Improved FCM algorithm for Clustering the IRIS data

K.Suresh[1] R.Madana Mohana[2] A.RamaMohanReddy[3]

[1] Department of Software Engineering, East China University of Technology,ECIT Nanchang Campus,
Nanchang, Jiangxi-330013, P.R.China.

[2] Department of Information Technology, Vardhaman college of Engineering,
Shamshabad, Hyderabad, A.P, India.

[3] Department of Computer Science and Engineering , S.V.University College of Engineering,
Tirupati, A.P, India.

Abstract

In this paper we present clustering method is very sensitive to the initial center values, requirements on the data set too high, and cannot handle noisy data the proposal method is using information entropy to initialize the cluster centers and introduce weighting parameters to adjust the location of cluster centers and noise problems. The improves clustering on web data efficiently using fuzzy c-means(FCM)clustering with iris data sets.

Keywords: Datasets, clutering, improved FCM clustering, webusage mining.

1. Introduction

One of the most challenging analysis problems in the data mining domains is organizing large amounts of information. One approach to this problem is to cluster information based on the content of a collection of documents. Clustering is a widely used technique in data mining application for discovering patterns in underlying data. Most traditional clustering algorithms are limited in handling datasets that contain categorical attributes. However, datasets with categorical types of attributes are common in real life data mining problem. For each pair of documents, a comparison vector is constructed that contains binary features that measure the overlap for highly informative but sparse features between the two documents and numeric features.

The aggregating the comparison vector into one value that belongs to interval. The aggregation step is performed by taking a weighted average the information gain has a tendency to favor features with many possible values over feature with fewer possible values ,we used a normalized version of information gain, called gain ration as weighting metric.

2. Related work

Clustering is of prime importance in data analysis, machine learning and statistics. It is defines as the process of grouping N item sets into distinct clusters based on similarity or distance function A good clustering technique may yield clusters thus have high inter cluster and low intra cluster distance[7].The objective of clustering is to maximize the similarity of the data points within each cluster and maximize dissimilarity across clusters.

Broadly speaking clustering algorithms can be divided into two types partitioned and hierarchical. Partitioning algorithms construct a partition of a database D of n objects into a set of clusters where k is a input parameter.

Hierarchical algorithms create decomposition of the database D. they are a Agglomerative and divisive. Hierarchical clustering builds a tree of clusters, also known as a dendrogram. Every cluster node contains child cluster. An agglomerative clustering starts with one-point (singleton) Clusters and recursively merges two or more most appropriate clusters. A divisive clustering starts with one cluster of all data points and recursively splits into the most appropriate clusters. The process continues until a stopping criterion is achieved. There are two main issues in clustering techniques. Firstly, finding the optimal number of clusters in a given dataset and secondly, given two sets of clusters, computing relative measure of goodness between them.

For both these purposes, a criterion function or a validation function is usually applied. The simplest and most widely used cluster optimization function is the sum of squared error [4]. Studies on the sum of squared error clustering were focused on the well-known k-Means algorithm[5] and its variants.

In conventional clustering objects that are similar are allocated to the same cluster while objects that differ are put in different clusters. These clusters are hard clusters. In soft clustering an object may be in more than two or more clusters.

Based on the simple clustering concept as described as, Suppose there are N subjects to be clustered. For each subject, there are P observations (random variables) representing the subjects' features. We can view each subject as a data point in a P-dimensional space. Imitating the aforementioned example, we can construct the following mechanism to move the data points (subjects). The movement of each subject is determined by the between subject proximity, which can be any measure such as the Euclidean distance or correlations.

The algorithm can be written as follows.

1. $X(0)^1$,$X(0)^2$, ··· ,$X(0)^N \in R^p$ to be clustered.

2. At time t + 1, every point is updated according to

$$X_i^{(t+1)}= \begin{matrix} \exp[-\frac{d}{\lambda}]....d \le r \\ 0............d > r \end{matrix}$$

3. Repeat (2) until every point converges.
f is a statistic that measures the between-subject proximity.
we propose to use

$$f(u,v)= \begin{matrix} \exp[-\frac{d}{\lambda}]....d \le r \\ 0............d > r \end{matrix}$$

where r and λ are fixed constants, and d is the Euclidean distance from u to v.

3. Improved FCM clustering algorithm

However, traditional FCM did not consider the spatial information therefore, it is very sensitive to noise and the results will be affected when different vectors contribute not the same to cluster. Considering problems above, some researchers proposed Improved (IFCM) which based on cluster information of data groups . In this paper, we improve IFCM by changing its way to select initial cluster centres according to image distribution. the cluster sample will now FCM algorithm[6] is very sensitive to the number of cluster centers, cluster centers initialization often artificially get significant errors, and even get the actual opposite results .FCM algorithm is hard on data sets too ,so the data sets must be quite regular, in order to solve problems, first of all we use information entropy to initialize the cluster centers to determine the number of cluster centers. it can be reduce some errors, and also can improve the algorithm introductions an weighting parameters .after that, combine with the merger of ideas, and divide the large chumps into small clusters. Then merge various small clusters according to the merger of the conditions, so that you can solve the irregular datasets clustering. document similarity measures .

```
The algorithm as follows
 Get the class prior prababities $\{P_r\}_{c=1}^{C}$
Set the class growth rate $n_c$=n X $Pr_c$
Where c=1,…..C
If $H^{(0)}$ I not given then
Construct an initial clusters of N
For t initialize $C_j$ (cluster centers)
Initialize $\alpha$ (threshold value)
Repeat
For i=1 to n :update $\mu_j(X_i)$
        For k=1 to p ;
          Sum=0
          Count=0
        For i=1 to n:
        If μ ($X_i$) is maximum in $C_k$   then
         If μ ($X_i$)>= $\alpha$
         Sum=sum+$X_i$
         Count= count+1
    $C_k$=sum/count
Until $C_j$ estimate stabilize.
```

The clustering framing as follows
Set value for cluster numbers
algorithm stop threshold $\varepsilon \geq 0$,

A set clusters $C=\{C_1, C_2, C_3,.......C_k\}$

The experiments use classical IRIS as the test data set to verify the validaty o the algorithm .Iris data set contains infroamtion on 150 species of iris ,and it is divided into three categories ,anemly,Iris-sectosa,Iris-versicolor and Iris-virginica.each type contains 50 kinds of data,and each data contains five kinds attributes.They are sepal length,sepal widht ,petal length,petal widht,classes.

Table 1 the classification of new algorithm and FCM algorithm for IRIS data set.

	Clustering centers	Wrong number of points	Error fraction

FCM	C1=(5.0039,3.4137,0.2537)
	C2=(5.8974,2.7656,4.3985,1.4041)
	C3=(6.7813,3.0543,2.0567)
		16	10.7%

Entropy weighting FCM	C1=(4.9979,3.3849,1.4623,0.2458)
	C2=(5.7839,2.7856,4.1786,1.2783)
	C3=(6.5129,2.9253,5.4226,2.0028)
		7	4.7%

The orginal location of IRIS data set's centers are:
C1=(5.00,3.42,1.46,0.24),C2=(5.93,2.77,4.26,1.32),C3=(6.58,2.97,5.55,2.02) table1 lists the clusters of two algorithms,and we can find the clustering centers of imporved algorithm are more reasonalbe than one of FCM algorithm.we also learn that error fraction of the entropy weighting FCM is lower than the one of orginal alogrithm.

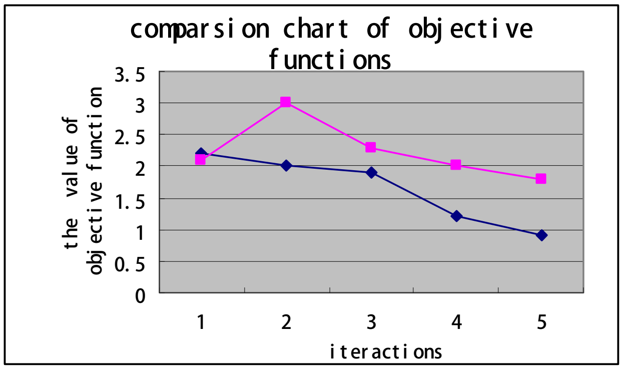

Fig 1 .The Comparison chart of objective functions

The objective function comparsion chart of the two algorithms,here before represents the objective function of FCM clustering map,and after represents the algorithm's objective function on behalf of distribution.Abscisssa presents the number of iterations,and vertical presents the distrubution function values .fig 1 show that the iteraction number of the improved algorithm is less than the one of the original algorithm.it means that the algorithm's efficiency has been greatly improved.

4. Conclusions

The suggested approach was used for efficacy contained a hard clustering of the iris data set and as the analysis indicated each of the clusters seems to contain observations with specific common charters tics and improve the algorithm efficiency ,ehich is able to identify clsuters of arbitary shape and handle noisy data to some extent.Experiments prove the improved algorithm has able to identify the initial cluster centers with help of entropy and to adopted information entropy to determine the number of redudant center,still there are some defects.to seek a better method is the focus and direction for the further work..

References

[1] J. Srivastava, R. Cooley, M. Deshpande, PN. Tan, Web usage mining: discovery and applications of usage patterns from web data. SIGKDD Explorations, Vol. 1, No. 2, 2000, pp.12–23.

[2] M. N. Garofalakis, R. Rastogi, S. Seshadri, K. Shim Data minino and the web: past, present and future, In Proc. of the second international workshop on webinformation and data management, ACM, 1999.

[3] A. K. Jain and R. C. Dubes, “Data clustering: A review.,” ACM Computing Surveys, vol. 31, 1999.

[4] U. Maulik and S. Bandyopadhyay, “Genetic algorithm based clustering technique,” Pattern Recognition, vol. 33, pp. 1455–1465, 2000.

[5] P. Zhang, X. Wang, and P. X. Song, “Clustering categorical data based on distance vectors,” The Journal of the American Statistical Association, vol. 101,no. 473, pp. 355–367, 2006.

[6] A. Vakali, J. Pokorný and T. Dalamagas, An Overview of Web Data Clustering Practices, EDBT Workshops, 2004, pp. 597-606.

[7] Lin Zhu, Fu-Lai Chung, Shitong Wang.Generalized Fuzzy C-Means Clustering Algorithm With Improved Fuzzy Partitions[J].IEEE Transactions on Systerms,2009:39-3.

[8] Cheul Hwang,Frank Chung-Hoon Rhee.Uncertain Fuzzy Clustering:Interval Type-2 Fuzzy Approach to C-Means[J]. IEEE Transcations on Fuzzy Systerms,2007.

K.Suresh received his Bachelor and Master degree from Jawaharlal Nehru Technological University, Hyderabad. He is currently working as Foreign Faculty in East China University of Technology, Nanchang Campus,Jiangxi-330013,P.R.China and Visiting Faculty for Jiangxi Normal University, P.R.China. Previously he is worked as Employee of AITS,Rajampet,A.P,India,he has published more than 20 national and International conference papers and journals. he received best paper award in 2008 in national wide paper presentation. his field of interest is data mining, database technologies, information retrieval system.

R. Madana Mohana received his B. Tech in Computer Science and Information Technology from Jawaharlal Nehru Technological University, Hyderabad in 2003 and M. E in Computer Science and Engineering from Sathyabama University ,Chennai in 2006. He is doing Ph. D in Computer Science and Engineering at Sri Venkateswara University. he is Associate Professor in the Information Technology department at Vardhaman College of Engineering, Hyderabad, Andhra Pradesh, India since 2007. He has 8 years of teaching experience at both UG and PG levels. He is a life member of ISTE Technical Association. His areas of interest include Data Mining, Automata Theory, Compiler Design and Database Systems.

Dr. A. Rama Mohan Reddy received the B. Tech. from JNT University, Hyderabad in 1986, M. Tech degree in Computer Science from National Institute of Technology in 2000 Warangal and Ph. D in Computer Science and Engineering in 2008 from Sri Venkateswara University, Tirupathi, Andhra Pradesh, India.He worked as Assistant Professor, Associate Professor and Presently working as Professor of Computer Science and Engineering, Sri Venkateswara University College of Engineering. He has 28 years of Industry and Teaching experience. Currently guiding twelve Ph. D scholars. He is life member of ISTE and IE. His research interests include Software Architecture, Software Engineering and Data Mining. He has 10 international publications and 14 international conference Publications at International and National level.