Expert Systems with Applications 40 (2013) 842–849

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

FUAT – A fuzzy clustering analysis tool

A. Selman Bozkir, Ebru Akcapinar Sezer *

Hacettepe University, Engineering Faculty, Department of Computer Engineering, 06800 Beytepe, Ankara, Turkey

ARTICLE INFO

Keywords:
Clustering analysis
Fuzzy c-means clustering
Validity index
Visual analysis

ABSTRACT

As it is known, fuzzy clustering is a kind of soft clustering method and primarily based on idea of segmenting data by using membership degrees of cases which are computed for each cluster. However, most of the current fuzzy clustering modules packaged in both open source and commercial products have lack of enabling users to explore fuzzy clusters deeply and visually in terms of investigation of different relations among clusters. Furthermore, without a decision maker or an expert, it is hard to decide the number of clusters in fuzzy clustering studies. Therefore, in this study, a desktop software, namely FUAT, is developed to analyze, explore and visualize different aspects of obtained fuzzy clusters which are segmented by fuzzy c-means algorithm. Moreover, to obtain and inform possible natural cluster number, FUAT is equipped with Expectation Maximization algorithm.

1. Introduction

Clustering is an unsupervised classification method to partition a collection of multivariate data points into meaningful groups, where all members with in a group represent similar characteristics and data points between different groups are dissimilar to each other (Tsai & Lin, 2011). There are many methods and algorithms for clustering, based on crisp, fuzzy, probabilistic and possibly approaches (Rezaee, 2010), i.e. k-means, c-means, hierarchical clustering. Fuzzy c-means clustering algorithm (FCM) is one of the popular clustering algorithms. FCM combines the c-means approach with the handling of the existing fuzziness in the data. This combination makes it more powerful, because fuzziness of the data affects the results in a disadvantageous way when creating crisp partitions. In general, soft clustering techniques aim to eliminate this situation and also, FCM is a kind of soft clustering method which is based on fuzzy set theory (Zadeh, 1965). In practical applications of clustering algorithms, several problems must be resolved, including determination of the number of clusters and evaluation of the quality of the partitions (Rezaee, 2010). Moreover, as stated by (Kim, Lee, & Lee, 2004), three major difficulties were drawn attention in fuzzy clustering: (1) determining the natural number of clusters to be created (most algorithms require the user to specify the number of clusters); (2) choosing the initial cluster centroids (most algorithms choose a random selection because such a selection is sure to converge the iterative process); (3) handling data characterized by large variations in cluster shape, cluster density, and the number of points in different clusters (Kim et al., 2004).

In this study, the tool, FUAT (fuzzy clustering analysis tool), is proposed to explore the clusters created with FCM clustering. The reason for the development of FUAT is sourced from the reported difficulties of FCM. First, FCM requires the cluster number as an input parameter but, to know this number is difficult for the decision maker in fact. Because, to forecast the distribution of the data points, which is gained from real world, in the space is hard job and sometimes impossible. In FUAT, natural clustering is embedded to give an advice to the user about possible cluster number. Second, the initial clusters have great effect on the resulted clusters. However, neither getting the resulted cluster centroids nor presentation of data with clusters numbers, and membership degrees is enough to assess clustering performance. Since, size or densities of the clusters, saturation and frequencies of the membership degrees in the clusters, closeness between clusters, intersection size or densities between clusters are required for performing detailed analyses on the clusters and parameters and, assessment of the clustering performance.

The problems and the critical points about fuzzy clustering have been discussed in literature, especially these studies are focused on the subject of validity index. For example partition coefficient (PC) (Bezdek, 1974a) and partition entropy (PE) (Bezdek, 1974b) are basic, simple, but efficient indices which are based on the fuzzy membership values of fuzzy partitions. Moreover, researchers have suggested many cluster validity indices that include both fuzzy membership values and the information of structures (Zalik, 2010). Most indices about validity employ compactness and separation concepts. Compactness is related with the closeness within the cluster, and separation is related with the isolation of clusters between each other. In other words, validity index for fuzzy clustering tries to reflect the ratio of overcoming the difficulties

* Corresponding author. Tel.: +90 3122977500.
E-mail address: ebru@hacettepe.edu.tr (E.A. Sezer).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.05.038

specified in Rezaee (2010). In fact, validity index in a necessity, because of the black box usage of fuzzy clustering algorithms and, their dependencies to initial parameters and structure. Before clusters obtained by fuzzy clustering algorithm, validity checking can be done by numerically by using selected validity index.

In fact FUAT has a complementary approach to concept of validity index. It is a tool and shows many characteristics of resulted clusters (compactness, separation, overlapping, case distribution, density) visually. In other words, by FUAT, we tried to convert effectively FCM based clusters from black box to transparent boxes for the users. Especially, we concentrate on the creating ability of the clusters analysis separately and all together for helping users to overcome difficulties of FCM usage as a black box. In FUAT design all characteristics of FCM are kept and different data types (integer, real) are supported.

2. Theory

In this study, two important clustering schemes are employed together. Fuzzy c-Means (FCM) and Expectation Maximization (EM) based clustering methods are used because of their soft clustering behaviors. Their major characteristics are explained below.

2.1. Fuzzy c-means clustering

Generalized Fuzzy c-means (FCM) (Bezdek, 1981) is one of the most popular unsupervised fuzzy clustering algorithm, which is widely used in pattern recognition, image recognition, gene classification, etc (Jingwei & Meizhi, 2008). As can be understood from the name of FCM, it is based on Zadeh's (Zadeh, 1965) fuzzy set theory and applies c-means clustering approach. By FCM, fuzzy clusters are constructed in that way, ith data, x_i, belongs to jth cluster, F_j, with degree of $\mu_{Fj}(x_i)$. In FCM, data points are partitioned into the c clusters by the minimization of the distance between data points, and the fuzzy cluster centroids iteratively. The general algorithm of FCM is as follows:

```
Specify number of clusters
Do
  Compute centroids of clusters
  For each case
    Compute membership degrees of case to clusters
While convergence criteria is not met
```

In FCM, centroid of ith cluster (c_i) is calculated with Eq. 1.

$$c_i = \frac{\sum_{j=1}^{N}(\mu_{ij})m \cdot x_j}{\sum_{j=1}^{N}(\mu_{ij})m} \quad (1)$$

where N is number of data, x_j is jth case in N and μ_{ij} is the membership degree of jth case to ith cluster. At this point, m parameter is used as the coefficient of distance, and it enables to control fuzziness of the clustering and value of 2 is suggested in (Bezdek, 1999). Let assume that μ_{ij} value is calculated with Eq. (2).

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\|x_j - c_i\|}{\|x_j - c_k\|}\right)^2 \Big/ (m-1)} \quad (2)$$

In general definition of FCM, Euclidean distance measurement is used to calculate distance between vectors, and it converts to the data space spherical view. However, it is possible to replace the distance measurement method with another one. The stopping criterion of FCM is dependent on the change in the vector of membership degrees from previous iteration to completed iteration. In other words, the change in membership values is smaller than convergence criteria value that the algorithm stops.

2.2. Expectation maximization

EM (Expectation Maximization) algorithm is an unsupervised clustering method based on discovering the appropriate parameters of a particular statistical model which defines the data. The employed model in this process is called as mixture models which view the data as a set of cases from a mixture of different probability distributions and are being modeled by using a number of statistical distributions that each represents a cluster (Tan, Steinbach, & Kumar, 2005), Moreover, as stated previously in (Tan et al., 2005) that, parameters of each distribution provide a description of the corresponding cluster.

In general lines, EM procedure consists of two important iterative steps: E-step and M-step. Probability of being a member of each case to each cluster is computed at E-Step (*Expectation*). At the next stage (*Maximization*), parameter vector of the probability distribution of each cluster is re-approximated. This iteration based mechanism finishes when the maximum number of iterations, or accepted error range to converge is reached. As a result of this procedure, total natural clusters can be obtained without specifying a cluster number.

Consequently, EM based clustering is segmentation method which utilizes maximum likelihood concept. On the other hand, similar to fuzzy clustering, it owns soft segmentation characteristic because of a point being member of more than one cluster with certain probability. Due to these facts and let the users know the probably true number of clusters in data, EM based clustering schema is included to FUAT. More detailed explanations about EM based clustering can be found at (DMR, 2011; Tan et al., 2005).

3. Components utilized

In development of the software subjected to this study, various components are utilized. They are listed below:

3.1. R

In this study, R *(http:www.r-project.org)*, a famous and well known statistical computing program, is employed for EM algorithm usage. As (Zupan & Demsar, 2008), reported that, R involves many techniques for statistics, predictive modeling and data visualization, and has become a defacto standard at open source library for statistical computing. The main benefit of R is having a script language inherited from *S* (Becker & Chambers, 1984), which enables users to program what they need in clear way. Further, R has got an extensive variety of freely accessible modules for various purposes located at "The Comprehensive R Archive Network" (CRAN, http://cran.r-project.org).

As the R mainly supports command line scripting, it has two important advantages: (1) whole analysis procedure can be operated by clearly stated definitions, and they can be stored for later use; (2) R can be accessed and directed via COM interfaces in supported programming languages and platforms such as C++, VB or .NET. Therefore, a seamlessly integrated component named R(D)COM which bridges a wide range of different languages (e.g., C#, C++, Python, VBA, VB, Java) to R is developed by Statconn (Statconn, 2011). Further, it supports various numbers of platforms (COM/DCOM, .NET, Uno, C, Web Services SOAP/http) to create and integrate solutions (Statconn, 2011). Statconn also provides other similar modules called RExcel which enables usage of R functions in Excel natively and ROOo at Open Office environment.