Applied Soft Computing 13 (2013) 1592–1607

Contents lists available at SciVerse ScienceDirect

Applied Soft Computing

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

A multivariate fuzzy *c*-means method

Bruno A. Pimentel, Renata M.C.R. de Souza*

Centro de Informática, Av. Jornalista Anibal Fernandes, s/n – Cidade Universitária 50.740-560, Recife (PE), Brazil

ARTICLE INFO

Article history:
Received 10 July 2012
Received in revised form
27 November 2012
Accepted 30 December 2012
Available online 11 January 2013

Keywords:
Fuzzy *c*-means method
Unsupervised pattern recognition
Clustering
Membership degree

ABSTRACT

Fuzzy *c*-means (FCMs) is an important and popular unsupervised partitioning algorithm used in several application domains such as pattern recognition, machine learning and data mining. Although the FCM has shown good performance in detecting clusters, the membership values for each individual computed to each of the clusters cannot indicate how well the individuals are classified. In this paper, a new approach to handle the memberships based on the inherent information in each feature is presented. The algorithm produces a membership matrix for each individual, the membership values are between zero and one and measure the similarity of this individual to the center of each cluster according to each feature. These values can change at each iteration of the algorithm and they are different from one feature to another and from one cluster to another in order to increase the performance of the fuzzy *c*-means clustering algorithm. To obtain a fuzzy partition by class of the input data set, a way to compute the class membership values is also proposed in this work. Experiments with synthetic and real data sets show that the proposed approach produces good quality of clustering.

1. Introduction

A growing number of application domains such as pattern recognition, machine learning, data mining, computer vision and computational biology have used clustering algorithms [1–3]. Clustering is a method of unsupervised learning whose objective is to group a set of elements into clusters such that elements within a cluster have a high degree of similarity, while elements belonging to different clusters have a high degree of dissimilarity. Mathematically, the degree of dissimilarity can be measured using, for instance, distance, angle, curvature, symmetry, connectivity or intensity with information from the data set [4]. Hierarchical and partitioning methods are the most popular clustering techniques. Hierarchical clustering find a sequence of partitions where the algorithm starts from one group with all objects and is executed until find singletons groups, or vice versa, whereas partitioning clustering directly divides data objects into some fixed number of clusters [5] using a suitable objective function. An advantage of the partitional method is its ability to manipulate large data sets, since the construction of dendrogram by the hierarchical method may be computationally impractical in some applications.

Partitioning clustering can be divided into hard and fuzzy methods. The concept of fuzzy set was initially explored by Zadeh [6] and applied to clustering by Ruspini [7]. Works with fuzzy set applications in cluster analysis were proposed and applied in several areas [8]. The concept of fuzzy set allowed works on industrial and academic fields [9]. Termini [10] used the definition of fuzzy sets to create an interaction with human sciences. Wong and Lai [11] described the applications of the fuzzy set theory in production and operations management, for example, planning, quality control and artificial intelligence (AI) techniques. Moreover, the work of Wong and Lai is based on the information of 402 articles published on journals between 1998 and 2009 that deal with application of fuzzy set theory techniques. In the medical field, Kuo et al. [12] used fuzzy set theory and, health care failure mode and effect analysis to study patient according the decision-making factors: severity, incidence, and detection.

In the hard approach each element of the data set can be associated to only one cluster, while in the fuzzy approach each element of the data set has a possibility of belonging to all cluster but with different membership degrees. Therefore, the calculation of membership functions is an important problem in fuzzy clustering. When each pattern is associated to the cluster with the largest measure of membership, the fuzzy clustering is equivalent to hard clustering. Three examples of categories of fuzzy used in the cluster analysis are: fuzzy clustering based on fuzzy relation, fuzzy clustering based on objective functions and the fuzzy generalized k-nearest neighbor rule [8]. The most popular fuzzy clustering method based on objective functions is the Fuzzy *c*-means (FCMs) [1,13]. An advantage of the FCM is that it may be used in applications where the clusters are overlapping [8].

There are several papers with related works to theory and applications of the FCM algorithm such as stochastic and numerical theorems, image processing, parameter estimation and many

* Corresponding author.
E-mail address: rmcrs@cin.ufpe.br (R.M.C.R. de Souza).

1568-4946/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2012.12.024

others [8]. Hathaway et al. [14] presented fuzzy c-means methods using general L_p norm distances whose main objective is to increase robustness to outliers. Jajuga [15] presented the fuzzy clustering algorithm based on the L_1 norm. Groenen and Jajuga [16] presented a new fuzzy clustering based on squared Minkowski distance. Oh et al. [17] proposed a new fuzzy clustering algorithm for categorical multivariate data. Xu and Wunsch [18] presented a replacement of the Euclidean norm by a new robust metric in c-means clustering diminishing the weaknesses of the classical FCM. Zhang and Chen [19] proposed the substitution of a kernel-induced distance metric for the original Euclidean distance in the FCM which allows implicitly to perform a nonlinear mapping to a high dimensional feature space. Kummamuru et al. [20] showed a modified version of the algorithm proposed by Oh et al. where this version can be applied in datasets containing large number of documents or words. Pal et al. [21] showed the production of memberships simultaneously in order to avoid various problems of the FCM. De Carvalho et al. [22] presented partitional fuzzy clustering methods based on different adaptive quadratic distances defined by fuzzy covariance matrices. Liu and Xu [23] obtained kernelized fuzzy attribute c-means clustering algorithm with kernel-induced distance. Chen et al. has studied the application of fuzzy methods [24–28]. Tang et al. [29] proposed a new kind of data weighted fuzzy c-means clustering approach. Mei and Chen [30] characterized each cluster by multiple medoids with the help of prototype weight. After, Mei and Chen [31] proposed fuzzy clustering with multiple weighted medoids. An important parameter that influences the robustness of the FCM is the weighting exponent m called fuzzifier. Yu [32] proposed a theoretical approach to measuring this parameter. Wu [33] introduced a new guideline for selecting the weighting exponent m.

In a fuzzy clustering, the memberships are calculated based on distances between clusters and prototypes and is assumed that these memberships are the same for all the features, i.e., the features are considered equally important for the definition of the memberships. However, this model can be restrictive since the features can have dissimilar dispersions and the fuzzy clustering algorithm can have its performance affected. The main contribution of this work is to introduce a new FCM method where the membership values are computed based on the inherent information in each feature. The idea of this method is to find a set of prototypes and a multivariate fuzzy partition that minimizes an objective function. Here, the multivariate memberships allows to take account the intra-structure of the clustering. Section 2 describes a motivation example. In Section 3, a fuzzy c-means method based on membership values computed by feature is presented. In order to validate the proposed method, Section 4 shows experiments with synthetic and real data sets. A comparative study of this method in relation to two fuzzy c-means clustering methods with adaptive distances introduced in [22] is performed. These adaptive distances consider weights for variables according to the intra-structure of the clustering and they are capable of increasing the performance of the fuzzy clustering algorithms since they can be the same for all clusters or different from one cluster to another. Moreover, the proposed method is compared with the Gustafson–Kessel algorithm [34] whose the main feature is its capacity of identify clusters of different sizes and shapes. In Section 5, the concluding remarks are given.

2. A motivation example

Fig. 1 shows a data set with two clusters. Cluster 1 has five patterns indexed from 1 to 5 (labeled as circles) and cluster 2 has five patterns indexed by 6–10 (labeled as stars). According to the fuzzy c-means method, is expected that the item 5 has membership value in cluster 2 slightly greater than the membership value in cluster 1 since the distance between the items 3 and 5 is greater than distance between the items 5 and 8. However, analyzing the feature y_2 separately, we can verify that the membership of the item 5 in cluster 1 is close to 1 and the membership in cluster 2 is close to 0. Whereas, investigating the feature y_1, the membership of the item 5 in cluster 2 has value close to 1, but the membership in cluster 1 has not close to 0 since the dispersion of the feature y_1 is greater than that of the feature y_2. In this case, an approach that computes multivariate membership values is needed.

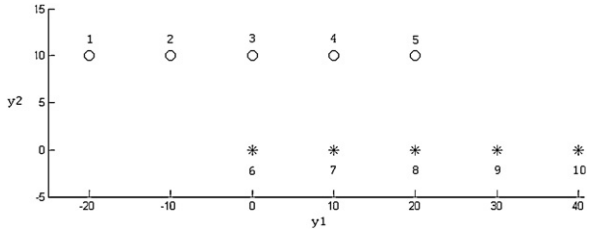

Fig. 1. Data set with two clusters.

The membership values that will be computed in this paper are such that: (a) they are able to take account the statistical information of each feature, in order to improve the performance of the algorithm, (b) they can change at each iteration and, they can be different from one feature to another and from one class to another, and (c) the algorithm is able to derive cluster prototypes optimizing an objective function based on multivariate membership values. In order to obtain a fuzzy partition by class of the input data set, a way of calculating the class membership values is also proposed in this paper.

3. Fuzzy c-means based on multivariate memberships

This section starts with a brief description of the classical fuzzy c-means and, subsequently, it introduces a fuzzy c-means algorithm that is able to find a multivariate fuzzy partition taking into account multiple membership matrices (denote here MFCM).

Consider Ω a set of n patterns indexed by k and formed by p features indexed by j. Each pattern k is represented by a quantitative feature vector $\mathbf{x}_k = (x_{1k}, \ldots, x_{pk})^t$. Let $L = \{\mathbf{y}_1, \ldots, \mathbf{y}_c\}$ be a set of c prototypes associated to a fuzzy partitioning into c clusters. Each prototype of a cluster $C_i (i = 1, \ldots, c)$ is also represented as a quantitative feature vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})^t$.

The Fuzzy C-means (FCM) algorithm proposed by Bezdek [4] aims to find a prototype data set L and a fuzzy partitioning $\mathbf{U} = [u_{ik}](i = 1, \ldots, c)(k = 1, \ldots, n)$ of the data set Ω, by minimizing an objective function given by:

$$J = \sum_{i=1}^{c}\sum_{k=1}^{n} (u_{ik})^m ||\mathbf{x}_k - \mathbf{y}_i||^2. \quad (1)$$

The fuzzy partitioning $\mathbf{U}$ represents a membership matrix where u_{ik} is the membership degree of a given point k which belongs to the cluster i under the following restrictions:

1. $u_{ik} \in [0, 1]$ for all i and k;
2. $0 < \sum_{k=1}^{n} u_{ik} < n$ for all i and
3. $\sum_{i=1}^{c} u_{ik} = 1$ for all k.

In the MFCM clustering algorithm, the memberships degrees are different from one feature to another and from one cluster to another. Therefore, it is necessary to consider an appropriate representation of the memberships and a way to calculate the distance between clusters and prototypes.