Information Sciences 279 (2014) 105–122

Contents lists available at ScienceDirect

Information Sciences

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

A new credibilistic clustering algorithm

M. Rostam Niakan Kalhori [a], M.H. Fazel Zarandi [a,b,*], I.B. Turksen [b,c]

[a] Department of Industrial Engineering, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran
[b] Knowledge Intelligent Systems Laboratory, University of Toronto, Toronto, Canada
[c] Department of Industrial Engineering, TOBB University of Economics and Technology, Sogutozu, Ankara, Turkey

ARTICLE INFO

Article history:
Received 13 April 2013
Received in revised form 9 March 2014
Accepted 26 March 2014
Available online 18 April 2014

Keywords:
Credibilistic clustering
Credibility measure
Objective function-based clustering

ABSTRACT

This paper focuses on credibilistic clustering approach. A data clustering method partitions unlabeled data sets into clusters and labels them for various goals such as computer vision and pattern recognition. There are different models for objective function-based fuzzy clustering such as Fuzzy C-Means (FCM), Possibilistic C-Mean (PCM) and their combinations. Credibilistic clustering is a new approach in this field. In this paper, a new credibilistic clustering model is introduced in which credibility measure is applied instead of possibility measure in possibilistic clustering. Also, in objective function, the separation of clusters is considered in addition to the compactness within clusters. The steps of clustering are designed based on this approach. Finally, the main issues about model are discussed, and the results of computational experiments are presented to show the efficiency of the proposed model.

1. Introduction

The aim of cluster analysis is to partition a given set of data or objects into clusters (subsets, groups, classes). This partition should have the following properties [12]:

- Homogeneity within the clusters, i.e., data that belongs to a cluster should be as similar as possible.
- Heterogeneity between clusters, i.e., data that belongs to different clusters should be as different as possible.

The concept of "similarity" has to be specified according to the data. Since data is in most cases real-valued vectors, the Euclidean distance between data can be used as a measure of the dissimilarity. One should consider that the individual variables (components of the vector) cannot be of different relevance. In particular, the range of values should be suitable scaled in order to obtain reasonable distance values [12]. A basic classification of clustering models assigns them into two groups: crisp and fuzzy. One of the most used fuzzy clustering models is FCM [2]. This model imposes sum of one for degree of membership of each data to clusters. In some cases, FCM considered as probabilistic clustering because of this constraint. Although FCM is a very useful clustering method, its memberships do not always correspond well to the degrees of belonging of the data, and it may be inaccurate in a noisy environment [18]. To deal with this weakness of FCM, and to obtain memberships that have a good explanation of the degrees of belonging for the data, in [18] a possibilistic approach to clustering

* Corresponding author at: Department of Industrial Engineering, Amirkabir University of Technology, P.O. Box 15875-3144, Tehran, Iran. Tel.: +98 21 64545378; fax: +98 21 64545378.
E-mail addresses: niakan@aut.ac.ir (M. Rostam Niakan Kalhori), zarandi@aut.ac.ir (M.H. Fazel Zarandi), bturksen@etu.edu.tr (I.B. Turksen).

http://dx.doi.org/10.1016/j.ins.2014.03.106
0020-0255/

has been introduced by Krishnapuram and Keller; it has considered a possibilistic type of membership function to obtain the degree of belonging (Possibilistic C-Means (PCM)). In [18], it was shown that algorithms with possibilistic memberships are more robust to noise and outliers than FCM. The problem of PCM is the coincident clusters, because by relaxing the constraint of FCM, it obtains the degree of belonging of each data to each cluster only by considering that cluster. So, it allows each data to belong to each cluster independently of the other data and the other clusters. The idea of relaxing the constraint of FCM using a measure which has the feature of this constraint intrinsically seems rational. This idea leads us to use credibility measure instead of possibility measure in objective function of PCM. Unlike the possibility measure, credibility measure is self-dual. That is, the credibility of belonging each data to each cluster in addition to credibility of belonging it to the other clusters is equal to 1. So, the credibility of belonging of each data to different clusters contributes to obtain the credibility of belonging it to a special cluster. This approach uses more information about the other clusters to construct a cluster. It should be noted that, in this paper this feature of credibility measure does not consider as a model constraint, because it is the intuitive property of credibility measure. Another contribution of this paper toward the literature is considering the separation of clusters in objective function. So, the contributions of this work can be summarized as follows:

- Utilization of the good characteristics of both PCM and FCM in addition to elimination of the deficiencies of them using the credibility measure in the proposed model.
- Considering the compactness within the clusters and the separation of them in objective function simultaneously.
- Designing different indicators for evaluation of the proposed model.

The rest of the paper is organized as follows: Related works are reviewed in Section 2. Proposed credibilistic clustering model is explained in Section 3. The experimental results by discussing the obtained results are presented in Section 4. Finally, some conclusions are drawn in Section 5.

2. Literature review

Suppose we have an unlabeled data set $X = \{x_1, x_2, \ldots, x_N\} \subseteq \Re^p$. Partitioning this data set to $c(1 < c < N)$ subgroup which assigns each data to one subgroup is known as clustering. Partition matrix U is a matrix with members u_{ik} which are degree of belonging x_k to the ith cluster. There are four basic methods and their combinations to obtain partition matrix U using objective function-based clustering: Crisp partitioning, Fuzzy C-Means, Possibilistic Clustering Method and their combinations, and Credibilistic Clustering. Generally, the objective function-based clustering models can be classified into two classes: crisp and fuzzy. Hard partitioning is crisp model and FCM, PCM, credibilistic clustering, and their combinations are in fuzzy class. In hard partitioning the degree of belonging data x_k to the ith cluster is 0 or 1. In this model each data can be assigned to one and only one cluster. In fuzzy clustering models, the u_{ik} takes value in [0,1].

In Section 3 main groups of objective function-based fuzzy clustering models are reviewed. Fuzzy C-Means (FCM) based models are the first group. The FCM algorithm recognizes spherical clouds of points in a p dimensional space. The clusters are assumed to be of approximately the same size. Each cluster is represented by its center. This representation of a cluster is also called a prototype, since it is often regarded as a representative of all data assigned to the cluster. As a measure for the distance, the Euclidean distance between a datum and a prototype is used. It is only supposed to clarify that the algorithm is intended for a fixed number of clusters, i.e. it does not determine that, number [12]. In FCM and its derivatives a least-squares type of criterion function constructed based on the relation between objects and their distance of cluster centers, is minimized. An important constraint of this model is that the sum of the membership degrees for each object equals 1. It means that each object gets the same weight in comparison to all other objects and, therefore, that all objects are (equally) included into the clusters. The reviewed works related to FCM are as following:

In FCM, u_{ik} is named as membership grade. FCM algorithm assigns memberships to the kth data (x_k) which are indirectly related to the relative distance of x_k to the c cluster centers (prototypes) in the FCM model. Applying the constraints $\sum_{i=1}^{c} u_{ik} = 1$ for each x_k forces the algorithm to assign unreal membership degree to noise data to satisfy this constraint.

There are different algorithms based on FCM. A modified version of fuzzy C-means, called fuzzy C-means with additional data (FCM-AD), was presented in [11] by Hirota and Iwama. The aim of the authors in [11] was achieving robustness against a few outliers. So, a standard pattern vector and a parameter which defined as a ratio of a term in FCM to one in the pattern-matching method was added to FCM objective function. In FCM-AD the standard pattern vector should be given beforehand. Also, another parameter depends on the standard pattern vector. The method is inefficient because these unknown parameters have important role in it.

Conditional FCM was proposed by Pedrycz in [30] in which a conditional variable was considered for each pattern. The structure of these patterns reveals considering their vicinity in feature space conditioned to the value of these assumed variables. In this algorithm, the constraint of FCM has been changed. Pedrycz and Waletzky [31] modified FCM using partial supervision. The main idea of the authors in [31] was exploitation of labeled data in order to cluster data set. The phenomenon of partial supervision occurs when in addition to a vast number of unlabeled patterns, one is also furnished with some (usually, few) labeled patterns. Definitely, these few already classified patterns, when carefully exploited, could provide some general guidance to the clustering mechanism [31]. In [31], the objective function of FCM has been modified in order to calculate the membership value of labeled and unlabeled data in clusters. This proposed clustering method can be used in