Expert Systems With Applications 58 (2016) 76–82

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Fuzzy c-means clustering algorithm for directional data (FCM4DD)

Orhan Kesemen*, Özge Tezel, Eda Özkul

Department of Statistics and Computer Sciences, Karadeniz Technical University, 61080 Trabzon, Turkey

ARTICLE INFO

Article history:
Received 10 February 2015
Revised 19 March 2016
Accepted 20 March 2016
Available online 1 April 2016

Keywords:
Angular difference
Clustering algorithm
Directional data
Fuzzy c-means algorithm

ABSTRACT

Cluster analysis is a useful tool used commonly in data analysis. The purpose of cluster analysis is to separate data sets into subsets according to their similarities and dissimilarities. In this paper, the fuzzy c-means algorithm was adapted for directional data. In the literature, several methods have been used for the clustering of directional data. Due to the use of trigonometric functions in these methods, clustering is performed by approximate distances. As opposed to other methods, the FCM4DD uses angular difference as the similarity measure. Therefore, the proposed algorithm is a more consistent clustering algorithm than others. The main benefit of FCM4DD is that the proposed method is effectively a distribution-free approach to clustering for directional data. It can be used for N-dimensional data as well as circular data. In addition to this, the importance of the proposed method is that it would be applicable for decision making process, rule-based expert systems and prediction problems. In this study, some existing clustering algorithms and the FCM4DD algorithm were applied to various artificial and real data, and their results were compared. As a result, these comparisons show the superiority of the FCM4DD algorithm in terms of consistency, accuracy and computational time. Fuzzy clustering algorithms for directional data (FCM4DD and FCD) were compared according to membership values and the FCM4DD algorithm obtained more acceptable results than the FCD algorithm.

1. Introduction

In the statistical analysis of random sampled data, it is assumed that the data came from a random variable. This random variable can exist in various measure spaces such as metric, time, color, angular etc. Univariate data in the angular (θ) space is called circular data. The directions of the winds; the directions of migrating birds or animals (Chang-Chien, Yang, & Hung, 2010); the orientation of objects in the plane can be held up as circular data. On the other hand, data which does not involve orientation but occurs in periodic process can be analyzed in the same class. Periodic data show the same characteristics within a certain period of time. A student's weekly study schedule and the amount of water consumed daily by living creatures on a yearly basis can be held up as periodic data. Data whose frequency changes periodically can be converted into circular data, although generally it is not circular data.

Generally, angular-based data is called directional data. If directional data has two variables, it is called spherical data. If directional data has more than two variables, it is called hyper-spherical data (Fisher, 1993).

Circular distribution of the data was first examined by von Mises in 1918 (von Mises, 1918). Then, studies on statistical inference from circular data were made by Watson and Williams (1956). After this study, the interest of many researchers in this field increased. Batschelet (1981), Fisher (1993) and Mardia and Jupp (2000) are major books on analysis of circular data which have applications in many fields such as biology, geology, medicine, meteorology, oceanography etc. In addition to these, Money, Helms, and Jolliffe (2003)) investigated circular data for a case study involving sudden infant death syndrome (SIDS). Carta, Bueno, and Ramirez (2008) studied statistical modeling of directional wind speeds. Lee (2010) compiled the methods that have been developed the last 50 years. Baayen, Klugkist, and Mechsner (2012) proposed a test of order-constrained hypotheses for circular data with applications to human movement science. Abraham, Molinari, and Servien (2013) studied unsupervised clustering of multivariate circular data which consist of the positions of five separate X-ray beams on a circle. Chen, Singh, Guo, Fang, and Liu (2013) improved a new method to identify flood seasonality and partition the flood season into sub-seasons. A study conducted by Costa, Koivunen, and Poor (2014) estimated directional probability distribution of wavefields observed by sensor arrays. Tasdan and Cetin (2014) carried out a simulation study on the influence of ties on uniform scores test for circular data. Hawkins and Lombard (2015) proposed an optimal method for segmentation of cir-

* Corresponding author. Tel.: +905378615941.
E-mail addresses: okesemen@gmail.com (O. Kesemen), ozge_tzl@hotmail.com (Ö. Tezel), eda.ozkul.gs@gmail.com (E. Özkul).

http://dx.doi.org/10.1016/j.eswa.2016.03.034
0957-4174/

cular data generated from von Mises distribution. Kitamura et al. (2015) proposed a new hybrid method that combined directional clustering and advanced nonnegative matrix factorization (NMF). They handled the problems in multichannel music signal separation. da Silva (2015) proposed a directional clustering approach based on mixtures of von Mises-Fisher (vMF) distributions to reduce uncertainty in estimating the orientation of neuronal pathways in diffusion magnetic resonance imaging. Yang, Chang-Chien, and Hung (2016) presented an unsupervised clustering algorithm for directional data on the unit hypersphere without initialization, for which it is not necessary to give the number of clusters a priori.

Clustering analysis is one of the most important issues in the data analysis. Clustering is used to separate a data set into a desired number of clusters. In this separation process, the data points in the same cluster are the most similar to each other and the data points in the different clusters are the most dissimilar.

When considered from the statistical point of view, clustering methods can generally be divided into two categories: the not distribution-free approach and the distribution-free approach. The most-used algorithms from the not distribution-free approaches are the expectation and maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977; McLachlan & Basford, 1988) and the fuzzy c-directions (FCD) algorithm (Yang & Pan, 1997). These algorithms can be applied to directional data. Chang-Chien, Hung, and Yang (2012) adapted the mean shift clustering algorithm, used for numeric data, for circular data by determining automatically the number of clusters. Then, Yang, Chang-Chien, and Kuo (2014) applied the mean shift clustering algorithm for circular data to hyperspherical data.

The most-used algorithms from the distribution-free approaches are partitional clustering methods. K-means (MacQueen, 1967) and fuzzy c-means (FCM) clustering algorithms are the most-common partitional clustering algorithms. These algorithms are applied to linear data. However, in this study, the FCM clustering algorithm is modified to apply to directional data. Different from the existing studies, the proposed method uses angular difference as the similarity measure. In addition, the proposed algorithm is a distribution-free approach.

In Section 2, classical FCM algorithm is explained. In Section 3, general definitions and similarity measures for directional data are described. In Section 3.1 and Section 3.2, the EM and the FCD algorithms for directional data are introduced. In Section 4, the modified FCM algorithm for directional data is given. In Section 5, the EM, the FCD and the FCM4DD algorithms are applied to the some numerical data, and their performances are compared. In Section 5.1, the membership values of the FCD and the FCM4DD are compared.

2. Fuzzy c-means clustering (Fcm) algorithm

The fuzzy c-means clustering (FCM) algorithm was proposed by Dunn in 1973 and improved by Bezdek in 1981 (Höppner, Klawonn, Kruse, & Runkler, 2000). The FCM algorithm, based on objective function, is subject to the principle that each data point belongs to more than one cluster with different membership values, ranging from [0,1]. Additionally, the sum of the membership values for each data point must be one. If a data point is in the cluster center, its membership value is one:

Let $X = \{x_1, x_2, \ldots, x_N\}$ be a sample of N observations in D-dimensional Euclidean space($x_i \in \mathbb{R}^D$). Clustering is process which separates this data set into C subsets and their cluster centers which are $\{v_1, v_2, \ldots v_j, \ldots, v_C\}$. The desired optimal criterion minimizes the objective function while separating the data set into subsets. The algorithm tries to minimize the following objective function which is the generalized form of the least-squared errors function (Höppner et al., 2000):

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} \mu_{ij}^m \left|\left|x_i - v_j\right|\right|^2, \ 1 < m < \infty \tag{1}$$

in which m is the weighting fuzziness parameter and is generally chosen as 2. μ_{ij} is the membership value of the i^{th} data to the j^{th}cluster and μ_{ij} must satisfy the following three conditions (Bezdek, 1981):

1. The membership value ranges between zero and one as given in Eq. (2):

$$\mu_{ij} \in [0, 1], \ \forall\, i, j \tag{2}$$

2. The sum of the membership values for each data point must be one as given in Eq. (3):

$$\sum_{j=1}^{C} \mu_{ij} = 1, \ \forall i \tag{3}$$

3. The sum of the all membership values in a cluster must be smaller than the number of data (N) as given in Eq. (4):

$$0 < \sum_{i=1}^{N} \mu_{ij} < N, \ \forall N \tag{4}$$

The FCM algorithm is a simple method and is the most common clustering algorithm in the all fuzzy clustering methods (Bezdek, Ehrlich, & Full, 1984). The FCM algorithm can be summarized as follows:

FCM Algorithm

Step 1. Fix $C \in [2, N)$, $(m > 0)$ and $(\varepsilon > 0)$.
Step 2. Give initials randomly $\mu_{ij}^{(0)} \sim U(0, 1)$ and let $t = 1$.
Step 3. Compute cluster centers (v_j) by using Eq. (5):

$$v_j = \frac{\sum_{i=1}^{N} \mu_{ij}^m x_i}{\sum_{i=1}^{N} \mu_{ij}^m}, \ (j = 1, 2, \ldots, C) \tag{5}$$

Step 4. Update μ_{ij} with v_j by using Eq. (6):

$$\mu_{ij} = \left(\sum_{k=1}^{C} \left(\frac{\|x_i - v_j\|}{\|x_i - v_k\|} \right)^{\frac{2}{m-1}} \right)^{-1}, \ (i = 1, 2, \ldots, N; \ j = 1, 2, \ldots, C) \tag{6}$$

Step 5. Compute $\|\mu^{(t)} - \mu^{(t-1)}\|$.
IF $\|\mu^{(t)} - \mu^{(t-1)}\| < \varepsilon$, STOP
ELSE $t = t + 1$ and return to Step 3.

3. Clustering methods for directional data

Generally, statistical data analysis is used for data on the linear axis (Kaufman & Rousseeuw, 1990). However, classical statistical methods used for these data cannot be applied inherently to directional data. This is because; directional data have a modular structure. If directional data are defined in the interval $[-\pi, \pi)$, they are continuous between the points (π) and $(-\pi)$; if directional data are defined in the interval $[0, 2\pi)$, they are continuous between the points (2π) and (0). In terms of numerical values, classical methods cannot be used, when the data is discontinuous within these boundaries. The best example of this is demonstrated by the following: distance between the angles 359° and 1° is 2°, but the numerical subtraction of these angles is 358°. Likewise, it might firstly appear that the mean of these angles is 0°, whereas it is 180° in reality.

Clustering algorithms use the distances between data as the similarity measure. There are angularly two distances between two