A New Clustering Algorithm Based on Local Purity and Class Mean Value

Jiemin Wu 15352337

***Abstract*—The purpose of data clustering is to divide elements into different categories according to their similarities. As a basic step of data analysis, it has a very wide range of applications. In this paper, a new clustering algorithm is proposed, which is based on the following two hypotheses: for a specific sample, a sample similar to it should belong to the same class as it. For a specific class, all samples belonging to this class should be as similar as possible to the class center. In this paper, we first give the definition of the symbol, then put forward our algorithm model and give the optimization method, finally by showing the algorithm in different datasets of the clustering effect to prove its feasibility.**

***Index Terms*—clustering, data analysis, unsupervised learning.**

I. Introduction

CLUSTERING is one of the basic experimental processes in data analysis. The use of clustering is extensive, and it is used in almost all natural and social sciences. In business, clustering can help market analysts separate consumer groups from the consumer database, and generalize the consumption patterns or habits of each type of consumer. On the other hand, as a module in data mining, it can be used as a separate tool to discover some deep information in the database, and to generalize the characteristics of each class. Furthermore, clustering analysis can be used as a preprocessing step for other analysis algorithms in Data mining algorithm.

The algorithm of clustering analysis can be divided into these categories, partitioning methods, hierarchical methods, density-based methods, grid-based methods, model-based Methods Given a dataset with n tuples or records, partitioning methods. The basic idea of partitioning methods is to regard the center of data points as the center of the corresponding cluster [1]. Typical partitioning methods include K-means [2], K-medoids [3], and CLARANS [4]. The basic idea of hierarchical methods is to construct the hierarchical relationship among data in order to cluster [5]. Typical hierarchical methods include BIRCH [6], CURE [7], and Chameleon [8]. The basic idea of density-based methods is that the data which is in the region with high density of the data space is considered to belong to the same cluster [9]. A fundamental difference between density-based methods and other methods is that it is not based on distance, but on density. Because the distance is symmetrical and the density is not, this can overcome the shortcoming of "circle-like" clustering which can only be found by the algorithm based on distance. Typical density-based methods include DBSCAN [10] and OPTICS [11].

In this paper, we present a new clustering algorithm based on the following two hypotheses: for a specific sample, a sample similar to it should belong to the same class as it. For a specific class, all samples belonging to this class should be as similar as possible to the class center. Like K-means, we cite Euclidean distances as a measure of similarity. Then we construct the corresponding objective function according to the above two hypotheses, and finally get the classification of all samples by optimizing the objective function. The experimental results show that our model can achieve a comparable mainstream model in clustering problem. Specifically, the main contributions of this paper can is summarized as follows:

- We propose a kind of clustering model framework which is different from the existing algorithm, by introducing the class standard matrix, constructing and optimizing the suitable objective function can make us get the reasonable clustering effect.
- We propose a clustering algorithm based on local purity and class mean, which not only can be used for clustering, but also can be used to optimize the results of other clustering algorithms.
- We conduct experiments on real-world datasets to evaluate the effectiveness of our clustering model. Experimental results show that the clustering effect obtained by this algorithm can be compared with the mainstream clustering algorithm.

In the following we start with some preliminaries . Then, we propose our clustering approach and report the experimental results. Finally, we conclude the paper .

II. Preliminaries

In this section we will give some definitions of the symbols, as well as the metrics used for similarity and the definition of neighbor points. These will be used to help us build the model of the algorithm.

A. Notation

Our problem is that now we have a dataset with the number of instances *N*, attribute is *M*, and we want to gather these data into *K* clusters based on the similarity between them. According to the above statement we can give the following symbols and their definitions.

data is the input data matrix($N \times M$), the number of instances *N*, attribute is *M*.

C is the class partitioning vector($N \times 1$), $C_i = k$ indicates that the i^th^ instance belongs to class k.

W is class indicator matrix($N \times K$), $W_{i,k}$ indicates the inclination of that i^th^ instance belongs to class k.

A is the adjacency matrix($N \times N$), $A_{i,j}$ is the Euclidean distance between i^th^ instance and j^th^ instance.

S is the neighbor indicator matrix, $S_{i,j} = 1$ indicates that i^th^ instance is a neighbor of j^th^ instance. Conversely if $S_{i,j} = 0$, then i^th^ instance and j^th^ instance are not neighbors.

e is the standard used to determine whether two points are neighbors, if the Euclidean distance between two points is less than e, that is, $D_{i,j} < e$, then they are neighbors, otherwise they are the opposite.

Suppose that we now have a point *p*, and the total number of all the neighbor points for *p* is *u*, where the number of neighbor points that belong to the same class as *P* is *g*.

λ is the learning rate of the algorithm.

B. Distance and Similarity

Distance (dissimilarity) and similarity are the basis for constructing clustering algorithms. In this paper, we use standard Euclidean distance as a measure of distance:

$$\left(\sum_{l=1}^{d}\left|\frac{x_{il}-x_{jl}}{S_l}\right|^2\right)^{\frac{1}{2}} \tag{1}$$

Based on the Euclidean distance between the two samples, we can determine the similarity of the two samples. The smaller the distance, the higher the similarity, and the greater the distance, the lower the similarity degree.

C. Neighbor

A point *q* with a distance of point *p* less than *e* is called a neighbor point *q* of *p*, and it is considered that the neighbor point of *p* has a similarity to *p*.

III. Formulation

We have mentioned two hypotheses above and we will write them in mathematical form. First, for a specific sample, a sample similar to it should belong to the same class as it. That is, for Point P, all of its neighbor points should be as good as the same class as it is, and according to the notation set above, we can use the mathematical form to describe the hypothesis as:

$$\max \frac{u}{g} \tag{2}$$

For all the input points, it's:

$$\max \sum_{i=1}^{N} \frac{u_i}{g_i} \tag{3}$$

$$\text{s.t. } \mathrm{u_i} = \sum_{j=1}^{N} (W_{j,C_i} * S_{i,j}) \tag{4}$$

$$\mathrm{g_i} = \sum_{j=1}^{N} S_{i,j} \tag{5}$$

$$\mathrm{S_{i,j}} = \begin{cases} 1, & D_{i,j} < 0 \\ 0, & D_{i,j} \geq 0 \end{cases} \tag{6}$$

We write the above formula in the form of a cost function, and take the logarithm of it for ease of calculation. Eventually we can get the following formula:

$$\min_{W} \sum_{i=1}^{N} \left[\log\left(\sum_{j=1}^{N} S_{i,j}\right) - \log\left(\sum_{j=1}^{N} W_{j,C_i} * S_{i,j}\right) \right] \tag{7}$$

$$s.t. \mathrm{S_{i,j}} = \begin{cases} 1, & D_{i,j} < 0 \\ 0, & D_{i,j} \geq 0 \end{cases} \tag{8}$$

Another assumption is that for a specific class, all samples belonging to this class should be as similar as possible to the class center. According to the definition of similarity above, we can express that the Euclidean distance between all the samples and the class center in a class should be as small as possible. We put this into a mathematical form:

$$\min \sum_{k=1}^{K} \sum_{j=1}^{n_k} \left(\sum_{l=1}^{d} \left| data_{j,l} - Center_{k,l} \right|^2 \right)^{\frac{1}{2}} \tag{9}$$

$$s.t.\ Center_k = \frac{1}{n_k} \sum_{j=1}^{n_k} data_j \tag{10}$$

n_k is the number of points that belong to the *k* class. Finally, our objective function can be written as follows:

$$\begin{aligned} \min_{W} \sum_{i=1}^{N} &\left[\log\left(\sum_{j=1}^{N} S_{i,j}\right) - \log\left(\sum_{j=1}^{N} W_{j,C_i} * S_{i,j}\right) \right] \\ &+ \sum_{k=1}^{K} \sum_{j=1}^{n_k} \left(\sum_{l=1}^{d} \left| data_{j,l} - Center_{k,l} \right|^2 \right)^{\frac{1}{2}} \\ s.t. \mathrm{S_{i,j}} &= \begin{cases} 1, & D_{i,j} < 0 \\ 0, & D_{i,j} \geq 0 \end{cases} \\ Center_k &= \frac{1}{n_k} \sum_{j=1}^{n_k} data_j \end{aligned} \tag{11}$$

IV. OPTIMIZATION

The model parameter set is $\{W, e\}$ while e is a super parameter that needs to be set artificially. So we need to determine the value of the W matrix after we find a suitable e. Gradient descent is used to optimize these parameters. But the latter part of the objective function (7)(8) does not explicitly use parameter W, we cannot use the batch gradient descent method, we then use an easy way to optimize it.

First for the formula (7) (8) We use the batch gradient descent method to solve the optimal parameter W. First we get the derivative for the parameter W:

$$\frac{\partial f}{\partial W_{i,k}} = \sum_{i=1}^{N} \frac{S_{i,j}}{\sum_{t=1}^{N} W_{t,k} S_{i,t}} \tag{12}$$

Therefore, the updated formula for gradient descent is:

$$\mathrm{W}_{i,k} = W_{i,k} - \lambda \sum_{i=1}^{N} \frac{S_{i,j}}{\sum_{t=1}^{N} W_{t,k} S_{i,t}} \tag{13}$$

For the later part of the objective function, we update using the following update formula:

$$\mathrm{W}_{i,k} = W_{i,k} - \lambda \sum_{j=1}^{n_k} \left(\sum_{l=1}^{d} \left| data_{j,l} - Center_{k,l} \right|^{2} \right)^{\frac{1}{2}} \tag{14}$$

To sum up, we write a comprehensive update formula on the parameter W:

$$\mathrm{W}_{i,k} = W_{i,k} - \lambda \left[\sum_{i=1}^{N} \frac{S_{i,j}}{\sum_{t=1}^{N} W_{t,k} S_{i,t}} + \sum_{j=1}^{n_k} \left(\sum_{l=1}^{d} \left| data_{j,l} - Center_{k,l} \right|^{2} \right)^{\frac{1}{2}} \right] \tag{15}$$

V. EXPERIMENTS

In this section we provide an overview of the datasets and methods which we will use in our experiments. Then we present an experimental analysis of our method compared with these methods on the datasets.

A. Datasets

We used some artificial datasets and three real-world datasets to evaluate the performance of our algorithm.

1) Artificial datasets

Artificial datasets include Flame dataset [12], Jain dataset [13], Path-based dataset [14], Compound dataset [15], and Aggregation dataset [16].

Flame dataset: N=240, K=2, M=2.
Jain dataset: N=373, K=2, M=2.
Path-based dataset: N=300, K=3, M=2.
Compound dataset: N=399, K=6, M=2.
Aggregation dataset: N=788, K=7, M=2.

2) Real-world datasets

Real-world datasets include Iris dataset, Wine dataset, Breast Cancer Wisconsin (Diagnostic) dataset [17].

Iris dataset: This is perhaps the best known database to be found in the pattern recognition literature. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

Wine dataset: These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The data set contains 3 classes of totally 178 instances. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

Breast Cancer Wisconsin (Diagnostic) dataset: The data set contains 2 classes of totally 569 instances. 32 kinds of features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

B. Baseline Methods

In this paper, we choose the classical Clustering algorithm K-means, DBSCAN, and Spectral Clustering [18] as baseline.

1) K-means

k-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

2) DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm proposed by Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu in 1996. It is a density-based clustering algorithm: given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors), marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away).

3) Spectral Clustering

In multivariate statistics and the clustering of data, spectral clustering techniques make use of the spectrum (eigenvalues) of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions. The similarity matrix is provided as an input and consists of a

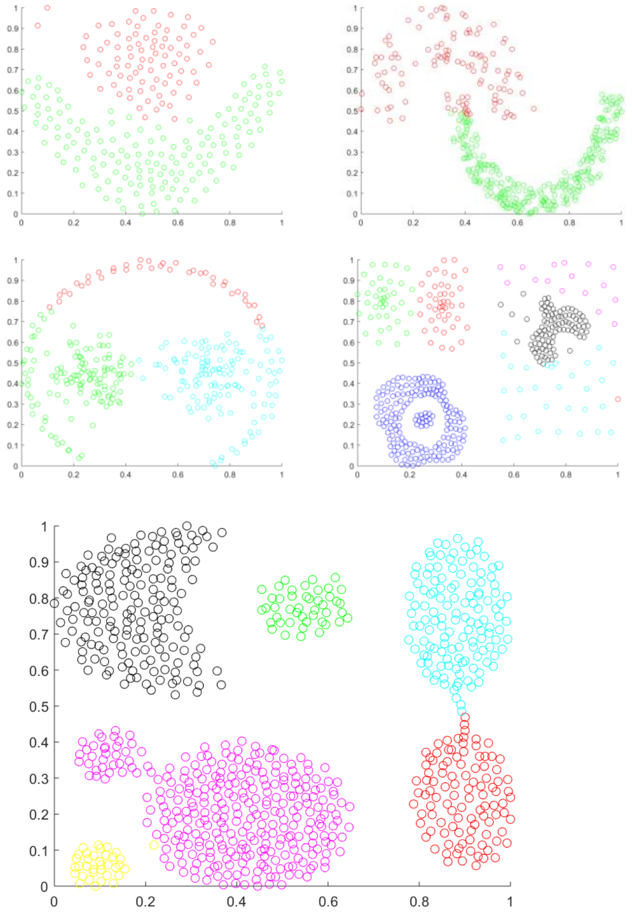

Fig. 1. This is the result of our algorithm on the artificial datasets. The first row in the first column of the graph is the Flame dataset, N=240, K=2. The first row in the second column of the graph is the Jain dataset, N=373, K=2.The second row in the first column of the graph is the path-based dataset, N=300, K=3. The second row in the second column of the graph is the compound dataset, N=399, K=6. The last one of the graph is the Aggregation dataset. N=788, K=7.

quantitative assessment of the relative similarity of each pair of points in the dataset.

C. *Experimental results*

First we evaluated our approach on the artificial datasets. Because the characteristic dimension of the artificial dataset is only 2 dimensions, we can give a visual display of the clustering results. Here are the clustering results in Fig. 1. We can see that the algorithm is performing well on the artificial data set, and can handle the flow data to some extent, and it will not appear that the clustering effect can only be spherical in the same way as K-means.

We then tested the accuracy(ACC) of our algorithm and baselines in artificial datasets and true datasets, standardized mutual information indices(NMI) [19], and purity. Here is the test result.

ACC(%)	wire	iris	wdbc	flame	jain	pathbased	compound	aggregation
K-means	0.9494	0.8867	0.9279	0.8583	0.882	0.7433	0.7719	0.8617
DBSCAN	0.7079	0.81	0.8506	0.6333	**0.9598**	**0.92**	**0.9373**	**0.9911**
Spectral Cluster	**0.97**	0.8733	**0.9297**	0.8417	0.8794	0.77	0.6591	0.7234
Our algorithm	0.9438	**0.9333**	**0.9297**	**0.9917**	**0.9598**	0.76	0.8922	0.9543

NMI(%)	wire	iris		flame	jain	pathbased	compound	aggregation
K-means	0.8374	0.7364	0.6115	0.4709	0.4926	0.5128	0.7921	0.7991
DBSCAN	0.4978	0.5935	0.3634	0.2639	**0.817**	**0.75**	**0.8885**	**0.9748**
Spectral Cluster	**0.8897**	0.7116	**0.6148**	0.415	0.5048	0.5446	0.6594	0.7439
Our algorithm	0.804	**0.8063**	**0.6148**	**0.932**	0.7444	0.51	0.8478	0.9125

Purity(%)	wire	iris	wdbc	flame	jain	pathbased	compound	aggregation
K-means	**0.9494**	0.8867	0.9279	0.8583	0.882	0.77	0.812	0.8617
DBSCAN	0.8202	0.8133	0.8506	0.7208	**0.9598**	**0.92**	**0.9373**	0.9911
Spectral Cluster	0.9719	0.8733	**0.9297**	0.8417	0.8794	0.77	0.6842	0.7652
Our algorithm	0.9438	**0.9333**	**0.9297**	**0.9917**	**0.9598**	0.76	0.9323	**0.9975**

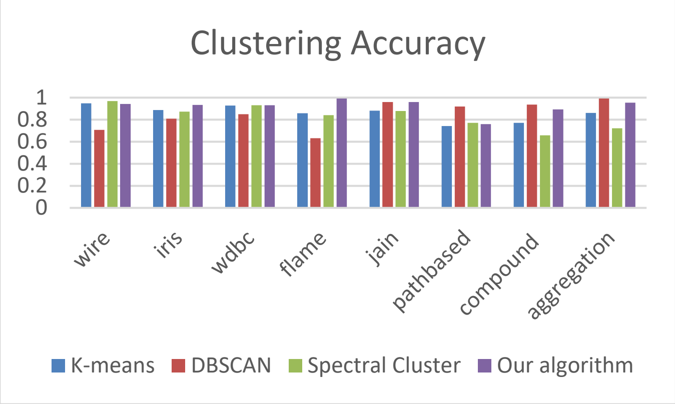

Fig. 2. This is the clustering accuracy rate of our algorithm and baselines in the artificial datasets and the real-world datasets.

The above results show that our algorithm can achieve the effect of the mainstream clustering algorithm in most cases, and can surpass them in some cases.

VI. Conclusion

In this paper, we propose a clustering algorithm based on local density and class mean value, which is built on the following two hypotheses. For a specific sample, a sample similar to it should belong to the same class as it. For a specific class, all samples belonging to this class should be as similar as possible to the class center. By establishing the appropriate mathematical model and proposing the corresponding optimization method, we can get the excellent clustering effect. Experiments on real-world datasets and shape datasets show that this algorithm can achieve very good performance. On the other hand, it can not only be used for clustering problem directly, but also can optimize the clustering effect of other algorithms. Finally, our future work should be in this new framework to optimize the stability of the algorithm output and the processing capacity for large data.

References

[1] Xu, D. & Tian, Y. A Comprehensive Survey of Clustering Algorithms. Ann. Data. Sci. (2015) 2: 165. https://doi.org/10.1007/s40745-015-0040-1

[2] MacQueen J (1967) Some methods for classification and analysis of multivariate observations. Proc Fifth Berkeley Symp Math Stat Probab 1:281–297

[3] Park H, Jun C (2009) A simple and fast algorithm for K-medoids clustering. Expert Syst Appl 36:3336–3341

[4] Ng R, Han J (2002) Clarans: a method for clustering objects for spatial data mining. IEEE Trans Knowl Data Eng 14:1003–1016

[5] Johnson S (1967) Hierarchical clustering schemes. Psychometrika 32:241–254

[6] Zhang T, Ramakrishnan R, Livny M (1996) BIRCH: an efficient data clustering method for very large databases. ACM SIGMOD Rec 25:103–104

[7] Guha S, Rastogi R, Shim K (1998) CURE: an efficient clustering algorithm for large databases. ACM SIGMOD Rec 27:73–84

[8] Karypis G, Han E, Kumar V (1999) Chameleon: hierarchical clustering using dynamic modeling. Computer 32:68–75

[9] Kriegel H, Kröger P, Sander J, Zimek A (2011) Densitybased clustering. Wiley Interdiscip Rev 1:231–240

[10] Ester M, Kriegel H, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the second ACM SIGKDD international conference on knowledge discovery and data mining, pp 226–231

[11] Ankerst M, Breunig M, Kriegel H, Sander J (1999) OPTICS: ordering points to identify the clustering structure. In: Proceedings on 1999 ACM SIGMOD international conference on management of data, vol 28, pp 49–60

[12] L. Fu and E. Medico, FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. BMC bioinformatics, 2007. 8(1): p. 3.

[13] A. Jain and M. Law, Data clustering: A user's dilemma. Lecture Notes in Computer Science, 2005. 3776: p. 1-10.

[14] H. Chang and D.Y. Yeung, Robust path-based spectral clustering. Pattern Recognition, 2008. 41(1): p. 191-203.

[15] C.T. Zahn, Graph-theoretical methods for detecting and describing gestalt clusters. IEEE Transactions on Computers, 1971. 100(1): p. 68-86.

[16] A. Gionis, H. Mannila, and P. Tsaparas, Clustering aggregation. ACM Transactions on Knowledge Discovery from Data (TKDD), 2007. 1(1): p. 1-30.

[17] Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[18] Jianbo Shi and Jitendra Malik, "Normalized Cuts and Image Segmentation", IEEE Transactions on PAMI, Vol. 22, No. 8, Aug 2000.

[19] Strehl, Alexander; Ghosh, Joydeep (2002), "Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions". The Journal of Machine Learning Research, 3 (Dec): 583–617