



A Modified Fuzzy C-Mean for Clustering Data with Attributes of Different Degrees of Importance

M. Namazifar¹, H. Javaheri Neyestanaki²

¹Computer Engineering and Information Technology Department, Amirkabir University of Technology

²Department of Mining and Metallurgy, Amirkabir University of Technology

Hafez St., Tehran, Iran

namazifar@ce.aut.ac.ir,

javaheri@aut.ac.ir

Abstract

In this paper we are about to modify the fuzzy c-mean (FCM) algorithm in a way which this algorithm be able to handle the cases in which the data set is multidimensional, and dimensions are not of equal degree of importance. In such cases, clusters obtained by FCM are not logically satisfying. So some modifications are absolutely needed.

Keywords: Clustering, Fuzzy C-Mean, Distance measure, Degree of Importance.

1. Introduction

The objective of cluster analysis is to classify experimental data in a certain number of sets where the elements of each set should be as similar as possible and dissimilar from those of other sets [1]. Since Zadeh [2] proposed fuzzy sets, which produced the idea of overlapping membership in two or more sets described by a membership function, fuzzy clustering has been widely studied and applied in various areas [3,4].

In the fuzzy clustering literature FCM is the most used method [5]. This method is applicable to data sets with multiple dimensions. In such spaces FCM assigns each single datum to clusters to some degrees based on its distance from each cluster center. In FCM method we implicitly assume that dimensions of data are equivalent, or more explicitly they are of equal degree of importance (*DOI*). But in real world it is highly plausible that this assumption be not a realistic one. For instance in a two-dimensional space with t and s dimensions, proximity of two data points in dimension s may be more important than the proximity of these points in dimension t .

To be able to deal with such cases, we propose a method that takes into account the degree of importance of each attribute in the data set that will be clustered. After a brief review of the FCM in section 2, a number of attribute ranking methods is described in section 3. These methods will be used in determining *DOI* of each attribute. In

section 4 distance measures are studied and a new one is proposed to handle the different importance of dimensions. In section 5 we proposed the modification of FCM for clustering spaces with in-equivalent dimensions. It should be notified that in this investigation attributes, dimensions and features are synonyms.

2. Fuzzy C-Mean Algorithm

Fuzzy c -mean (FCM) is an unsupervised clustering algorithm that has been applied successfully to a wide range of problems involving feature analysis, clustering and classifier design. FCM has a wide domain of applications such as agricultural engineering, astronomy, chemistry, geology, image analysis, medical diagnosis, shape analysis, and target recognition [6]. Unlabeled data are classified by minimizing an objective function based on a norm and clusters prototype. Although the description of the original algorithm dates back to 1973 [7,8] derivatives have been described with modified definitions for the norm and prototypes for the cluster centers [9,10,11]. The FCM minimizes an objective function J_m , which is the weighted sum of squared errors within groups and is defined as follows:

$$J_m(U, V; X) = \sum_{k=1}^n \sum_{i=1}^c u_{ik}^m \|x_k - v_i\|_A^2, \quad 1 < m < \infty \quad (2.1)$$

Where $V=(v_1, v_2, \dots, v_c)$ is a vector of unknown cluster prototype (centers) $v_i \in R^p$. The value of u_{ik} represents the grade of membership of data point x_k of set $X= \{x_1, x_2, \dots, x_n\}$ to the i th cluster. The inner product defined by a norm matrix A defines a measure of similarity between a data point and the cluster prototypes. A non-degenerate fuzzy c -mean partition of X is conveniently represented by a matrix $U=[u_{ik}]$. It has been shown by

Bezdek [3] that if $\|x_k - v_i\|_A^2 > 0$ for all i and k , then (U, V) may minimize J_m only, when $m > 1$ and

$$v_i = \frac{\sum_{k=1}^n (u_{ik})^m x_k}{\sum_{k=1}^n (u_{ik})^m} \quad \text{for } 1 \leq i \leq c, \quad (2.2)$$

$$u_{ik} = \frac{1}{\sum_{j=1}^c \left(\frac{\|x_k - v_i\|_A^2}{\|x_k - v_j\|_A^2} \right)^{\frac{1}{m-1}}} \quad (2.3)$$

for $1 \leq i \leq c$, $1 \leq k \leq n$.

Among others, J_m can be minimized by the Picard iteration approach. This method minimizes J_m by initializing the matrix U randomly (or predefined) and computing the cluster prototypes (Eq. (2.2)) and the membership values (Eq. (2.3)) after each iteration. The iteration is terminated when it reaches a stable condition. This can be defined for example, when the changes in the cluster centers or the membership values at two successive iteration steps is smaller than a predefined threshold value. The FCM algorithm always converges to a local minimum or a saddle point. A different initial guess of u_{ij} may lead to a different local minimum. Finally, to assign each data point to a specific cluster, defuzzification is necessary, e.g., by attaching a data point to a cluster for which the value of the membership is maximal [12].

3. Determining DOI of Attributes

In section 1 we mentioned that we propose a clustering method for a data set with in-equivalent dimensions (attributes), which means that data with attributes of different DOI should be clustered. A key question that arises here is how can we determine the degree of importance of each attribute. In other words, we are about to assign a weight to each attribute so that the weight of each attribute determines the DOI of that attribute.

To determine the DOI of attributes of a data set two major approaches can be adopted: Direct approach and Indirect approach. In direct approach we determine the DOI of each attribute based on negotiation with an expert individual who has enough experience and knowledge in the field that is the subject of clustering. On the other hand, in indirect approach we use the data set itself to determine the DOI of its dimensions. We will discuss more about these approaches in next lines.

Direct approach: As is described above, in direct approach by negotiating with an expert, we choose DOI of each attribute. This approach has some advantages and some drawbacks. In some cases, using the data set itself to determine the DOI of each attribute may fail to achieve the real DOI's, and direct approach should be adopted to

determine the DOI of each attribute. Figure 1 demonstrates a situation this case happens.

Suppose figure 1 shows a data set in which DOI of attribute a is two times DOI of attribute b in reality. Since indirect approach uses the position of data points in the data space to determine the DOI of attributes, using data set itself to determine the DOI of attributes a and b (indirect approach) will lead to equal DOI's for a and b . Although this case (data set with homogeneously and equidistantly distributed data points) rarely happens in real world and is somehow an exaggerated one, it shows that, sometimes, direct approach is the better choice.

On the other hand, direct approach has its own drawbacks. We cannot guarantee that the behaviors that are observed by a human expert and used to determine the DOI's include all situations that can occur due to disturbances, noise, or plant parameter variations. Also suppose situation in which there is no human expert for negotiation to determine DOI's. How does this problem should be dealt with?

Structure the signal can be found using linear transforms. This approach does not take into account that the system has some structure. In the time domain, filtering is a linear transformation. The Fourier, Wavelet, and Karhunen-Loeve transforms have compression Capability and can be used to identify some structure in the signals. When we are using these transforms, we do not take into account any structure in the system.

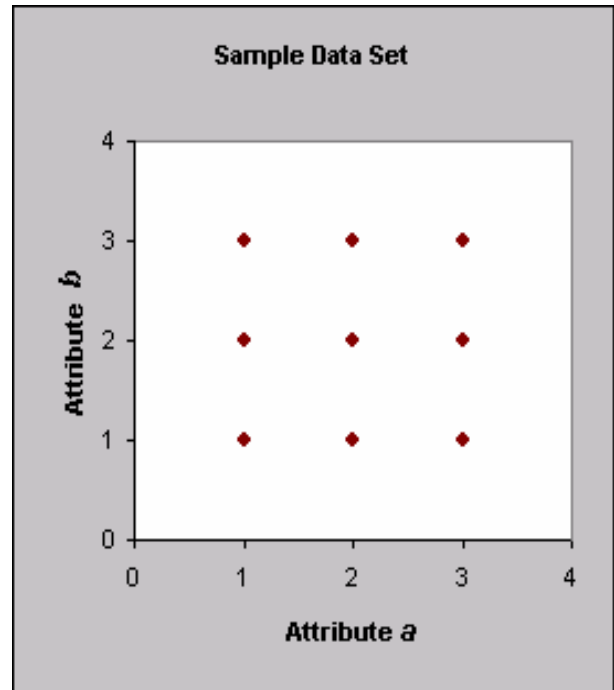


Figure 1. A data set with two dimensions

Indirect approach: Several methods based on fuzzy set theory [14-16], artificial neural network (ANN) [17-21], fuzzy-rough set theory [22], and neuro-fuzzy methods [23] and [24] have been reported for feature evaluation. Some of the mentioned methods just rank attributes, but with some modifications they will be able to calculate the DOI of attributes.

