Enhancement of Fuzzy C-Means Clustering using EM Algorithm

Esh Narayan
Lovely Professional University
Jalandhar (Punjab)

Yogesh Birla
Lovely Professional University
Jalandhar (Punjab)

Gaurav Kumar Tak
Lovely Professional University
Jalandhar (Punjab)

ABSTRACT

In present time many clustering techniques are use the data mining. The clustering gives the best performance in all research proposals. We know that k_ means clustering is the best clustering techniques in the data mining, but some disadvantage of k_ means clustering. Although clustering algorithms can be classified into two categories: hard clustering algorithms and fuzzy clustering algorithms and The Fuzzy-C-Means (FCM) clustering algorithm is also known as Soft K-Means so we have proposed new enhancement of C-means clustering. Proposed algorithms will be gave the best result in fuzzy C-means clustering algorithms. We use the mouse data set to test these performances in Matlab tools.

Keywords

Feature selection, C-mean clustering, EM algorithms, mouse dataset, Matlab, unsupervised learning

1. INTRODUCTION

The method to arrange the set of objects into classes of similar (which are having same behavior) objects, is defined as clustering. We can categorize the objects into 2 categories, which is provided similar or dissimilar cluster from the documents.

- Documents within a cluster should be similar.
- Documents from different clusters should be dissimilar.

Cluster analysis is identifying such grouping (or clusters) in an unsupervised manners, in unsupervised approach are divides a set of objects into homogeneous groups. There have been many clustering algorithms scattered in publications in very diversified areas such as pattern recognition, artificial intelligence, information technology, image processing, biology, psychology, and marketing [2]. Clustering algorithms can be classified into main two categories: hard clustering algorithms and fuzzy clustering algorithms. Unlike hard clustering algorithms, which require that each data point of the data set belong to one and only one cluster, fuzzy clustering algorithms allow a data point to belong to two or more clusters with different probabilities. There is also a huge number of published works related to fuzzy clustering [9].The hard clustering is sometimes called non-fuzzy clustering where as the soft clustering is referred to fuzzy clustering. From a general perspective, the hard and soft clustering can be seen as differing in the assigning of values for a partition matrix.But clustering is useful in a visualization of data and hypothesis generation. In fig. 1 are to shown the data patterns can change into the clusters.

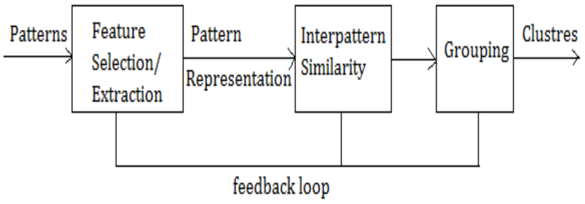

Fig.1: To change the patterns into the clusters

Basically clustering has many types but clustering algorithms can be classified into main two categories: hard clustering algorithms and fuzzy clustering algorithms.

1.1 Hard clustering

Basically hard clustering has each document belongs to exactly one cluster. In hard clustering we make a hard partition of the data set Z. The hard clustering approach only assigns a value of 1 and 0.

$$\cup_{i=1}^{e} Ai = ZandAi \cap Aj = \emptyset foralli \neq j \quad \ldots\ldots\ldots\ldots\ldots\ (1.1)$$

Also none of the set Ai may be empty and $i \neq j$ for all value of i=1 to e.

1.2 Fuzzy clustering

Fuzzy clustering also called soft clustering. In fuzzy clustering we make a fuzzy partition of the data set.Fuzzy clustering, each point has a degree of belonging to clusters, as in fuzzy logic, rather than belonging completely too just one cluster. Fuzzy clustering uses membership function in partition data set.

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1, \ \ldots..,n \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ (1.2)$$

This function is called membership function and its value between 0 and 1.

2. FUZZY C-MEANS CLUSTERING

Fuzzy C-means Clustering (FCM) is also known as Fuzzy ISODATA, is a clustering technique which is separated from hard k-means that employs hard partitioning. The FCM employs fuzzy partitioning such that a data point can belong to all groups with different membership grades between 0 and 1.

FCM is an iterative algorithm. The aim of FCM is to find cluster centers (centroid) that minimize a dissimilarity function. This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point [3][5].

To accommodate the introduction of fuzzy partitioning, the membership matrix (U) is randomly initialized according to Equation 2.1

$\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1, \ldots, n$…………………(2.1)

To find the value of centroid (c_i) to help of membership matrix (u_{ij})

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}{}^{m} x_j}{\sum_{j=1}^{n} u_{ij}{}^{m}}$$ ………..…..…. (2.2)

The dissimilarity function which is used in FCM is given Equation

$$J(U, c_1, c_2, \ldots, c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}{}^{m} d_{ij}{}^{2}$$ ……………..… (2.3)

u_{ij} is between 0 and 1

d_{ij} is the Euclidian distance between i_{th} centroid (c_i) and j_{th} data point;

Where dij = $\sqrt{\sum_{i=1}^{n}(xi - ci)^2}$……….......… (2.3.1)

$$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{d_{ij}}{d_{kj}}\right)^{2/(m-1)}}$$ …………..…. (2.3.2)

The equation (2.3.1), we have fined the Euclidian distance and the equation (2.3.2) have fined the member ship value, which value is important for the further steps.

If|| U (k+1) - U (k) ||<∈ …………..……………. (2.4)

In this equation we have to check the value of present classes and the next classes of the membership function. If values are satisfied, Then STOP; otherwise return to step 2.

2.1 Fuzzy c-means algorithms

Basically fuzzy c-means clustering depends on equations (2.1) to (2.4). Thesealgorithm works by assigningmembership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point, this algorithms works on fallowing these four steps.

Step1.In these step we find the membership matrix (U) initialise randomly in equation2.1

Step2. Calculate the centroids (ci) in equation 2.2

Steps3. Using dissimilarity function to calculate the dissimilarities between centroid and data points in equation2.3

Step4. Check threshold value in equation2.4 then stop if we find the correct threshold value.

If|| U (k+1) - U (k) ||<∈ …..……Then STOP;otherwise back in step2 and also find new (u_{ij}).

3. EM ALGORITHMS

An EM (Expectation maximization) algorithm is very use full in statically model. The most common algorithm uses an iterative refinement technique. These algorithms are giving the best result in clustering method; it is also referred to as LoyardAlgo particularly in the computer science community. EM algorithms given an initial set of c-means $m_1^{(1)}\ldots m_k^{(1)}$, the algorithm proceeds by alternating between two steps [9].

3.1 Assignment step

In these steps Assign each observation to the cluster with the closest means, it is according to Voronoi diagrams for finding new mean [8].

$S_i^{(l)} = \{X_p : \| X_p - m_i^{(l)} \| \leq \| X_p - m_j^{(l)} \| \ \forall\ 1 \leq j \leq k$ ….....……………..………………….…………………… (3.1)

Where each $X_p$goes into exactly oneS_i^l, even if it could go in two of them

3.2 Update step by

Calculate the new means to be the centroid of the observations in the cluster.

$$m_i^{(l+1)} = \frac{1}{\left|S_i^{(l)}\right|}\sum_{X_j \in S_i^{(l)}} X_j$$ ……..........……… (3.2)

In this equation (3.2) we find new mean for new cluster. So we can say these algorithms is deemed to have converged when the assignment no longer change. And it gave the best performance of initial means. EM algorithms commonly used initialization methods are random partition.

The "assignment" step (3.1) is also referred to as expectation step, the "update step" (3.2) as maximization step, making this algorithm a variant of the generalized. The Forgy method randomly chooses k observations from the data set and uses these as the initial means. The Random Partition method first randomly assigns a cluster to each observation and then proceeds to the Update step, thus computing the initial means to be the centroid of the cluster's randomly assigned points. These are basic points of the Forgy methods

- Demonstration of the standard algorithm
- tends to spread the initial means out, while Random Partition places all of them
- Close to the center of the data set.
- According to Random Partition method is generally preferable.

4. ARTIFICIAL INTELLIGENCE

It is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence. AI is the intelligence to machine learning and that aims to make to the capable of machine learning. Mainly use of the artificial intelligence in the machine learning, which is provide to control and gave the good performance, but AI does not have to confine itself to methods that are biologically observable [3].

4.1 Supervised

A machine learning technique for deducing a function from training data, also we can say learning with teacher. In supervised learning, the model defines the effect one set of observations, called inputs, has on another set of observations, called outputs. In other words, the inputs are assumed to be at the beginning and outputs at the end of the causal chain. The models can include mediating variables between the inputs and outputs. Supervised learning includes the classification and regression; these techniques are very important in machine learning, image processing, data mining etc.

4.2 Unsupervised

A class of problems in which one seeks to determine how the data are organized and also we can learn without teacher unsupervised learning is the best technique in the machine learning approach. We can say unsupervised learning gives the best performance in artificial intelligence [2].

Unsupervised learning is likely to be much more common in the brain than supervised learning. Then we can say no information is provided to the algorithm on which data points belong to which clusters.

Main approaches of unsupervised learning.

- Clustering (k_ means, c-means, hierarchical clustering)
- Feather extraction (PCA, SOM)

PCA (Principle component analysis) is basically use for the dimensionally reduction.

5. PROPOSED ALGORITHMS

My proposal to enhancing the performance of c-means clustering using EM algorithms. My proposed algorithm is called Expectation maximization fuzzy c-means clustering (EMFCM). EMFCM algorithms is very help full to increase the performance of machine learning, all data mining approaches, image processing, network security etc.

Proposed algorithms used the clustering techniques and EM algorithms, which provided the sufficient result for the cluster analysis in maximum mean to calculate the fixed centroid and correct threshold value.

Proposed algorithms have these steps.

Step1.In these step we find the membership matrix (U) initialise randomly in equation5.1.

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1, \ldots, n \qquad \ldots\ldots\ldots\ldots\ldots\ldots (5.1)$$

This equation represent the membership matrix, it has taken the value equal to 1.

Step2. Calculate the centroids (ci) in equation 5.2

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}{}^{m} x_j}{\sum_{j=1}^{n} u_{ij}{}^{m}} \qquad \ldots\ldots\ldots\ldots\ldots\ldots (5.2)$$

Centroid is main point of the cluster analysis system, in clustering this value of c_i is depends on the member matrix function and related parameter ofx_i .

Steps3. Using dissimilarity function to calculate the dissimilarities between centroid and data points in equation5.3 and check threshold value in equation5.4 then stop if we find the correct threshold value.

$$J(U, c_1, c_2, \ldots, c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}{}^{m} d_{ij}{}^{2} \qquad \ldots\ldots\ldots\ldots\ldots\ldots (5.3)$$

In this steps check the threshold value using membership matrix and Euclidian distance between i_{th} centroid (c_i) and j_{th} data point.

If

$$\| U(k+1) - U(k) \| < \in \ldots\ldots\ldots\ldots\ldots\ldots (5.4)$$

In this equation we have to check the value of present classes and the next classes of the membership function. Check the value of membership matrix in next membership matrix in correct threshold value, which have been done by dissimilarities between centroid and data points. If values are satisfied, then we forwarded the next steps.

Step4. If threshold value is not correct we find the new mean (m_i) using EM algorithm that has constraints in equation (5.5) using this equation we can find new mean, which is provided the correct threshold value for the dissimilarity function, So we can say these algorithms is deemed to have converged when the assignment no longer change. And it gave the best performance of initial means. EM algorithms commonly used initialization methods are random partition.

$$m_i^{(l+1)} = \frac{1}{\left|S_i^{(l)}\right|} \sum_{X_j \in S_i^{(l)}} X_j \ldots\ldots\ldots\ldots\ldots\ldots (5.5)$$

Step5. In these steps assign each observation to the cluster with the closest means in equation (5.6)

$$S_i^{(l)} = \{X_p : \| X_p - m_i^{(l)} \| \leq \| X_p - m_j^{(l)} \| \; \forall \; 1 \leq j \leq k .. (5.6)$$

In this equation (5.5) we find new mean for new cluster. So we can say these algorithms is deemed to have converged when the assignment no longer change.

6. CONCLUSION AND FURTHER WORK

We know that clustering gives the best performance in all research proposals in present time but some casesclustering are not give the better result so we enhance the clustering techniques, which algorithms gives the better performance. Clustering gives the best performance in unsupervised learning. In proposed are algorithms expectation maximization fuzzy c-means clustering (EMFCM), proposed algorithms will provide the better result to fuzzy c-means clustering by avoiding the looping problems and save the time. This algorithm will be helping of to increase the performance of intrusion detection system. And also EMFCM algorithms is very help full to increase the performance of machine learning, all data mining approaches, image processing, network security etc.

In future work implements the proposed algorithms and check the performance also comparing the fuzzy c-means clustering. For implement we can use the mouse data set in MATLAB tool and check the performance of EMFCM algorithms.

7. ACKNOWLEDGEMENTS

I wish to express my science profound gratitude to Mr. Gaurav Kumar Tak Asst. Professor, whose supervision & guidance in this investigation has been carried out, without whose guidance and constant supervision. It is not possible for me to complete this research paper successfully.

8. REFERENCES

[1]. Ozyilmaz, L.; Yildirim, T.; , "Diagnosis of thyroid disease using artificial neural network methods," Neural Information Processing, 2002. ICONIP '02. Proceedings of the 9th International Conference on , vol.4, no., pp. 2033- 2036 vol.4, 18-22 Nov. 2002 doi: 10.1109/ICONIP.2002.1199031

[2]. Abdel-Azim, M.; Abdel-Fatah, A.I.; Awad, M.; "Performance analysis of artificial neural network intrusion detection systems," Electrical and Electronics Engineering, 2009. ELECO 2009. International Conference on , vol., no., pp.II-385-II-389, 5-8 Nov. 2009

[3]. PengShanguo; Wang Xiwu; ZhongQigen; , "The study of EM algorithm based on forward sampling," Electronics, Communications and Control (ICECC), 2011 International Conference on , vol., no., pp.4597-4600, 9-11 Sept. 2011 doi: 10.1109/ICECC.2011.6067693

[4]. Fisher, D.; Ling Xu; Carnes, J.R.; Reich, Y.; Fenves, J.; Chen, J.; Shiavi, R.; Biswas, G.; Weinberg, J.; , "Applying AI clustering to engineering tasks," IEEE Expert , vol.8, no.6, pp.51-60, Dec. 1993 doi: 10.1109/64.248353

[5]. J.-S. R. Jang, C.-T. Sun, E.Mizutani, Neuro-Fuzzy and Soft Computing, p (426-427)Prentice Hall, 1997

[6]. Maria Colmenares& Olaf WolkenHauer, “An Introduction into Fuzzy Clustering”, http://www.csc.umist.ac.uk/computing/clustering.htm, July 1998, last update 03 July,2000

[7]. http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html

[8]. www.ics.uci.edu/pub/ml-repos/machine-learning-database/, 2001

[9]. Neal, Radford; Hinton, Geoffrey (1999). Michael I. Jordan. ed."A view of the EM algorithm that justifies incremental, sparse, and other variants". Learning in Graphical Models (Cambridge, MA: MIT Press): 355–368.

[10]. Jain, A. K. And Dubs, R.C. Algorithms for clustering data (N. J. & Cliffs; Prentice Hall) in 1999.