Comparative Research of Clustering Algorithms for Prediction of Academic Performance of Students

Ankita R. Chordiya,
PG Student E &TC Department,
Lt. G. N. Sapkal College of Engineering
Savitribai Phule Pune University

Prof. S. B. Bagal
Associate Professor & Head, E &TC Department,
Lt. G. N. Sapkal College of Engineering,
Savitribai Phule Pune University

***Abstract*—- Clustering is a task of assigning a set of objects into groups called clusters. In general the clustering algorithms can be classified into two categories. One is hard clustering; another one is soft (fuzzy) clustering. Hard clustering, the data's are divided into distinct clusters, where each data element belongs to exactly one cluster. In soft clustering, data elements belong to more than one cluster, and associated with each element is a set of membership levels In order to monitor the progress of students efficiently, different clustering algorithms are applied to the academic results of students so as to categorize in appropriate class as per their performance. We proposed the use of FCM and KFCM clustering algorithms for prediction of students' academic performance. Euclidean distance as a measure of similarity measurement is taken into consideration. These algorithms are applied and performance is evaluated on the basis of clustering output. FCM allows data points to belong to more than one cluster where each data point has a degree of membership of belonging to each cluster. The KFCM whereas uses a mapping function and gives better performance as compared to FCM. The summarized result shows that KFCM gives better performance than FCM**

Keywords—FCM, KFCM, clustering, academic performance, membership function

I. INTRODUCTION

Pattern recognition is a branch of machine learning that focuses on the recognition of patterns and regularities in data clustering and classification are the major subdivisions of pattern recognition technique although clustering and classification are often used for purposes of segmenting data records, they have different objectives and achieve their segmentations through different ways. Classification is supervised learning technique used to assign predefined. tag to instance on the basis of features. So classification algorithm requires training data. Classification model is created from training data, then classification model is used to classify new instances. Clustering is unsupervised technique used to group similar instances on the basis of features. Clustering does not require training data. Clustering does not assign per-defined label to each and every group.

Fast and robust clustering algorithms play an important role in extracting useful information in large databases. The aim of cluster analysis is to partition a set of N object into C clusters such that objects within cluster should be similar to each other and objects in different clusters are should be dissimilar with each other. Clustering can be used to quantize the available data, to extract a set of cluster prototypes for the compact representation of the dataset, into homogeneous subsets.

Clustering is a mathematical tool that attempts to discover structures or certain patterns in a dataset, where the objects inside each cluster show a certain degree of similarity. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Cluster analysis is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization. It will often necessary to modify preprocessing and parameter until the result achieves the desired properties. In Clustering, one of the most widely used algorithms is fuzzy clustering algorithms. Fuzzy set theory was first proposed by Zadeh in 1965 [1]& it gave an idea of uncertainty of belonging which was described by a membership function. The use of fuzzy set provides imprecise class membership function. Applications of fuzzy set theory in cluster analysis were early proposed in the work of Bellman, Zadeh, and Ruspini [8]. This paper opens door step of fuzzy clustering [2]. Integration of fuzzy logic with data mining techniques has become one of the key constituents of soft computing in handling challenges posed by massive collections of natural data. The central idea in fuzzy clustering is the non-unique partitioning of the data into a collection of clusters. The data points are assigned membership values for each of the clusters and fuzzy clustering algorithm allow the clusters to grow into their natural shapes [3].

The fuzzy clustering algorithms can be divided into two types . The FCM is the soft extension of the traditional hard c-means clustering[1]. Each cluster was considered as fuzzy set and the membership function measures the possibility that each training vector belongs to a cluster. so the vectors may be assigned to multiple clusters. Thus, it overcomes some drawbacks of hard clustering but it is effective only when the data is non-overlapping. By the contrast to the crisp c-partitions, in a fuzzy case elements can belong to several clusters and to different degrees [16]. 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. More the data is near to the cluster center more is its membership towards the particular cluster center.

In KFCM, both the data and the cluster centers are mapped from the original space to a new space by ϕ [15] . An important fact about kernel function is that it can be directly constructed in the original input space without knowing the form of ϕ.That is, a kernel function implicitly defines a nonlinear mapping function [13].

The remainder of this paper is organized as follows: Section II provides the basic algorithm of FCM. Section III provides kernel based FCM. Section IV proposes the application of FCM and KFCM in prediction of students academic performance and the results are compared for each of the algorithm in terms of cluster efficiency. Section V presents the conclusions drawn for the comparison of the FCM and KFCM on the basis of cluster efficiency.

II. Fuzzy C-Means (FCM)

Fuzzy c-means (FCM) is a method of clustering which allows one piece of data to belong to two or more clusters. This method (developed by Dunn in 1973 and improved by Bezdek in 1981) is frequently used in pattern recognition. Straightly speaking, this algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster and the data point. More the data is near to the cluster center more is its membership towards the particular cluster center. Clearly, summation of membership of each data point should be equal to one [8]. The algorithm is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^m \left| x_i - c_j \right|^2 \qquad (1)$$

where m (the Fuzziness Exponent) is any real number greater than 1, N is the number of data, C is the number of clusters, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the ith of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership u_{ij} and the cluster centers c_j by:

$$u_{ij} = \frac{1}{\sum_{i=1}^{c}\left[\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right]^{2/m-1}} \qquad (2)$$

where $\|x_i - c_j\|$ is the Distance from point i to current cluster centre j, $\|x_i - c_k\|$ is the Distance from point i to other cluster centers k.

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m . x_i}{\sum_{i=1}^{N} u_{ij}^m .} \qquad (3)$$

When $\max_{ij}\left\{u_{ij}^{k+1} - u_{ij}^{k}\right\} < \varepsilon$, where ε is a termination criterion between 0 and 1, whereas k are the iteration steps, the iteration stops. This procedure converges to a local minimum or a saddle point J_m.

The algorithm is composed of the following steps:

1. Randomly select cluster centre
2. Initialize $U=[u_{ij}]$ matrix, $U^{(0)}$
   Calculate the the u_{ij} using:

$$u_{ij} = \frac{1}{\sum_{i=1}^{c}\left[\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right]^{2/m-1}}$$

3. At k-step: calculate the centres vectors $C^{(k)}=[c_j]$ with $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m . x_i}{\sum_{i=1}^{N} u_{ij}^m .}$$

4. Update $U^{(k)}$, $U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{i=1}^{c}\left[\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right]^{2/m-1}}$$

5. If $\| U^{(k+1)} - U^{(k)}\| < \varepsilon$ or the minimum J is achieved, then STOP; otherwise return to step 2.

III. Kernel fuzzy C-means clustering

The KFCM algorithm adds kernel information to the traditional fuzzy c-means algorithm and it overcomes the disadvantage that FCM algorithm can't handle the small differences between clusters. The main idea of fuzzy kernel c-means algorithm (KFCM) is described as follows. The kernel method maps nonlinearly the input data space into a high dimensional feature space.

Given a dataset, $X= \{x ,..., xn\} \subset R^p$, the original FCM algorithm partitions X into c fuzzy subsets by minimizing the following objective function

$$J_m(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^m \left\| x_k - v_i \right\|^2 \qquad (4)$$

where *c* is the number of clusters and selected as a specified value in this paper, *n* the number of data points, u_{ik} the membership of x_k in class *i*, satisfying $\sum_{i=1}^{c} u_{ik} = 1$, *m* the quantity controlling clustering fuzziness, and *V* the set of cluster centers or prototypes ($v_i \in R^p$). The function *J m* is minimized by a famous alternate iterative algorithm. Now consider the proposed kernel fuzzy c-means (KFCM) algorithm. Define a nonlinear map as $\Phi : x \rightarrow \Phi(x) \in F$, where $x \in X$. *X* denotes the data space, and *F* the transformed feature space with higher even infinite dimension. KFCM minimizes the following objective function :-

$$J_m(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} \left\|\phi(x_i) - \phi(v_i)\right\|^2 \quad (5)$$

where $\left\|\phi(x_i) - \phi(v_i)\right\|^2$ = K(ξκ, ξκ) + K(ϖι ,ϖι) □ 2K(ξκ , ϖι)
(6)

Where $K(x, y) = \Phi(x)^T\Phi(y)$ is an inner product kernel function. If we adopt the Gaussian function as a kernel function, i.e.

K(ξ,ψ)= εξπ ($-\left\|x - y\right\|^2 / 2\sigma^2$), τηεν $K(\xi, \xi) = 1$,

according to Eq. (5), Eq. (6) can be rewritten as,

$$J_m(U,V) = 2\sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m}\left(1 - K(x_k, v_i)\right) \quad (7)$$

Minimizing Eqs. (4) under the constraint of u_{ik} , we have

$$\upsilon_{ik} = \frac{\left(1/\left(1 - K(x_k, v_i)\right)\right)^{1/m-1}}{\sum_{j=1}^{c}\left(1/\left(1 - K(x_k, v_i)\right)\right)^{1/m-1}} \quad (8)$$

$$v_i = \frac{\sum_{k=1}^{n} u_{ik} K(x_k, v_i) x_k}{\sum_{k=1}^{n} u_{ik}^{m} K(x_k, v_i)} \quad (9)$$

Here we just use the Gaussian kernel function for simplicity. If we use other kernel functions, there will be corresponding modifications in Eq. (5) and (6). In fact, Eq.(3) can be viewed as kernel-induced new metric in the data space, which is defined as the following :

$$\left\|\phi(x) - \phi(y)\right\| = \sqrt{2\left(1 - K(x, y)\right)}$$

And it can be proven that *d*(*x*, *y*) defined in Eq. (7) is a metric in the original space in case that *K(x,y)* takes as the Gaussian kernel function. According to Eq. (6), the data point x_k is endowed with an additional weight $K(x_k, v_i)$, which measures the similarity between x_k and v_i , and when x_k is an outlier, i.e., x_k is far from the other data points, then $K(x_k, v_i)$ will be very small, so the weighted sum of data points shall be more robust.

KFCM Algorithm

Step 1: Fix *c* ,t_{max}, *m* >1 and ε >0 for some positive constant;
Step 2: Initialize the memberships u_{ik}^{0}
Step 3: For *t* =1,2,…,t_{max}, do:
(a) Update all prototypes v_i^t with Eqs. (9);
(b) Update all memberships u_{ik}^{t} with Eqs. (8);
(c) Compute $E^t = \max_{i,k}\left|u_{ik}^{t} - u_{ik}^{t-1}\right|$,if $E^t \leq \varepsilon$,stop; else t=t+1

IV. RESULTS

We applied the model on the data set (academic result of one semester) of a University of Pune. Table I shows the dimension of the data set (Student's scores)in the form N by M matrices, where N is the rows (# of students) and M is the column (# of courses) offered by each student. In table II, Performance index is specified as per the average score of every individual so as to categorize in different class

The result generated is shown in tables III and IV. The corresponding algorithm is applied individually to the dataset and the respective count i.e the number of samples in each cluster are evaluated. The count values obtained in each of the clustering algorithm are thus compared with the actual values and the overall performance of each cluster is calculated. The summarized results shows that KFCM has better performance as compared to FCM.

TABLE I. STATISTICS OF DATA USED

Student's score	***Number of students***	***Dimension(Total number of subjects)***
Data	78	5

TABLE II. PERFORMANCE INDEX

70 & above	**Excellent**
60-69	Very Good
50-59	Good
45-49	Very Fair
40-44	Fair

TABLE III. INDIVIUAL CLUSTER EFFICIENCY FOR FCM

Sr. No.	***Cluster***	***Actual***	***Count***	***Cluster Efficiency (%)***
1	1	14	18	94
2	2	9	25	84
3	3	18	5	87
4	4	25	9	84
5	5	12	43	69

TABLE IV. INDIVIDUAL CLUSTER EFFICIENCY FOR KFCM

Sr. No.	***Cluster***	***Actual***	***Count***	***Cluster Efficiency (%)***
1	1	14	15	99
2	2	9	13	96
3	3	18	14	96
4	4	25	20	95
5	5	12	38	74

V. CONCLUSIONS

Thus we studied and applied different clustering algorithms for the purpose of result analysis of students academic performance. The results of the paper in terms of cluster accuracy confirmed that KFCM has a better performance than FCM when applied to evaluate the academics result of students. Also in future,numerical interpretation of the results based on the clustering algorithms will be shown which will be helpful in making an effective decision by the academic Planners.

REFERENCES

[1] James C. Bezdek, Robert Ehrlich and William Full, " FCM: The Fuzzy c-means Clustering Algorithm" , Computer and Geosciences Vol. 10 , No. 2-3, pp. 191-203, 1984.

[2] Oyelade, Oladipupo and Obagbuwa, "Application of k-means Clustering Algorithm for Prediction of Students' Academic Performance", Internation Journal of Computer science and Information Security, Vol. 7, No. 1, 2010.

[3] SoumiGhosh and Sanjay kumarDubey, "Comparative Analysis of k-means and Fuzzy c-means Algorithms", International Journal of Advanced Computer Scienceand Applications", Vol. 4, No. 4, 2013.

[4] MamtaBharadwaj, AnkitaWalia and HemantTulsani, " Comparative Research on Fuzzy c-means and k-means Clustering Segmentation", International Journalof Computer Application and Information Technology, Vol. 3, Issue II, 2013, ISSN : 2278-7720.

[5] G. Raghotham Reddy, K.Ramudu, Syed zaheeruddin and Dr. R.RameshwarRao, "Image Segmentation Using Kernel Fuzzy c-means Clustering on Level Set Method on Noisy Images.

[6] Dao-Qiang Zhang and Song-Can Chen, "Kernel-based Fuzzy and Possibilities c-means Clustering".

[7] Tara.Saikumar, B.K.Anoop and P.S.Murthy, "Robust Adaptive Thresold Algorithm based on Kernel Fuzzy Clustering on Image Segmentation".

[8] R.Suganya, R.Shanthi, "Fuzzy C- Means Algorithm-AReview"International Journal of Scientific and ResearchPublications, Volume 2, Issue 11, November 2012 1 ISSN2250-3153

[9] Bezdek JC, "Cluster validity with fuzzy sets," Journal ofCybernetics, Vol.3, 1974, pp. 58-73

[10] D.Q. Zhang, S.C. Chen, "A novel kernelized fuzzy c-meansalgorithm with application in medical imagesegmentation,"Art. Intell. Med, vol. 32, 2004, pp. 37-50

[11] Bezdek JC, "Pattern recognition with fuzzy objective functionalgorithm," New York: Plenum Press, 1981

[12] D.L. Pham, J.L. Prince, "An adaptive fuzzy c-meansalgorithm for Imagesegmentation in the presence of intensityinhomogeneities," PatternRecogn. Lett. vol. 20, 1999, pp.57-68.

[13] Z. Wang, S. e. Chen, and T.K.Sun," Multiple kernel learning algorithm,"IEEE Transactions on patterns analysis and Machine Intelligence, Vol.30,no.2,pp.348-353,Feb2008

[14] Hossein Yousefi-Banaem, Saeed Kermani,Davood Khodadad, "An Improved Spatial FCM Algorithm forCardiac Image Segmentation", 13thIranian Conference on Fuzzy Systems (IFSC), 2013

[15] Abdenour Mekhmoukh, Karim Mokrani and Mohamed Cheriet, "A modified Kernelized Fuzzy C-Means algorithm for noisyimages segmentation: Application to MRI images",IJCSI International Journalof Computer Science Issues, Vol. 9, Issue 1, No 1, January 2012.

[16] K. M. Bataineh, M. Naji, M. Saqer,"A Comparison Study between Various Fuzzy Clustering Algorithms", Jordan Journal of Mechanicaland Industrial Engineering, Volume 5, Number 4, Aug. 2011 ISSN 1995-6665 Pages 335 – 343