Efficient Clustering with Multi-class Point Identification

U. Maulik
Department of CSE, Jadavpur University
Jadavpur 700 032, INDIA
e-mail: ujjwal_maulik@yahoo.com,dr_maulik@jdvu.ac.in,

A. Mukhopadhyay
Department of Computer Science and Engineering
University of Kalyani, Kalyani 741 235, INDIA
e-mail: anirbanbuba@yahoo.com,

S. Bandyopadhyay
Machine Intelligence Unit
Indian Statistical Institute, Kolkata 700 108, INDIA
e-mail: sanghami@isical.ac.in

G. Chakraborty and B. Chakraborty
Iwate Prefectural University
Japan
e-mail: {goutam,basabi}@soft.iwate-pu.ac.jp

Abstract

This article deals with the development of an improved clustering technique that is based on the identification of points having significant membership to multiple classes. Cluster assignments of such points are difficult, and they often affect the actual partitioning of the data. As a consequence, it may be more effective if the points that are associated with maximum confusion regarding their cluster assignments are first identified and excluded from consideration while clustering. Thereafter, these points may be assigned to one of the identified clusters based on a nearest neighbor criterion. Such an approach is described in the present article. The well-known fuzzy C-Means (FCM) algorithm and a recently proposed genetic scheme are utilized as the underlying clustering technique when the number of clusters is known *a priori*. The performance of the proposed clustering algorithm has been compared with the average linkage hierarchical clustering algorithm, in addition to the FCM and genetic clustering scheme, to prove its effectiveness on a variety of data sets.

Keywords: Cluster validity indices, fuzzy clustering, multi-class membership, Minkowski score, genetic algorithm.

1 Introduction

Clustering [JD88]-[Har75] is a popular unsupervised pattern classification technique which partitions the input space into K regions depending on some similarity/dissimilarity metric. Any clustering technique is intended to evolve a $K \times n$ partition matrix $U(X)$ of a data set X ($X = \{x_1, x_2, \ldots, x_n\}$) in N-dimensional space $\mathcal{R}^N$, representing its partitioning into K clusters $(C_1, C_2, \ldots, C_K)$. Let $(z_1, z_2, \ldots, z_K)$ represent the K cluster centroids. The partition matrix $U(X)$ can be represented as $U = [\mu_{k,j}]$, $k = 1, \ldots, K$, and $j = 1, \ldots, n$, where $\mu_{k,j}$ is the membership of pattern x_j to cluster C_k. In case of fuzzy clustering [Bez81, Dun74], $0 \leq \mu_{k,j} \leq 1$, and $\sum_{k=1}^{K} \mu_{k,j} = 1$ for $j = 1, \ldots, n$. Greater value of $\mu_{k,j}$ implies that the degree of belongingness of point x_j to cluster C_k is more. Fuzzy C-Means (FCM) [Bez81, PB95] is a widely used technique that uses the principles of fuzzy sets to evolve a partition matrix $U(X)$ while minimizing a fuzzy

functional given in Eqn. 1. The FCM algorithm often gets stuck at suboptimal solutions based on the initial configuration of the system. In order to overcome this, a genetic algorithm [Gol89] based fuzzy clustering technique has been proposed in [MB03].

It has been observed that, in general, the performance of clustering algorithms degrade with more and more overlaps among clusters in a data set. This is because in such situations there are several points in the data set which have significant belongingness to more than one cluster, leading to a lot of confusion regarding their cluster assignments. As such, it may be beneficial if these points are first identified and excluded from consideration while clustering the data set. They could, thereafter, be assigned to one of the clusters using some nearest neighbor criterion. Such a clustering algorithm is proposed in this article, that utilizes the concept of points having significant multiclass membership. Performance of the proposed clustering method is compared with the average linkage method [TG74], in addition to the conventional FCM and the genetic algorithm based method, for several artificial and real-life data sets in terms of the Minkowski scores [BHG03].

2 Clustering Techniques

In this section, some clustering algorithms used in the article are described briefly.

2.1 Fuzzy C-means

Fuzzy C-Means (FCM) [Bez81] is a widely used technique that uses the principles of fuzzy sets to evolve a partition matrix $U(X)$ while minimizing the measure

$$J_m = \sum_{j=1}^{n}\sum_{k=1}^{K} \mu_{k,j}^m D^2(z_k, x_j), \qquad (1)$$

where n is the number of data objects, K represents number of clusters, μ is the fuzzy membership matrix (partition matrix) and m denotes the fuzzy exponent. Here x_j is the j^{th} data point and z_k is the center of k^{th} cluster, and $D(z_k, x_j)$ denotes the distance of point x_j from the center of the k^{th} cluster. In this article, the Euclidean norm is taken as a measure of the distance between two points.

FCM algorithm starts with random initial K cluster centers, and then at every iteration it finds the fuzzy membership of each data points to every cluster using the following equation [Bez81]:

$$\mu_{i,k} = \frac{1}{\sum_{j=1}^{K} (\frac{D(z_i,x_k)}{D(z_j,x_k)})^{\frac{2}{m-1}}}, \qquad (2)$$

for $1 \leq i \leq K$; $1 \leq k \leq n$, where $D(z_i, x_k)$ and $D(z_j, x_k)$ are the distances between x_k and z_i, and x_k and z_j respectively. m is the weighting coefficient. (Note that while computing $\mu_{i,k}$ using Eqn. 2, if $D(z_j, x_k)$ is equal to zero for some j, then $\mu_{i,k}$ is set to zero for all $i = 1, \ldots, K$, $i \neq j$, while $\mu_{j,k}$ is set equal to one.) Based on the membership values, the cluster centers are recomputed using the following equation [Bez81]:

$$z_i = \frac{\sum_{k=1}^{n} (\mu_{i,k})^m x_k}{\sum_{k=1}^{n} (\mu_{i,k})^m}, \quad 1 \leq i \leq K. \qquad (3)$$

The algorithm terminates when there is no further change in the cluster centers. Finally, each data point is assigned to the cluster to which it has maximum membership.

2.2 Genetic Algorithm Based Clustering

Here we briefly discuss the use of genetic algorithms (GAs) for clustering. In GAs, the parameters of the search space are encoded in the form of strings (called *chromosomes*). A collection of such strings is called a *population*. Initially a random population is created, which represents different points in the search space. An *objective/fitness* function is associated with each string that represents the degree of *goodness* of the solution encoded in the string. Based on the principle of survival of the fittest, a few of the strings are selected and each is assigned a number of copies that go into the mating pool. Biologically inspired operators like *crossover* and *mutation* are applied on these strings to yield a new population. The process of selection, crossover and mutation continues for a fixed number of generations or till a termination condition is satisfied.

In GA-based fuzzy clustering, the chromosomes are made up of real numbers which represent the coordinates of the centers of the partitions [MB00]. If chromosome i encodes the centers of K clusters in N dimensional space then its length

l is $N * K$. For initializing a chromosome, the K centers are randomly selected points from the data set while ensuring that they are distinct.

The fitness of a chromosome indicates the degree of goodness of the solution it represents. In this article we use the Xie-Beni (XB) cluster validity index [XB91] for this purpose. The XB index is defined as a function of the ratio of the total variation σ to the minimum separation sep of the clusters. Here σ and sep can be written as $\sigma(U, Z; X) = \sum_{i=1}^{K} \sum_{k=1}^{n} \mu_{i,k}^2 D^2(z_i, x_k)$, and $sep(Z) = \min_{i \neq j}\{||z_i - z_j||^2\}$, where $||.||$ is the Euclidean norm, and $D(z_i, x_k)$, as mentioned earlier, is the distance between the pattern x_k and the cluster center z_i. The XB index is then written as

$$XB(U, Z; X) = \frac{\sigma(U,Z;X)}{n \ sep(Z)} = \frac{\sum_{i=1}^{K}(\sum_{k=1}^{n} \mu_{i,k}^2 D^2(z_i,x_k))}{n(\min_{i \neq j}\{||z_i - z_j||^2\})}. \quad (4)$$

Note that when the partitioning is compact, value of σ should be low while sep should be high, thereby yielding lower values of the Xie-Beni (XB) index. The objective is therefore to minimize the XB index for achieving proper clustering.

Given a chromosome, the centers encoded in it are first extracted. Let the chromosome encode K centers, and let these be denoted as $z_1, z_2, \ldots, z_K$. The membership values $\mu_{i,k}$, $i = 1, 2, \ldots, K$ and $k = 1, 2, \ldots, n$ are computed as in Eqn. 2. The corresponding XB index is computed as in Eqn. 4. The fitness function for a chromosome is then defined as $\frac{1}{XB}$. Note that maximization of the fitness function will ensure minimization of the XB index. Subsequently, the centers encoded in a chromosome are updated using Eqn. 3 [MB00].

Conventional proportional selection implemented by the roulette wheel strategy is applied on the population of strings. The standard single point crossover is applied stochastically with probability μ_c. The cluster centers are considered to be indivisible, i.e., the crossover points can only lie in between two clusters centers.

Each gene position of a chromosome is subjected to mutation with a fixed probability μ_m, resulting in the overall perturbation of the chromosome. A number δ in the range [0, 1] is generated with uniform distribution. If the value at a gene position is v, after mutation it becomes $(1 \pm 2 * \delta) * v$, when $v \neq 0$, and $\pm 2 * \delta$, when $v = 0$. The '+' or '-' sign occurs with equal probability. Note that because of mutation more than one cluster center may be perturbed in a chromosome.

The algorithm is termination after it has executed a fixed number of generations. The elitist model of GAs has been used, where the best string seen so far is stored in a location within the population. The best string of the last generation provides the solution to the clustering problem.

2.3 Average Linkage Hierarchical Clustering Technique

Agglomerative clustering techniques [TG74] begin with singleton clusters, and combine two least distant clusters at every iteration. Thus in each iteration two clusters are merged, and hence the number of clusters reduces by one. This proceeds iteratively in a hierarchy, providing a possible partitioning of the data at every level. When the target number of clusters (K) is achieved, the algorithms terminate. Single, average and complete linkage agglomerative algorithms differ only in the linkage metric used. For the single linkage algorithm, the distance between two clusters C_i and C_j is computed as the smallest distance between all possible pairs of data points p and q, where $p \in C_i$ and $q \in C_j$. For average and complete linkage algorithms, the linkage metric is taken as the average and largest distances respectively.

3 The Proposed Technique

In this section, the proposed clustering algorithm is described in detail. First, the technique for identifying the multi-class points has been discussed.

3.1 Identification of Multi-class Points

FCM, as well as genetically guided fuzzy clustering techniques, assigns membership values to each pattern that indicates the degree of belongingness to different clusters. This results in a fuzzy membership matrix $U(X)$. The fuzzy partitioning matrix may be used to find out the multi-class points i.e., the points which are situated at the overlapping regions of two or more clusters, and hence they cannot be assigned to any cluster with a reasonable amount of certainty. Suppose some clustering algorithm partitions the data set $X = \{x_1, x_2, \ldots, x_n\}$ into K clusters

$\{C_1, C_2, \ldots, C_K\}$, and produces the partition matrix $U(X)$ where $U = [\mu_{k,j}]$, $k = 1, \ldots, K$, and $j = 1, \ldots, n$. Let us assume that a particular point x_j has the highest membership value for cluster q, and next highest membership value for cluster r. i.e., $\mu_{q,j} \geq \mu_{r,j} \geq \mu_{k,j}$ where $k = 1, \ldots, K$, and $k \neq q, k \neq r$. Suppose the difference in the membership values $\mu_{q,j}$ and $\mu_{r,j}$ is δ_j, i.e., $\mu_{q,j} - \mu_{r,j} = \delta_j$. It is evident that smaller the value of δ_j, greater is the confusion regarding the class assignment of the point x_j. Thus a threshold τ is selected, such that for every x_j, $j = 1, 2, \ldots, n$, if $\delta_j < \tau$, then x_j is said to be a multi-class point. Let

$$B = \{x_j |\ \delta_j < \tau,\ j = 1, 2, \ldots, n\}. \qquad (5)$$

3.2 The Clustering Algorithm

The proposed algorithm has two different stages. In the first stage, the underlying data set is partitioned using either FCM or genetically guided fuzzy clustering algorithms. From the resulting partition matrix the multi-class points are identified using the technique discussed in Section 3.1. In the subsequent stage, the proposed technique excludes these points from the data set and reclusters the remaining points into K clusters. Finally, in the resulting cluster solution, each excluded point is assigned to one of the clusters using nearest neighbor rule.

4 Experimental Results

The experimental results of clustering using the proposed approach are provided for two artificial data sets (*Data 1* and *Data 2*), and two real-life data sets (*Iris* and *Cancer*). These are first described below, followed by the performance measure used for comparison. Finally, the results are provided.

4.1 Data Sets

Data 1: This is a overlapping two dimensional data set where the number of clusters is five. It has 250 points. The value of K is chosen to be 5. The data set is shown in Fig. 1(a).

Data 2: This is an overlapping two dimensional triangular distribution of data points having nine classes where all the classes are assumed to have equal *a priori* probabilities ($= \frac{1}{9}$). It has 900 data points. This data set is shown in Fig. 1(b).

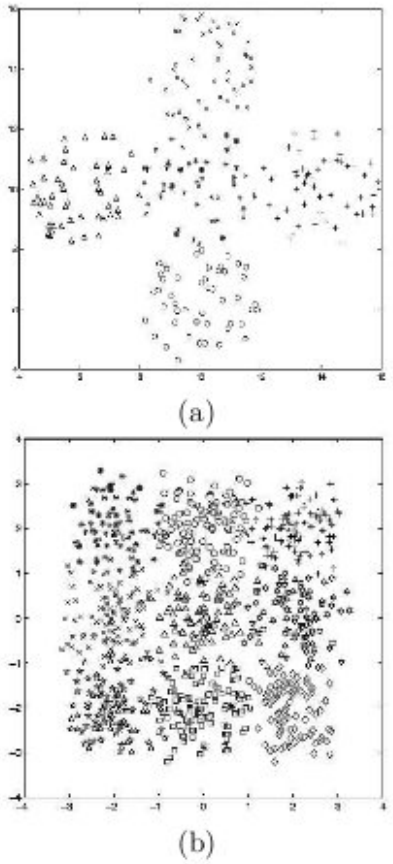

Figure 1: Data Sets: (a) *Data 1* (b) *Data 2*

Iris: This data consists of 150 patterns divided into three classes of Iris flowers namely, Setosa, Virginia and Versicolor. The data is in four dimensional space (sepal length, sepal width, petal length and petal width).

Cancer: It has 683 patterns in nine features (clump thickness, cell size uniformity, cell shape uniformity, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli and mitoses), and two classes malignant and benign. The two classes are known to be linearly inseparable.

The *Iris* and *Cancer* data sets are available in [http://www.ics.uci.edu/~mlearn/MLRepository.html].

Note that as two fuzzy clustering algorithms are used in two levels of the proposed algorithm, there may be four combinations possible in these two steps. These are referred to as FCM-FCM, FCM-GA,GA-FCM and GA-GA.

4.2 Performance Metric

Here, the performances of the clustering algorithms are evaluated in terms of the *Minkowski score* [BHG03]. A clustering solution for a set of n elements can be represented by an $n \times n$ matrix C, where $C_{i,j} = 1$ if point i and j are in the same cluster according to the solution, and $C_{i,j} = 0$ otherwise. The Minkowski score (MS) of a clustering result C with reference to T, the matrix corresponding to the true clustering, is defined as

$$MS(T,C) = \frac{\| T - C \|}{\| T \|} \quad (6)$$

where $\| T \| = \sqrt{\sum_i \sum_j T_{i,j}}$.

The Minkowski score is the normalized distance between the two matrices. Lower Minkowski score implies better clustering solution, and a perfect solution will have a score zero.

4.3 Results

Table 1 shows the comparative Minkowski scores obtained by the different algorithms for the four data sets. The values of τ chosen for the different data sets is shown in Table 2. Although the performance of the algorithms were relatively robust to the exact choice of τ, intuitively it is evident that as the number of clusters in a data set increases, τ should be made smaller. This is a consequence of the condition that the sum of the membership values of a data point to the different clusters equals 1. Empirical analysis also confirmed this fact. Fig. 2 shows, for the purpose of illustration, the points identified by the proposed method as having multi-class memberships for *Data 2*. As is evident from the figure, these points are situated at the overlapping regions of more than one cluster.

As can be seen from Table 1, irrespective of the clustering method used (viz., FCM or GA) in the proposed algorithm, the performance gets improved after the application of the second level of clustering. For example, in case of *Data 1*, the Minkowski score after the application of GA in the first level is 0.4398 while this gets improved to 0.4243 at the end. Similarly, when FCM is applied in the first level, the score is 0.4404 which gets improved to 0.4269 (with FCM in the second level) and 0.3851 (with GA in the second level). The final Minkowski scores are also better than those obtained using the average linkage method. The results demonstrate the utility of adopting the approach presented in this paper, irrespective of the clustering method used.

Data Set	1st stage algo.	1st stage MS	2nd stage algo.	Final MS	K	Avg. Link.
Data 1	GA	0.4398	FCM	0.4243		
			GA	0.4243		
	FCM	0.4404	FCM	0.4269	5	0.4360
			GA	0.3851		
Data 2	GA	0.5348	FCM	0.5290		
			GA	0.5304		
	FCM	0.5314	FCM	0.5290	9	0.6516
			GA	0.5304		
Iris	GA	0.5583	FCM	0.5307		
			GA	0.5307		
	FCM	0.5987	FCM	0.5666	3	0.5666
			GA	0.5307		
Cancer	GA	0.3936	FCM	0.3666		
			GA	0.3556		
	FCM	0.3926	FCM	0.3666	2	0.4445
			GA	0.3556		

Table 1: Comparative results for the different data sets

Data Set	K	τ
Data 1	5	0.225
Data 2	9	0.2
Iris	3	0.25
Cancer	2	0.3

Table 2: Choice of τ for the different data sets

5 Discussions and Conclusions

A fuzzy clustering method that is based on the identification of points which are associated with the maximum confusion regarding their cluster assignments has been proposed in this article. Experimental results indicate that this approach, with a suitable choice of a single parameter, is likely to yield better results irrespective of the actual clustering technique adopted.

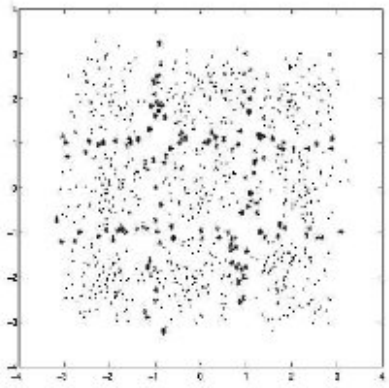

Figure 2: *Data 2* with the points identified as belonging to multiple classes marked as '*'

There are several directions in which this work needs to be extended in the future. First of all, a detailed comparison with other competing techniques needs to be carried out. Secondly, a sensitivity analysis of the choice of τ should be performed. Finally, a theoretical analysis needs to be carried out to provide a functional form for τ.

References

[Bez81] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*, Plenum, New York, 1981.

[BHG03] A. Ben-Hur and I. Guyon, *Detecting stable clusters using principal component analysis in methods in molecular biology*, Humana press, 2003, pp. 159–182.

[Dun74] J. C. Dunn, *Well separated clusters and optimal fuzzy partitions*, J. Cyberns. **4** (1974), 95–104.

[Eve93] B. S. Everitt, *Cluster analysis*, Halsted Press, third edition, 1993.

[Gol89] D. E. Goldberg, *Genetic algorithms in Search, optimization and machine learning*, Addison-Wesley, New York, 1989.

[Har75] J. A. Hartigan, *Clustering algorithms*, Wiley, 1975.

[JD88] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[MB00] U. Maulik and S. Bandyopadhyay, *Genetic algorithm based clustering technique*, Pattern Recognition **33** (2000), 1455–1465.

[MB03] ———, *Fuzzy partitioning using a real-coded variable-length genetic algorithm for pixel classification*, IEEE Transactions on Geoscience and Remote Sensing **41** (2003), no. 5, 1075– 1081.

[PB95] N. R. Pal and J. C. Bezdek, *On cluster validity for the fuzzy c-means model*, IEEE Transactions on Fuzzy Systems **3** (1995), 370–379.

[TG74] J. T. Tou and R. C. Gonzalez, *Pattern recognition principles*, Addison-Wesley, Reading, 1974.

[XB91] X. L. Xie and G. Beni, *A validity measure for fuzzy clustering*, IEEE Transactions on Pattern Analysis and Machine Intelligence **13** (1991), 841–847.