Knowledge-Based Systems 74 (2015) 119–132

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Credal c-means clustering method based on belief functions

Zhun-ga Liu [a,b,*], Quan Pan [a], Jean Dezert [c], Grégoire Mercier [b]

[a] School of Automation, Northwestern Polytechnical University, Xi'an, China
[b] Telecom Bretagne, CNRS UMR 6285 Lab-STICC/CID, Brest, France
[c] ONERA – The French Aerospace Lab, F-91761 Palaiseau, France

ARTICLE INFO

Article history:
Received 16 April 2014
Received in revised form 9 November 2014
Accepted 10 November 2014
Available online 20 November 2014

Keywords:
Fuzzy c-means (FCM)
Belief functions
Credal partition
Data clustering
Uncertain data

ABSTRACT

The recent credal partition approach allows the objects to belong to not only the singleton clusters but also the sets of clusters (i.e. meta-clusters) with different masses of belief. A new credal c-means (CCM) clustering method working with credal partition has been proposed in this work to effectively deal with the uncertain and imprecise data. In the clustering problem, one object simultaneously close to several clusters can be difficult to correctly classify, since these close clusters appear not very distinguishable for this object. In such case, the object will be cautiously committed by CCM to a meta-cluster (i.e. the disjunction of these close clusters), which can be considered as a transition cluster among these different close clusters. It can well characterize the imprecision of the class of the object and can also reduce the misclassification errors thanks to the use of meta-cluster. CCM is robust to the noisy data because of the outlier cluster. The clustering centers and the mass of belief on each cluster for any object are obtained by the optimization of a proper objective function in CCM. The effectiveness of CCM has been demonstrated by three experiments using synthetic and real data sets with respect to fuzzy c-means (FCM) and evidential c-means (ECM) clustering methods.

1. Introduction

Fuzzy c-means (FCM) [2] remains so far the most popular data clustering method, and it works with fuzzy partition under the probabilistic framework. In the clustering of imprecise data, some data points (objects) can be simultaneously close to several clusters, and these clusters sometimes become undistinguishable for the objects. So it makes these objects really difficult to correctly classify. The imprecision of information cannot be well captured by the probabilistic framework [17], whereas belief function (BF) theory [34–36], also called evidence theory, is good at modeling the imprecision and uncertainty. Belief functions have been successfully applied to deal with the uncertain and imprecise data in many fields including classification [10,11,14,20,22,28,33,37,38], clustering [13,29–31] and information fusion [18,19,24–26], etc.

A concept of partition named credal partition [12] has been recently proposed by Denœux and Masson for data clustering under the belief function framework. A credal partition can be considered as an extension of the existing concepts of hard [23], fuzzy [15] and possibilistic partition [21], since it allows that the objects belong to not only the singleton clusters in the set of clusters $\Omega = \{w_1, \dots w_c\}$, but also to any subsets of Ω (i.e. meta-clusters) with different masses of belief. This additional flexibility of credal partition is able to gain a deeper insight in the data and to improve robustness with respect to outliers [31]. An EVidential CLUStering (EVCLUS) [13] algorithm working with credal partition has been developed for relational data, and evidential C-Means (ECM) [31] clustering method inspired from FCM [2] and Noise-Clustering (NC) algorithm [7–9] is also proposed for credal partition of object data. The relational version of ECM (RECM) [32] has been derived for dealing with relational data. RECM and EVCLUS are compared in [32], and it is pointed that RECM provides similar results to those given by EVCLUS, but the optimization procedure of RECM is computationally much more efficient than the gradient-based procedure of EVCLUS. The constrained ECM (CECM) [1] method has been also recently introduced to take into account the pairwise constraints.

ECM [31] working with credal partition can produce three kinds of cluster: singleton (specific) clusters (e.g. w_i), meta-clusters (e.g. $w_j \cup \dots \cup w_k$) defined by disjunction (union) of several singleton clusters, and the outlier cluster represented by $\emptyset$. Each cluster (e.g. $w_i, i = 1, \dots, c$) corresponds to one clustering center (prototype) (e.g. $\mathbf{v}_i, i = 1, \dots, c$), and the meta-cluster's center is obtained

* Corresponding author at: School of Automation, Northwestern Polytechnical University, Xi'an, 710072, China.

E-mail addresses: liuzhunga@gmail.com (Z.-g. Liu), jean.dezert@onera.fr (J. Dezert), Gregoire.Mercier@telecom-bretagne.eu (G. Mercier).

http://dx.doi.org/10.1016/j.knosys.2014.11.013
0950-7051/

by arithmetic mean value of the prototype vectors of the singleton clusters included in the meta-cluster. As a result, the different cluster centers may be very close. For example, one singleton cluster's center $\mathbf{v}_i$ can be very close to an incompatible[1] meta-cluster's center $\mathbf{v}_{j,k}$ (corresponding to the meta-cluster $w_j \cup w_k$), once it holds $\mathbf{v}_i \approx (\mathbf{v}_j + \mathbf{v}_k)/2 = \mathbf{v}_{j,k}$. Moreover, the incompatible meta-cluster centers can also be overlapped. The mass of belief on each cluster mainly depends on the distance between the object and the corresponding clustering center taking into account the cardinality.[2] When the different cluster centers are close, ECM will produce very counterintuitive results because some objects belonging to a singleton cluster can be wrongly committed into an incompatible meta-cluster whose center is close to the singleton cluster's center. The main interesting idea of ECM is the introduction of the meta-clusters to FCM, but their use (as proposed in ECM) remains questionable. In our previous related works, we developed a method called belief c-means (BCM) [27] to deal with the close clustering centers by considering the meta-class as sets of the objects far from the specific classes included in the meta-class, but much farther to the other class. This interpretation of meta-clusters, which is different from ECM standpoint, had been proposed in BCM mainly for outliers detection, but it is complicate to use and to implement, which makes BCM not very attractive for the potential users.

In this work, we propose a new evidential version of FCM called credal c-means (CCM) to overcome the limitation of ECM. The use of meta-clusters is presented and well justified. CCM also works with credal partition for the clustering of imprecise data based on belief functions. The singleton cluster in CCM corresponds to the objects very close to the center of this cluster, which is similar to FCM and ECM. CCM is also robust to the noisy data (i.e. outlier) because of the outlier cluster, and it is mainly determined according to a given outlier threshold. In CCM, the meta-cluster is considered as a kind of transition cluster among the different close singleton clusters. Thanks to meta-cluster, credal partition provides an effective tool for the clustering of the imprecise data that are hard to be correctly committed to a particular cluster, and it can also reduce the error occurrences. If one object is considered in a meta-cluster, it must be simultaneously close to the singleton clusters included in the meta-cluster, which means this object is not likely to belong to the other incompatible clusters, and this mainly depends on the distances between the object and these singleton clusters' centers. Meanwhile, these singleton clusters should be undistinguishable for the object, which indicates the objects are hard to correctly classify, and this mainly depends on the distance between this object and the meta-cluster's center (i.e. the mean value of the involved singleton cluster centers). Thus, in the determination of the mass of belief on the meta-cluster, we should take into account not only the distance to the meta-cluster's center but also the distances to the involved singleton clusters' centers.

This new CCM method differs from our previous BCM method because it is based on a distinct underlying principle and a different interpretation of the meta-clusters. To illustrate the difference of CCM with respect to BCM, let us consider two clusters w_1 and w_2 with the corresponding centers $\mathbf{v}_1$ and $\mathbf{v}_2$ as shown in Fig. 1. The object (e.g. $\mathbf{x}_1$) lying in the overlapped zone of w_1 and w_2 will be considered quite uncertain by BCM because it is close to the clustering centers $\mathbf{v}_1$ and $\mathbf{v}_2$ with the similar distance. Finally, $\mathbf{x}_1$ will be committed to a singleton cluster[3] (either w_1, or w_2) according to the maximum mass of belief based on the hard credal partition obtained from BCM. The object (e.g. $\mathbf{x}_2$) being far from both $\mathbf{v}_1$ and $\mathbf{v}_2$ will be committed by BCM to the meta-cluster $w_1 \cup w_2$ because the clustering centers $\mathbf{v}_1$ and $\mathbf{v}_2$ cannot be clearly discerned by $\mathbf{x}_2$. The commitment principle of CCM is different because with CCM the object $\mathbf{x}_2$ will be likely considered as an outlier depending on the chosen outlier threshold,[4] and the object $\mathbf{x}_1$ will be committed to the meta-cluster $w_1 \cup w_2$ with the biggest mass of belief because it is both close to the center of meta-cluster $\mathbf{v}_{1,2}$, and close to the centers $\mathbf{v}_1$ and $\mathbf{v}_2$.

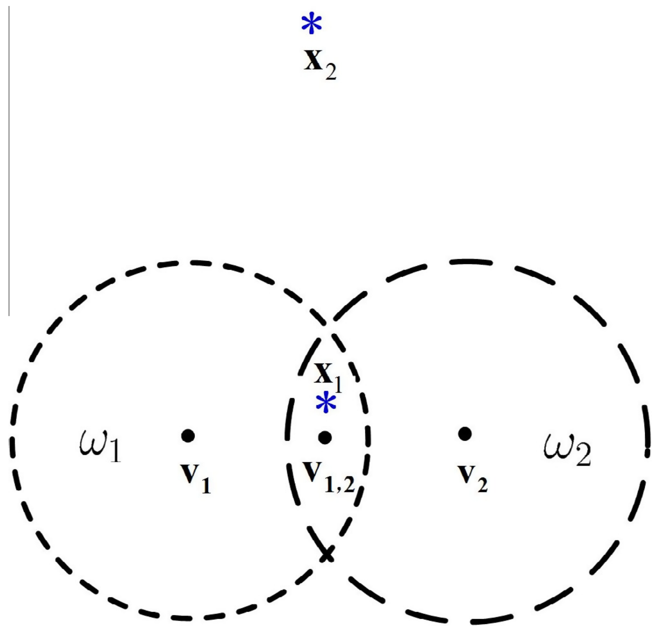

Fig. 1. Illustration of the difference between BCM and CCM.

In CCM, if one object is in a meta-cluster, it indicates that this object belongs to one of the singleton clusters included in the meta-cluster, but the available information used for making the classification is not sufficient enough to obtain a clear (specific) class of the object. CCM can well reveal the imprecision degree of the object belonging to different classes and can also reduce the misclassification errors due to the meta-cluster. This is advantageous for many applications, specially those related to defense and security, like in target classification and tracking. In these applications, it is better to get a robust (partly imprecise) result that will need to be precisiated with additional techniques, than to obtain directly with high risk a wrong precise classification from which an erroneous fatal decision would be drawn. The output of CCM is not always used to provide a final decision about classification of an object. In fact, it can be seen as an interesting source of information to be combined with some other complementary information sources in order to get more precise clustering results if necessary.

The objective function of CCM is defined following this basic principle. The clustering centers and the mass of belief on each cluster for the objects can be obtained by the optimization (minimization) of this objective function. For a c-class data set, the credal partition produces 2^c clusters, and its computation complexity is very high when c is big. In real applications, the classification of the imprecise object is usually unspecific among a very small number (e.g. two or three) of singleton clusters, and there are very few objects belonging to the meta-clusters having big cardinalities. So a

[1] Clusters A and B are said compatible if $A \cap B \neq \emptyset$, and they are said incompatible if $A \cap B = \emptyset$.

[2] The cardinality of a cluster A, denoted by $|A|$, is the number of the singleton clusters included in A. For example, if $A = \{w_i\}$, then $|A| = 1$. If $A = \{w_i, w_j\}$ then $|A| = 2$. Using Shafer's notations [34], the set $A = \{w_i, w_j\}$ is usually written as $A = w_i \cup w_j$ in the BF framework, and we also use this notation in the sequel of this paper.

[3] because the center of meta-cluster is not involved in BCM approach.

[4] as similarly done in ECM.