W}_{ij} = \frac{1}{\text{In}(1 + d(x_i, x_j)) + 1}$ ,where the element $\mathbf{W}_{ii} = 0, \ 1 \le i, j < n$ on the diagonal.
2: To construct Laplacian matrix $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ , where D is a diagonal matrix,
   $\mathbf{D}_{ii} = \sum_{j=1}^{n} \mathbf{W}_{ij}$ .
3: Seeking the feature vectors $v_1, v_2, \cdots v_k$ corresponding to the k largest feature value in a Laplacian
   matrix L, and to constructing matrix $\mathbf{V} = [v_1, v_2, \cdots v_k] \in R^{n \times k}$ , where V is the column vector.
4: Unitization the row of V, and get matrix Y , where $\mathbf{Y}_{ij} = \mathbf{V}_{ij} / (\sum_j \mathbf{V}_{ij}^2)$ .
5: Each row of Y is regard as a point in $R^k$ , which be clustered into k classes by k-means
   algorithm to optimize the initial cluster centers.
6: If the $i$-th row of Y belongs to class $j$, the original data points are classified into class $j$.
```

According to study the k-means algorithm and its relevant improved methods, we designed a simple and effective method of selecting initial cluster centers. The method is only to do a simple change based on the traditional k-means that is increasing random number (*e.g.* 50) in the step of random initializing the clustering center. For each randomly initialized k cluster centers, computing European cluster each other. Then save and choose the largest distance of k clustering centers as the initial clustering center. In result, the cluster center made clustering results stable and making the clustering result has been greatly improved.

The core of the improved algorithm as follow: random choose k data objects as the initial cluster centers from the dataset **X**; calculate the Euclidean distance between k clustering centers; and then repeat selected k data objects randomly; calculated Euclidean distance between k clustering centers again. If the distance is larger than the last time, we need to save the k cluster centers and the corresponding distance, or without modification for the next random selection until reach the set random number. Eventually, the algorithm will get a better initial cluster center. The pseudo is given in algorithm 2.

4 Experimental Analysis

We coded the algorithm with MATLAB 7.10.1（R2010a）in windows 7 system. The used datasets mainly came from UCI, and showed the detail of the datasets in Table 1. In order to evaluate the quality of the clustering by using the Normalized Mutual Information (NMI) , and matching accuracy index of the extent of the clustering quality measure: ACC.

First of all, let **X** and **Y** are respectively the clustering results p_a, and random variables of the predetermined categories results p_b, and **H**(**X**) and **H**(**Y**) are **X** and **Y** entropy, let **I**(**X**, **Y**) for the mutual information between **X** and **Y**, we can find that **I**(**X**, **Y**) is no limit, and the **H**(**X**) =**I**(X, X), so in the literature of **I**(**X**, **Y**) have been standardized, finally get the NMI as follows:

$$\mathrm{NMI} = \frac{\mathbf{I}(\mathbf{X},\mathbf{Y})}{\sqrt{\mathbf{H}(\mathbf{X})\mathbf{H}(\mathbf{Y})}}. \tag{4}$$

Table 1. Detailed information of benchmark data

Dataset	Wine	Iris	Ionosphere	glass
Instances	178	150	351	214
Features	13	4	34	9
classes	3	3	2	6

In order to exhibit the reliability, advantages and disadvantages of the two similarity measures, and the improvement effect of the k-means algorithm, the experiments were divided into three sub experiments, to facilitate comparison.

Experiment 1: algorithm using the similarity measure of Eq. (2), and the step (5) with the traditional k-means algorithm.

Experiment 2: algorithm using the similarity measure of Eq. (3), and the step (5) with the traditional k-means algorithm.

Experiment 3: algorithm adopts the similarity measure of Eq. (3), and the step (5) with the improved k-means algorithm.

Each part of the experiments were repeated 20 times, and then take the average of NMI and ACC. In addition, experiment 1, experiment 2 are recorded the algorithm running time (s) of each time, compared the shortest path effect on the running time of the algorithm.

(1) Experiment 1:

We summarized the clustering quality evaluation indicators results in Table 2. Control of the Table 2 and the Fig 2, Fig 3 in terms of the clustering quality indexes of NMI and ACC, we clearly found that the Fig 2 and Fig 3 reflects the stability of the traditional clustering algorithm exists some shortcomings during the clustering. And our improved method though conducted many times experiments, but, little difference between the results of each experiment, and the experimental result reached the highest value in the stable interval of the traditional spectral clustering algorithm in Fig 2, Fig 3. Especially, the iris dataset of NMI 0.7578 is the maximum among four datasets, the performance is evident.

Table 2. The test results in terms of two evaluation index value

Evaluate indicators	Wine	iris	Ionosphere	glass
NMI	0.4021	**0.7578**	0.1969	0.3553
ACC	71.7977	**90**	72.1368	49.3458
TIME	1510	**634.1531**	45000	3720

(2) Experiment 2:

In experiment 2, we obtained the clustering quality evaluation indicators shown in Table 3. Compared Table 3 with Table 2 in terms of NMI, ACC, TIME, we can easily find TIME in Table 2 is much larger than Table 3, the reason is that the algorithm in experiment 1 needs to calculate the shortest path, and the time complexity of computing for the shortest path algorithm requires too much time. And in terms of NMI and ACC, exception NMI and ACC of the iris dataset is same, the other data sets in Table 3 in terms of the index values were slightly smaller than the Table 2. This shows that the introduction of the shortest path in experiment 1 plays a good role, but not obvious. From the TIME point of view, experiment 2 algorithm is more suitable for practical application.

Table 3. The test results in terms of two evaluation index value

Evaluate index	Wine	iris	Ionosphere	glass
NMI	0.3976	**0.7578**	0.1963	0.3491
ACC	71.6292	**90**	72.0798	48.5514
TIME	0.9878	**0.8096**	5.4177	1.4794

In order to facilitate the experiment, experiment 3 also uses similarity measure, but not introduce the shortest path metric method, and compared with experiment 2.

(3) Experiment 3:

This part experiment, the introduced k-means algorithm to optimize the initial cluster center, and we proposed to use similarity measure methods in Eq. (3), the purpose is to further improve the stability of the improved algorithm. Comparing Table 4 and Table 3, we can see that the evaluation index of NMI and ACC have a small range increase. This is because of the k-means on the initial clustering center is dependent and may cause the clustering quality difference. So the improved k-means algorithm is introduced in this paper greatly, we can get the initial cluster center of high quality, diversity and the clustering result is not significant, so the average index evaluation experiments will be improved. Especially, the glass dataset improved most significantly, the dataset contains the most data categories, thus easily divided class mistake, so the traditional k-means behave more unstable. In Table 4, we can see that the clustering results of the glass data set ACC is least, this also shows that the instability of clustering, the clustering result is relatively more stable after the improved k-means algorithm is introduced than the clustering method in experiment 1 and experiment 2.

Table 4. The test results in terms of two evaluation index value

Evaluate index	Wine	iris	Ionosphere	glass
NMI	0.4122	**0.7632**	0.2135	0.3893
ACC	73.0337	**90.6667**	72.4986	52.8037

5 Conclusion

In view of the density sensitive similarity measure, we designed two methods to measure the similarity to improve the traditional spectral clustering algorithm. The experiments demonstrate the second method of similarity measure has a better performance than the first one. The proposed method solved the sensitive problem of Gauss kernel function parameter σ, the method is a non-parameter method. Lastly, the last stage of spectral clustering is sensitive to the initial clustering center of k-means algorithm to the selected feature vector clustering, so we designed an improved k-means algorithm that is very effective to optimize the initial cluster. The experimental result verified the improved spectral clustering method has better performance than other clustering methods.

References

1. Ding, C., He, X.: k-Nearest-Neighbor consistency in data clustering: Incorporating local information into global optimization. In: ACM Symposium on Applied Computing, pp. 584–589 (2004)
2. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data vis the EM algorithm. Journal of Royal Statistical Society Series B 39(1), 1–38 (1997)
3. Guha, S., Rastogi, R., Shim, K.: CURE: An efficient clustering algorithm for large databases. ACM SIGMOD Record 27(2), 73–84 (1998)

4. Gelbard, R., Goldman, O., Spiegler, I.: Investigating diversity of clustering methods: An empirical comparison. Data & Knowledge Engineering, 155–156 (2007)
5. Huang, Z.: Extensions to the k-means algorithm for clustering large datasets with categorical values. Data Mining and Knowledge Discovery 2, 283–304 (1998)
6. Jain, A.: Data clustering: 50 years beyond k-means. In: ICPR, pp. 651–666 (2010)
7. Michael, K., Joyce, C.: Clustering categorical data sets using tabu search techniques. Pattern Recognition 35, 2783–2790 (2002)
8. Queen, J.M.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkley Symposium Math. Stat. Prob., vol. 1, pp. 281–297 (1967)
9. Qin, Y., Zhang, S., Zhu, X., Zhang, J., Zhang, C.: Semi-parametric optimization for missing data imputation. Appl. Intell. 27(1), 79–88 (2007)
10. Sun, Y., Zhu, Q., Chen, Z.: An iterative initial-points refinement algorithm for categorical data clustering. Pattern Recognition Letters 23, 875–884 (2002)
11. Strehl, A., Ghosh, J.: Cluster ensembles-a knowledge reuse framework for combining partitioning‘s. Journal of Machine Learning Research 3, 583–617 (2002)
12. Wagstaff, K., Cardie, C., Rogers, S., Schroedl, S.: Constrained k-means clustering with background knowledge. In: ICML, pp. 577–584 (2001)
13. Wang, L., Bo, L., Jiao, L.: Density-Sensitive Semi-Supervised Spectral Clustering. Journal of Software 18(10), 2412–2422 (2007)
14. Wang, L., Bo, L., Jiao, L.: Density-Sensitive Spectral Clustering. Acta Electronica Sinica 35(8), 1577–1581 (2007)
15. Xiang, T., Gong, S.: Spectral clustering with eigenvector selection. Pattern Recognition 41(3), 1012–1029 (2008)
16. Wu, X., Zhang, S.: Synthesizing High-Frequency Rules from Different Data Sources. IEEE Trans. Knowl. Data Eng. 15(2), 353–367 (2003)
17. Wu, X., Zhang, C., Zhang, S.: Efficient mining of both positive and negative association rules. ACM Trans. Inf. Syst. 22(3), 381–405 (2004)
18. Wu, X., Zhang, C., Zhang, S.: Database classification for multi-database mining. Inf. Syst. 30(1), 71–88 (2005)
19. Zhang, S., Zhang, J., Zhu, X., Qin, Y., Zhang, C.: Missing Value Imputation Based on Data Clusteri ng. Transactions on Computational Science 1, 128–138 (2008)
20. Zhang, S., Chen, F., Wu, X., Zhang, C., Wang, R.: Mining bridging rules between conceptual clusters. Applied Intelligence 36(1), 108–118 (2012)
21. Zhang, J., Zhu, X., Li, X., Zhang, S.: Mining item popularity for recommender systems. In: Motoda, H., Wu, Z., Cao, L., Zaiane, O., Yao, M., Wang, W. (eds.) ADMA 2013, Part II. LNCS (LNAI), vol. 8347, pp. 372–383. Springer, Heidelberg (2013)
22. Zhang, S., Zhang, C., Yan, X.: Post-mining: maintenance of association rules by weighting. Inf. Syst. 28(7), 691–707 (2003)
23. Zhang, S., Qin, Z., Ling, C., Sheng, S.: "Missing Is Useful": Missing Values in Cost-Sensitive Decision Trees. IEEE Trans. Knowl. Data Eng. 17(12), 1689–1693 (2005)
24. Zhao, Y., Zhang, S.: Generalized Dimension-Reduction Framework for Recent-Biased Time Series Analysis. IEEE Trans. Knowl. Data Eng. 18(2), 231–244 (2006)
25. Zhu, X., Zhang, S., Jin, Z., Zhang, Z., Xu, Z.: Missing Value Estimation for Mixed-Attribute Data Sets. IEEE Trans. Knowl. Data Eng. 23(1), 110–121 (2011)
26. Zhu, X., Zhang, L., Huang, Z.: A Sparse Embedding and Least Variance Encoding Approach to Hashing. IEEE Transactions on Image Processing 23(9), 3737–3750 (2014)
27. Zhu, X., Huang, Z., Shen, H., Zhao, X.: Linear cross-modal hashing for efficient multimedia search. In: ACM Multimedia, pp. 143–152 (2013)

28. Zhu, X., Suk, H., Shen, D.: A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis. NeuroImage 100, 91–105 (2014)
29. Zhu, X., Suk, H., Shen, D.: Matrix-Similarity Based Loss Function and Feature Selection for Alzheimer's Disease Diagnosis. In: CVPR, pp. 3089–3096 (2014)
30. Zhu, X., Huang, Z., Yang, Y., Shen, H., Xu, C., Luo, J.: Self-taught dimensionality reduction on the high-dimensional small-sized data. Pattern Recognition 46(1), 215–229 (2013)
31. Zhu, X., Huang, Z., Cui, J., Shen, H.: Video-to-Shot Tag Propagation by Graph Sparse Group Lasso. IEEE Transactions on Multimedia 15(3), 633–646 (2013)
32. Zhu, X., Huang, Z., Cheng, H., Cui, J., Shen, H.: Sparse hashing for fast multimedia search. ACM Trans. Inf. Syst. 31(2), 9 (2013)
33. Zhu, X., Huang, Z., Shen, H., Cheng, J., Xu, C.: Dimensionality reduction by Mixed Kernel Canonical Correlation Analysis. Pattern Recognition 45(8), 3003–3016 (2012)