A Ranking Fuzzy Cluster Algorithm for Multidimensional Data

Guo, Jia

(出版者 / Publisher)
法政大学大学院情報科学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 情報科学研究科編 / 法政大学大学院紀要. 情報科学研究科編
(巻 / Volume)
11
(開始ページ / Start Page)
1
(終了ページ / End Page)
4
(発行年 / Year)
2016-03-24
(URL)
https://doi.org/10.15002/00012872

A Ranking Fuzzy Cluster Algorithm for Multidimensional Data

JIA GUO
Graduate School of Computer and Information Sciences
Hosei University
Tokyo 184-8584, Japan
jia.guo.ud@stu.hosei.ac.jp

***Abstract*—Data classification is an important problem in various scientific fields. Data analysis algorithm such as the fuzzy c-means algorithm and k-means algorithm are wildly used on multidimensional data but still suffers from several drawbacks. To get higher classification accuracy, the Ranking Fuzzy Cluster (RFC) algorithm which uses the method Dimension Ranking (DR) and a new center moving formula is proposed in this paper. The method DR is used to calculate the weight of each instance among all the data set in each dimension. The new center moving formula which uses the non-gradient descent iterative to avoid the local minimum problems of FCM. In the experiment, five benchmark data sets are used to test the classification accuracy of RFC. The fuzzy c-means algorithm and k-means algorithm are also used in the experiment comparison. Experimental result shows that the RFC has higher classification accuracy than fuzzy c-means and k-means with multidimensional data.**

Keywords—Fuzzy Cluster; Multidimensional Data; Dimension Ranking; Cluster Center Moving;

I. INTRODUCTION

Cluster analysis becomes an important problem since apes exist. What kind of fruit is able to eat, what kind of animal is easy to fight. These may be the first and coarsest cluster analysis. Thousands years later, the object of cluster analysis change from real objects into abstract data, and the one who makes the analysis changes from humans to machines. This reduces the accuracy of the classification and also to provide the possibility to handle large data.

Cluster analysis aims at identifying groups of similar objects, and helps to discover distribution of patterns and interesting correlations in large data sets [1]. The development of clustering methodology has been a truly interdisciplinary endeavor. Taxonomists, social scientists, psychologists, biologists, statisticians, mathematicians, engineers, computer scientists, medical researchers, and others who collect and process real data have all contributed to clustering methodology [2]. Now cluster analysis is widely used in machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics [3]. Decades ago, the term "k-means" was first used by James MacQueen in 1967 [4], though the idea goes back to Hugo Steinhaus in 1957 [5]. The standard algorithm was first proposed by Stuart Lloyd in 1957 as a technique for pulse-code modulation, though it wasn't published outside of Bell Labs until 1982 [6]. In 1965, E.W.Forgy published essentially the same method, which is why it is sometimes referred to as Lloyd-Forgy [7]. A more efficient version was proposed and published in Fortran by Hartigan and Wong in 1975/1979 [8][9].

In fuzzy clustering (also called soft clustering), different with hard clustering, each element belong to every cluster center and a membership level is used to tell the correlation between them. These indicate the strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters. Fuzzy C-Means (FCM) is one of the most widely used fuzzy cluster analysis algorithm, it still suffers from several drawbacks, like being very sensitive to noise data, great affected by initial value and so on. With the advances of technology, it was found that links between the data is not absolute. In fuzzy clustering, every point has a degree of belonging to clusters, as in fuzzy logic, rather than belonging to just one cluster. Thus, points on the edge of a cluster, may be in the cluster to a lesser degree than points in the center of cluster. An overview and comparison of different fuzzy clustering algorithms is available [10].

II. RELATED WORK

A. K-means Algorithm

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 partitions a data set by minimizing a sum-of-squares cost function. A coordinate descend method is then used to find local minima [11].

The problem is computationally difficult (NP-hard) [12]; however, there are efficient heuristic algorithms that are commonly employed and converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both algorithms. Additionally, they both use cluster centers to model the data; however, k-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

K-means clustering has been used as feature learning (or dictionary learning) step, in either (semi-)supervised learning or unsupervised learning [13]. The basic approach is first to train a k-means clustering representation, using the input training data (which need not be labelled). Then, to project any input datum into the new feature space, we have a choice of "encoding"

Supervisor: Prof. Yuji Sato

functions, but we can use for example the threshold matrix-product of the datum with the centroid locations, the distance from the datum to each centroid, or simply an indicator function for the nearest centroid [14], or some smooth transformation of the distance [15]. Alternatively, by transforming the sample-cluster distance through a Gaussian RBF, one effectively obtains the hidden layer of a radial basis function network [16].

B. Fuzzy C-means

Fuzzy C-Means (FCM) is an unsupervised clustering algorithm has a long history. FCM clustering algorithm proposed by Dunn [17] and then extended by Bezdek [18]. FCM attempts to find the most characteristic point in each cluster, which can be considered as the "centroid" of the cluster and, then, the grade of membership for each object in the clusters. Such aim is achieved by minimizing the objective function. A commonly used objective function is membership weighted within cluster error defined as follow:

$$\text{Minimize } J_m(U,V) = \sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}^{m} * ||x_j - v_i||^2 \qquad (1)$$

where n is the total number of instances in a given data set and c is the number of clusters; $X = \{x_1, x_2,...,x_n\} \subset R^s$ and $V = \{v_1,...,v_c\} \subset R^s$ are the feature data and cluster centroids; and $U = [u_{ij}]_{c\times n}$ is a fuzzy partition matrix composed of the membership grade of pattern x_j to each cluster i. $||x_j - v_i||$ is the Euclidean norm between x_j and v_i. The weighting exponent m is called the fuzzifier which can have influence on the clustering performance of FCM. Good choice of m makes the calculating accurate and fast. The cluster centroids and the respective membership functions that solve the constrained optimization problem in (1) are given by the following equations:

$$v_i = \frac{\sum_{j=1}^{n}(u_{ij})^m * x_j}{\sum_{j=1}^{n}(u_{ij})^m} \qquad (2)$$

$$u_{ij} = [\sum_{k=1}^{c}(\frac{||x_j - v_i||^2}{||x_j - v_k||^2})^{1/(m-1)}]^{-1} \qquad (3)$$

Equation (2) and (3) constitute an iterative optimization procedure. The goal is to iteratively improve a sequence of sets of fuzzy clusters until no further improvement in $J_m(U, V)$ is possible.

FCM is an efficient and accurate algorithm but still suffers from several drawbacks [19]. It is very sensitive to noise data and the initial value of cluster centers. Yang points out that the objective function is a non-linear multimodal function with a number of local minima. There is a possibility of getting stuck at local minima [20].

III. PROPOSAL OF RANKING FUZZY CLUSTER ALGORITHM

The Ranking Fuzzy Cluster Algorithm (RFC) is proposed in this section. The RFC is an unsupervised clustering algorithm which aims at finding the center of each cluster. The optimal solution of center makes instances in same class gathered close and instances in different classes are spares. Numerical speaking, RFC try to find the best cluster centers which make the objective function achieve its minimum. The objective function $J(U, V)$ defined as follows [18]:

$$J(U,V) = \sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}^{m} * ||x_j - v_i||^2 \qquad (4)$$

where n is the number of instances in a given data set and c is the number of cluster; $X = \{x_1, x_2,...,x_n\}$ and $V = \{v_1,...,v_c\}$ are the feature data and cluster centers; and $U = [u_{ij}]_{c\times n}$ is a fuzzy partition matrix composed of the membership grade of pattern x_j to each cluster i. $||x_j - v_i||$ is the Euclidean norm between x_j and v_i. The weighting exponent m is called the fuzzifier which can have influence on the clustering accuracy.

A. Dimension Ranking

As the objective function stands for the summation of distance between instances and cluster centers, it will be affected by the value of instances. Each instance has several dimensions in multidimensional data. When one or some dimensions' absolute value is much bigger than others, its influence on the objective function will be greater than the other dimensions. To solve this problem, the method Dimension Ranking (DR) is proposed by RFC to calculate the weight of each instance among all the data set in each dimension. The DR is defined in this paper as follows:

$$DR(i,j) = \left(\frac{D(i,j) - Q(j)}{P(j) - Q(j)}\right)^{m-1} \quad i \in [1,n], j \in [1,f] \qquad (5)$$

where n is the number of instances and f is the number of attributes, D is the given data set, Array $P = \{P_1, P_2,...,P_f\}$ and $Q = \{Q_1, Q_2,...,Q_f\}$ is the maximum and minimum values of instance's each feature, m is the weighted index and suitable choices of m can make the calculation more precise. On the basis of keeping the characteristic of original data, DR translates all data into a scaled number between 0 and 1.

B. Initial Cluster Centers

Cluster centers will be output as a result of clustering algorithm. Optimal centers make the objective function achieve the minimum value. Different with FCM setting the initial cluster centers randomly, RFC average divide 1 into c pieces, sets the centroids as the center. The amount of calculation can be reduced while starting calculating from the specified cluster centers. The initial center setting formula is defined in this paper as follows:

$$V(i,j) = \frac{2i-1}{2c} \qquad i \in [1,c], j \in [1,f] \qquad (6)$$

where c is the number of clusters and f is the number of attributes of each instance. Each feature of initial centers has the same value. Average designated cluster centers can simplify the calculation and makes the algorithm more stable. For the average distribution of data, this setting enables fast and accurate classification.

C. Membership Matrix U

The membership matrix *U* describes the membership of each instance corresponding to each cluster center. The *U* is calculated as follows [18]:

$$K(j,i) = \frac{1}{||x_i - v_j||} \qquad i \in [1,n] \tag{7}$$

$$U(j,i) = \frac{K(j,i)}{\sum_{j=1}^{c} K(j,i)} \qquad i \in [1,n], j \in [1,c] \tag{8}$$

where n is the number of instances in the given data set and c is the number of clusters, $DR = \{x_1, x_2,...,x_n\}$ and $V = \{v_1,...,v_c\}$ are the feature data and cluster centers, $||x_i - v_j||$ is the Euclidean norm between x_j and v_i. The matrix K is the reciprocal of the distance from instance to each cluster center. Equation (9) is a normalized translation. It makes the summation of one instance's membership to all cluster centers is 1.

D. Center Moving Formula

The new cluster center is the summation of the old cluster center and walking distance. The walking distance is calculated by the DR, *U* and old cluster. The center moving formula is defined in this paper as follows:

$$\Delta v_j = \frac{\sum_{i=1}^{n} (\frac{x_i - v_j}{2}) * U_{ij}^m}{n} \qquad i \in [1,n], j \in [1,c] \tag{9}$$

$$New\ v_j = v_j + \Delta v_j \qquad j \in [1,c] \tag{10}$$

where n is the number of instances in the given data set and c is the number of clusters, Δv_j is the moving distance, $DR = \{x_1, x_2,...,x_n\}$ and $V = \{v_1,...,v_c\}$ are the feature data and cluster centers, the weighting exponent m is called the fuzzifier which can have influence on the clustering accuracy. This formula aims at reducing the value of the objective function by center moving. It is shown in (5) that the distance from high value instance to center has higher influence to the objective function. The center moves more to instances has higher membership and all instances are considered in each movement in (9). This formula aims at reducing the distance between centers and instances according to the membership between them.

E. Process of RFC

The RFC algorithm is executed in the following steps:

Step 1: Given a pre-selected number of cluster c, a chosen value of m, a certain threshold ε, the maximum number of iterations k, a data set which has n instances and f dimension for each instance.

Step 2: Calculate the DR using (5), the initial fuzzy cluster center $V = \{v_1,...,v_c\}$ using (6), the initial fuzzy membership $U = [u_{ij}]_{c \times n}$ using (7) and (8), initial value of $J(U, V)$ using (4) .

Step 3: Calculating the new center using (9) and (10).

Step 4: Update the fuzzy membership *U using* (7) and (8). Recording the value of $J(U, V)$.

Step 5: If the improvement in $J(U, V)$ is less than ε or iterations is over k, then halt and output the final cluster centers; otherwise go to step 3.

TABLE I. EVALUATION TEST ENVIRONMENT

OS	Windows 7
Processor	Intel® Core™ i7
Memory	16G
System type	Microsoft Corporation
Tool	MATLAB R2015a, Python 2.7

TABLE II. BENCHMARK DATA SET DESCRIPTION

Data Set Name	Instances	Dimension	Cluster Number	Type
Iris	150	4	3	Life
Wine	178	13	3	Physical
Seeds	210	7	3	Life
Lung Cancer	27	56	3	Life
Contraceptive Method Choice	1473	9	3	Life

IV. EXPERIMENT

A. Experimental Method

Five benchmark data base, Iris, Wine, Seeds, Lung Cancer, Contraceptive Method Choice, are used in the experiment in order to verify the accuracy of classification of RFC. Comparative experiment is done by the Fuzzy C-Means (FCM) algorithm and K-means algorithm. All benchmark data sets are from the University of California, Irvine Machine Learning Repository.

Value for each parameter: threshold $\varepsilon=10^{-5}$; the maximum number of iterations k=100 and weighting exponent m=2. The test environment is shown in Table I. The description of all data sets is shown in Table II.

B. Experimental Results

FCM and K-means algorithm have a same correct rate at 89.33% when facing data set Iris. The RFC gives a better accuracy at 92.67% at Iris. FCM and K-means give the correct rate at 68.54% and 56.74% when facing data set wine, RFC gives a 72.46% correct rate. Three fuzzy cluster algorithms give the same accuracy at 89.52% when facing the data base Seeds. FCM and K-means gives similar result, accuracy at 62.96% and 59.25%, when facing data set Lung Cancer. RFC has 48.15% correct rate dealing with Lung Cancer. FCM and K-means have a correct rate at 39.17% and 39.37% when facing the last data set in this experiment, Contraceptive Method Choice. RFC has the higher correct rate at 40.80% when facing the Contraceptive Method Choice.

Specially, to the data set Seeds, RFC can give 189 instances correct classification in all 210 instances after 122 iterations. The max iterations number in this experiment is 100. All results of the experiment are shown in Table III.

TABLE III. EXPERIMENTAL RESULT

	Accuracy of K-means	Accuracy of FCM	Accuracy of RFC
Iris	134/150	134/150	139/150
Wine	101/178	122/178	129/178
Seeds	188/210	188/210	188/210
Lung Cancer	16/27	17/27	13/27
Contraceptive Method Choice	580/1473	577/1473	601/1473

C. Discussion

The result of experiment shows that RFC has higher classification accuracy than FCM and k-means when facing data set Iris, Wine and Contraceptive Method Choice. This is because the center calculating formula of FCM only considers about the instance and membership, but the center moving formula of RFC also considers about the distance from centers to all instances. This formula reduce the distance between centers and instances according to the membership between them. Different with FCM using the gradient descent, the center moving formula of RFC considers the influences of all instances while calculating. This makes the RFC has higher classification accuracy than FCM and k-means when facing data set Iris, Wine and Contraceptive Method Choice.

These three algorithms have same classification accuracy at 89.52% when facing the data set Seeds. This means the three kinds of algorithm are all very suitable at this data set. Specially, RFC gets the 188 instances right in all 210 instances after 100 iterations and it can improve the number to 189 after the 122 iterations. This shows that the center moving formula can make the objective function reach the minimum value and the calculating speed can be improved in future work.

To the data set Lung Cancer, RFC gives lower classification accuracy than FCM and k-means. This is because this data set has missing value. Five instances with missing value are deleted from the data set, this leads the low classification accuracy of RFC. The center moving formula of RFC considers the distance from all instances to cluster centers, which means RFC may affected by missing value very much. How to reduce the impact of missing value is one of the future points.

V. CONCLUSION

In this thesis, we have proposed a Ranking Fuzzy Cluster (RFC) algorithm which uses the method Dimension Ranking (DR) and a new center moving formula. Five benchmark data sets are used in experiment in order to verify the accuracy of classification of RFC. Specially, Fuzzy c-means (FCM) algorithm and k-means algorithm are also used in the experiment as comparison. The experiment result shows that RFC can give higher accuracy of classification than FCM and k-means when facing multidimensional data.

VI. ACKNOWLEDGEMENTS

The work of this thesis is under the guide of my supervisor, Professor Yuji Sato. His scientific methods of work and tirelessly guidance give me a great deal of help. His rigorous academic attitude and extraordinary wisdom give me a lot of guidance. He provided a number of key recommendations to help me fulfill this thesis. And I would like to show my deepest gratitude to Mrs. Shiqin Yang for her guidance enlightening and detailed inspection. This thesis wouldn't be completed without her help. There for I would like to show my deepest gratitude to Professor Yuji Sato and Mrs. Shiqin Yang.

REFERENCES

[1] Wang, Weina, and Yunjie Zhang. "On fuzzy cluster validity indices." Fuzzy sets and systems 158.19 (2007): 2095-2117.

[2] Jain, Anil K. "Data clustering: 50 years beyond K-means." Pattern recognition letters 31.8 (2010): 651-666.

[3] Ahmed, Mohamed N., et al. "A modified fuzzy c-means algorithm for bias field estimation and segmentation of MRI data." Medical Imaging, IEEE Transactions on 21.3 (2002): 193-199.

[4] MacQueen, James. "Some methods for classification and analysis of multivariate observations." Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Vol. 1. No. 14. 1967.

[5] Steinhaus, Hugo. "Sur la division des corp materiels en parties." Bull. Acad. Polon. Sci 1 (1956): 801-804.

[6] Lloyd, Stuart P. "Least squares quantization in PCM." Information Theory, IEEE Transactions on 28.2 (1982): 129-137.

[7] Forgy, Edward W. "Cluster analysis of multivariate data: efficiency versus interpretability of classifications." Biometrics 21 (1965): 768-769.

[8] Hartigan, John A. Clustering algorithms. John Wiley & Sons, Inc., 1975.

[9] Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." Applied statistics (1979): 100-108.

[10] Nock, Richard, and Frank Nielsen. "On weighting clustering." Pattern Analysis and Machine Intelligence, IEEE Transactions on 28.8 (2006): 1223-1235.

[11] Zha, Hongyuan, et al. "Spectral relaxation for k-means clustering." Advances in neural information processing systems. 2001

[12] Pardalos, Panos M., and Stephen A. Vavasis. "Quadratic programming with one negative eigenvalue is NP-hard." Journal of Global Optimization 1.1 (1991): 15-22.

[13] Coates, Adam, and Andrew Y. Ng. "Learning feature representations with k-means." Neural Networks: Tricks of the Trade. Springer Berlin Heidelberg, 2012. 561-580.

[14] Csurka, Gabriella, et al. "Visual categorization with bags of keypoints." Workshop on statistical learning in computer vision, ECCV. Vol. 1. No. 1-22. 2004.

[15] Coates, Adam, Andrew Y. Ng, and Honglak Lee. "An analysis of single-layer networks in unsupervised feature learning." International conference on artificial intelligence and statistics. 2011.

[16] Schwenker, Friedhelm, Hans A. Kestler, and Günther Palm. "Three learning phases for radial-basis-function networks." Neural networks 14.4 (2001): 439-458.

[17] Dunn, Joseph C. "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters." (1973): 32-57.

[18] Bezdek, James Christian. "Fuzzy mathematics in pattern classification." (1973).

[19] Vasuda, P., and S. Satheesh. "Improved Fuzzy C-Means algorithm for MR brain image segmentation." International Journal on Computer Science and Engineering 2.5 (1713): 2010.

[20] Yang, Shiqin, and Yuji Sato. "Fuzzy Clustering with Fitness Predator Optimizer for Multivariate Data Problems." Simulated Evolution and Learning. Springer International Publishing, 2014. 155-166.