Knowledge-Based Systems 000 (2017) 1–16

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

k-CEVCLUS: Constrained evidential clustering of large dissimilarity data☆

Feng Li[a], Shoumei Li[a], Thierry Denœux[a,b,*]

[a] College of Applied Sciences, Beijing University of Technology, Beijing, China
[b] CNRS, Sorbonne Universités, Université de Technologie de Compiègne, Heudiasyc (UMR 7253), France

ARTICLE INFO

Article history:
Received 2 August 2017
Revised 21 November 2017
Accepted 22 November 2017
Available online xxx

Keywords:
Evidence theory
Dempster–Shafer theory
Belief functions
Relational data
Credal partition
Constrained clustering
Instance-level constraints

ABSTRACT

In evidential clustering, cluster-membership uncertainty is represented by Dempster–Shafer mass functions. The EVCLUS algorithm is an evidential clustering procedure for dissimilarity data, based on the assumption that similar objects should be assigned mass functions with low degree of conflict. CEVCLUS is a version of EVCLUS allowing one to use prior information on cluster membership, in the form of pairwise must-link and cannot-link constraints. The original CEVCLUS algorithm was shown to have very good performances, but it was quite slow and limited to small datasets. In this paper, we introduce a much faster and efficient version of CEVCLUS, called *k*-CEVCLUS, which is both several orders of magnitude faster than EVCLUS and has storage and computational complexity linear in the number of objects, making it applicable to large datasets (around 10^4 objects). We also propose a new constraint expansion strategy, yielding drastic improvements in clustering results when only a few constraints are given.

1. Introduction

Cluster analysis, also called data segmentation, is one of the basic tasks in data mining and machine learning. The goal of cluster analysis is to segment a collection of objects into clusters in such a way that similar objects belong to the same cluster, while dissimilar ones are assigned to different clusters. Typically, two data types are considered: *attribute* and *dissimilarity* data. Dissimilarity data, also known as *relational* data or *proximity* data, are composed of distances, or dissimilarities between objects. Attribute data can always be transformed into dissimilarity data by using a suitable metric. In this paper, we focus mostly on dissimilarity data.

Several approaches to clustering have been developed over the years. In hard clustering, each object is assigned with full certainty to one and only one cluster; the *c*-means algorithm is the reference method in this category. In contrast, "soft" clustering algorithms [22] are based on different ways of representing cluster-membership uncertainty. These include fuzzy [3], possibilistic [14] and rough [17] clustering. *Evidential clustering* [8,9,18,20] is a recent approach to soft clustering, in which uncertainty is represented by Dempster–Shafer mass functions [25]. The resulting clustering structure is called *credal partition*. Thanks to the generality of Dempster–Shafer theory, evidential clustering can be shown to extend all other soft clustering paradigms [7]. Some of the recent advances in evidential clustering are briefly summarized here. The Evidential *c*-Means (ECM) algorithm [20] is an extension of the hard and fuzzy *c*Means, in which prototypes are defined not only for clusters, but also for sets of clusters. A cost function is minimized in turn with respect to the prototypes, and with respect to the credal partition. A version of ECM for dissimilarity data, called RECM, was proposed in [21]. In [18], another variant of the ECM algorithm (called CCM) was proposed, based on an alternative definition of the distance between a vector and the prototype of a meta-cluster. This modification produces more sensible results in situations where the prototype of a meta-cluster is close to that of singleton cluster. In [27], Zhou et al. introduced yet another variant of ECM, called Median Evidential *c*-means (MECM), which is an evidential counterpart to the median *c*-means and median fuzzy *c*-means algorithms. An advantage of this approach is that it does not require the dissimilarities between objects to verify the axioms of distances. Denœux et al. [8] proposed another evidential clustering method, called E*k*-NNclus, which is based on evidential *k*-nearest neighbor rule [5]. Evidential clustering has been successfully applied in various fields, including machine prognosis [24], medical image processing [15,16,19] and analysis of social networks [27].

☆ This research was supported by grant No.11571024 from NSFC, and by the Overseas Talent program from the Beijing Government.

* Corresponding author at : CNRS, Sorbonne Universités, Université de Technologie de Compiègne, Heudiasyc (UMR 7253), France.
E-mail address: tdenoeux@utc.fr (T. Denœux).

https://doi.org/10.1016/j.knosys.2017.11.023
0950-7051/

The notion of credal partition was first introduced in [9], together with the first evidential clustering algorithm, called EVCLUS. The EVCLUS algorithm is similar in spirit to multidimensional scaling procedures [4]. It attempts to build a credal partition such that the plausibility of two objects belonging to the same cluster is higher when the two objects are more similar. This result is achieved by minimizing a stress, or cost function using a gradient-based optimization procedure. A constrained version of EVCLUS allowing for the utilization of prior knowledge about the joint cluster membership of object pairs was later proposed in [1] under the name CEVCLUS. In the CEVCLUS method, pairwise constraints are formalized in the belief function framework and translated as a penalty term added to the stress function of EVCLUS.

Both EVCLUS and CEVCLUS were shown to outperform state-of-the-art clustering procedures [1,9]. However, their high space and time complexity restricted their application to small datasets with only a few hundred objects. Recently, a new version of EVCLUS, called *k*-EVCLUS, has been proposed [10]. *k*-EVCLUS is based on an iterative row-wise quadratic programming (IRQP) algorithm, which makes it much faster than EVCLUS. It also uses only a random sample of the dissimilarities, which reduces the time and space complexity from quadratic to linear, making it suitable to cluster large datasets.

In this paper, we carry out similar improvements to the CEVCLUS algorithm. We show that the cost function composed of a stress term and a penalty term encoding pairwise constraints can also be minimized using the IRQP algorithm, which is several orders of magnitude faster than the gradient-based procedure used in [1]. Together with dissimilarity sampling, this modification makes the new version of CEVCLUS (called *k*-CEVCLUS) applicable to large datasets composed of tens of thousands of objects with pairwise constraints. We also introduce a new constraint expansion strategy, which brings considerable improvements in clustering results when only a few constraints are provided. Altogether, the contributions reported in this paper considerably extend the applicability of constrained evidential clustering to real-world datasets of realistic size.

The rest of this paper is organized as follows. Basic notions on belief functions and credal partitions, as well as the *k*-EVCLUS and CEVCLUS algorithms are first recalled in Section 2. The new *k*-CEVCLUS algorithm and the constraint expansion procedure are then described in Section 3.1, and experimental results are reported in Section 4. Finally, Section 5 concludes the paper.

2. Background

The purpose of this section is to provide the reader with background information so as to make the paper self-contained. Basic notions of Dempster–Shafer theory are first recalled in Section 2.1, and the concept of credal partition is introduced in Section 2.2. The *k*-EVCLUS and CEVCLUS algorithms are then presented in Sections 2.3 and 2.4, respectively.

2.1. Mass functions

Let $\Omega = \{\omega_1, \ldots, \omega_c\}$ be a finite set. A *mass function* on Ω is a mapping from the power set 2^Ω to [0, 1], satisfying the condition

$$\sum_{A \subseteq \Omega} m(A) = 1. \tag{1}$$

Each subset A of Ω such that $m(A) > 0$ is called a *focal set*. In Dempster–Shafer theory, a mass function encodes a piece of evidence about some question of interest, for which the true answer is assumed to be an element of Ω. For any nonempty focal set A, $m(A)$ is a measure of the belief that is committed exactly to A [25]. The mass $m(\emptyset)$ assigned to the empty set has a special interpretation: it is a measure of the belief that the true answer might not belong to Ω. As we will see, this quantity is very useful in clustering to identify outliers. A mass function is said to be

- *Bayesian* if all its focal sets are singletons;
- *Logical* if it has only one focal set;
- *Certain* if it is both logical and Bayesian;
- *Consonant* if its focal sets are nested.

Given a mass function m, the corresponding *belief* and *plausibility* functions are defined, respectively, as

$$Bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B)$$

and

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B),$$

for all $A \subseteq \Omega$. The quantity $Bel(A)$ represents the degree of total support in A, while $Pl(A)$ can be interpreted as the degree to which the evidence is consistent with A.

The *degree of conflict* [25] between these two mass functions m_1 and m_2 defined on the same frame Ω is

$$\kappa = \sum_{A \cap B = \emptyset} m_1(A) m_2(B). \tag{2}$$

If m_1 and m_2 are mass functions representing evidence about two distinct questions with the same set of possible answers Ω, then the plausibility that the two questions have the same answer is equal to $1 - \kappa$ [9].

2.2. Credal partition

Let $\mathcal{O} = \{o_1, \ldots, o_n\}$ be a set of n objets. We assume that each object belongs to at most one of c clusters. The set of clusters is denoted by $\Omega = \{\omega_1, \ldots, \omega_c\}$. In evidential clustering, the uncertainty about the cluster membership of each object o_i is represented by a mass function m_i on Ω. The n-tuple $\mathcal{M} = (m_1, \ldots, m_n)$ is called a *credal partition*. The notion of credal partition is very general and it encompasses most other types of soft clustering structures [7]. In particular,

- If all mass functions m_i are certain, then we have a hard partition, where object o_i is assigned to cluster ω_k if $m_i(\{\omega_k\}) = 1$.
- If all mass functions m_i are Bayesian, then the evidential partition is equivalent to a fuzzy partition; the degree of membership of object o_i to cluster ω_k is then $u_{ik} = m_i(\{\omega_k\})$, for $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, c\}$.
- If all mass functions m_i are logical with a single focal set $A_i \subseteq \Omega$, then we get a rough partition. The lower and upper approximations of cluster k can be defined, respectively, as $\underline{\omega}_k = \{o_i \in \mathcal{O} | A_i = \{\omega_k\}\}$ and $\overline{\omega}_k = \{o_i \in \mathcal{O} | \omega_k \in A_i\}$.
- If each m_i is consonant, then it is equivalent to a possibility distribution, and it can be uniquely represented by the plausibility of the singletons $pl_{ik} = Pl_i(\{\omega_k\})$ for $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, c\}$. Each number pl_{ik} is the plausibility that object i belongs to cluster k; these numbers form a possibilistic partition of the n objects.

Because a credal partition is more general than other types of hard or soft partitions, it can be converted into any other type [7]. For instance, we obtain a fuzzy partition by defining the degree of membership u_{ik} of object o_i to cluster ω_k as

$$u_{ik} = \frac{pl_{ik}}{\sum_{\ell=1}^{c} pl_{i\ell}}. \tag{3}$$

This fuzzy partition can then be converted to a hard partition by assigning each object to the cluster with the highest membership degree.

Please cite this article as: F. Li et al., *k*-CEVCLUS: Constrained evidential clustering of large dissimilarity data, Knowledge-Based Systems (2017), https://doi.org/10.1016/j.knosys.2017.11.023