Knowledge-Based Systems 106 (2016) 179–195

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Evidential clustering of large dissimilarity data

Thierry Denœux[a,1,*], Songsak Sriboonchitta[b], Orakanya Kanjanatarakul[c]

[a] Sorbonne Universités, Université de Technologie de Compiègne, CNRS, UMR 7253 Heudiasyc, France
[b] Faculty of Economics, Chiang Mai University, Thailand
[c] Faculty of Management Sciences, Chiang Mai Rajabhat University, Thailand

ARTICLE INFO

Article history:
Received 9 April 2016
Revised 20 May 2016
Accepted 22 May 2016
Available online 27 May 2016

Keywords:
Dempster-Shafer theory
Evidence theory
Belief functions
Unsupervised learning
Credal partition
Relational data
Proximity data
Pairwise data

ABSTRACT

In evidential clustering, the membership of objects to clusters is considered to be uncertain and is represented by Dempster-Shafer mass functions, forming a credal partition. The EVCLUS algorithm constructs a credal partition in such a way that larger dissimilarities between objects correspond to higher degrees of conflict between the associated mass functions. In this paper, we present several improvements to EVCLUS, making it applicable to very large dissimilarity data. First, the gradient-based optimization procedure in the original EVCLUS algorithm is replaced by a much faster iterative row-wise quadratic programming method. Secondly, we show that EVCLUS can be provided with only a random sample of the dissimilarities, reducing the time and space complexity from quadratic to roughly linear. Finally, we introduce a two-step approach to construct credal partitions assigning masses to selected pairs of clusters, making the algorithm outputs more informative than those of the original EVCLUS, while remaining manageable for large numbers of clusters.

1. Introduction

Clustering data into groups is one of the fundamental tasks in data mining and machine learning. Clustering algorithms can be distinguished according to the input data they can process, and according to the outputs they produce.

Typically, two categories of input data are considered: attribute (vectorial) data and dissimilarity (proximity, relational, pairwise) data. In the former case, each object is described by a vector of numerical or categorical attributes. In the latter, the data takes the form of a matrix of dissimilarities between objects. Attribute data can be easily transformed into dissimilarity data by choosing a suitable distance. The inverse transformation (from dissimilarity to attribute data) is generally more difficult, except in the special case of metric dissimilarities, i.e., dissimilarities that are exact Euclidean distances between vectors in a latent space, a case not so frequent in practice. Finding an attribute representation of a set of objects, such that distances between objects approximate a given dissimilarity matrix is often a difficult task (referred to as multidimensional scaling – MDS), which requires to solve a large scale nonlinear optimization problem [3,4]. Most clustering algorithms, such as the *c*-means algorithms and its numerous variants, are designed to handle attribute data. A smaller number of algorithms, referred to as *relational clustering* methods, can directly handle dissimilarity data [9–11].

As for the clustering outputs, we can distinguish between partitional clustering, which aims at finding a partition of the objects, and hierarchical clustering, which finds a sequence of nested partitions. Over the years, the notion of partitional clustering has been extended to several important variants, including fuzzy [2] and possibilistic [16] clustering, and more recently, rough [20,27] and evidential [7,25] clustering. Contrary to classical (hard) partitional clustering, in which each object is assigned unambiguously and with full certainty to a single cluster, these variants allow for ambiguity, uncertainty or doubt in the assignment of objects to clusters. For this reason, they are referred to as "soft" clustering methods [28], in contrast with classical, "hard" clustering. Among soft clustering paradigms, *evidential clustering* describes the uncertainty in the membership of each object to clusters using a Dempster-Shafer mass function [30], which assigns a mass to each subset of clusters. This is a rich and informative description of the clustering structure of a data set, which can be shown to include hard, fuzzy and rough partitions as special cases. Recently, evidential clustering has been successfully applied in various domains such as machine prognosis [29], medical image processing [17,24] and analysis of social networks [34]. Similar ideas have also been exploited in supervised classification (see, e.g., [18,21,22]).

* Corresponding author.
E-mail address: tdenoeux@utc.fr, tdenoeux@hds.utc.fr (T. Denœux).
[1] Fax: +33 344234477.

http://dx.doi.org/10.1016/j.knosys.2016.05.043
0950-7051/

In [7], one of us (the first author) introduced EVCLUS, an evidential clustering algorithm that handles (non necessarily metric) dissimilarity data. EVCLUS is based on the natural assumption that the plausibility of two objects belonging to the same cluster is higher when the two objects are more similar. This assumption translates into the search for a credal partition minimizing a cost function. A variant of EVCLUS allowing one to use prior knowledge in the form of pairwise constraints was later introduced in [1].

The EVCLUS algorithm has several advantages. It is conceptually simple and it can handle non metric dissimilarity data (even expressed on an ordinal scale). It was also shown to outperform some of the state-of-the-art relational clustering techniques on a number of datasets [7]. On the minus side, the main drawback of EVCLUS is its computational complexity. As other relational clustering algorithms, it requires to store the whole dissimilarity matrix; the space complexity is thus $O(n^2)$, where n is the number of objects, which precludes application to datasets containing more than a few thousand objects. Furthermore, each iteration of the gradient-based optimization procedure implemented in the EVCLUS algorithm requires $O(f^3n^2)$ operations, where f is the number of focal sets of the mass functions, i.e., the number of subsets of clusters being considered. In the worst case, $f = 2^c$, where c is the number of clusters. To make the method usable even for moderate values of c, we need to restrict the form of the mass functions so that masses are only assigned to focal sets of size 0, 1 or c, which prevents us from fully exploiting the potential generality of the method.

In this paper, we propose some improvements to the EVCLUS algorithm, making it applicable to very large datasets. These improvements are threefold. First, the gradient-based optimization procedure in the original EVCLUS algorithm is replaced by an adaptation of the much faster iterative row-wise quadratic programming method proposed in [31]. Secondly, we show that EVCLUS does not need to be provided with the whole dissimilarity matrix, reducing the time and space complexity from quadratic to roughly linear. Finally, we introduce a two-step approach to construct credal partitions assigning masses to selected pairs of clusters, making the algorithm outputs more informative than those of the original EVCLUS, while remaining manageable for large numbers of clusters.

The rest of the paper is organized as follows. The background on belief functions, evidential clustering and the EVCUS algorithm will first be recalled in Section 2. The new optimization procedure will be described and evaluated in Section 3. Improvements of EVCLUS making it applicable to problems with large numbers of objects and large numbers of clusters will then be described, respectively, in Sections 4 and 5. Finally, Section 6 will conclude the paper.

2. Background

In this section, a brief reminder on Dempster-Shafer theory will first be provided in Section 2.1. Credal partitions and related necessary notions will then be recalled in Section 2.2, and the EVCLUS algorithm will be presented in Section 2.3.

2.1. Mass functions

Let $\Omega = \{\omega_1, \ldots, \omega_c\}$ be a finite set representing the possible answers to some question Q, one and only one of which is true. The true answer is denoted by $\boldsymbol{\omega}$. A *mass function* Ω is a mapping from the power set 2^Ω to [0, 1] such that

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

Each number $m(A)$ represents the degree of support attached to the proposition $\boldsymbol{\omega} \in A$, and to no more specific proposition [30]. The subsets A of Ω such that $m_i(A) > 0$ are called the *focal sets* of m. A mass function m is said to be

- *normalized* if $\emptyset$ is not a focal set;
- *logical* if it has only one focal set;
- *Bayesian* if its focal sets are singletons;
- *certain* if it is both logical and Bayesian, i.e., if it has only one focal set, and this focal set is a singleton;
- *consonant* if its focal sets are nested.

To each mass function m, we may associate belief and plausibility functions from 2^Ω to [0, 1] defined, respectively, as follows,

$$Bel(A) = \sum_{\emptyset\neq B\subseteq A} m(B) \tag{2a}$$

$$Pl(A) = \sum_{B\cap A\neq\emptyset} m(B), \tag{2b}$$

for all $A \subseteq \Omega$. These two functions are linked by the relation $Pl(A) = Bel(\Omega) - Bel(\bar{A})$, for all $A \subseteq \Omega$. The quantity $Bel(A)$ is a measure of how much the proposition "$\boldsymbol{\omega} \in A$" is supported by the available evidence. In contrast $Bel(\Omega) - Pl(A) = Bel(\bar{A})$ is a measure of how much the complementary hypothesis $\bar{A}$ is supported, so that $Pl(A)$ can be seen as a measure of lack of support for $\bar{A}$. The function pl: $\Omega \to [0, 1]$ that maps each element ω of Ω to its plausibility $pl(\omega) = Pl(\{\omega\})$ is called the *contour function* associated to m.

If m is Bayesian, then $Bel = Pl$, and this function is a probability measure; the contour function is thus the usual probability mass function, i.e., $Bel(A) = Pl(A) = \sum_{\omega\in A} pl(\omega)$ for all $A \subseteq \Omega$. If m is consonant, then Pl is a possibility measure, i.e., we have $Pl(A \cup B) = \max(Pl(A), Pl(B))$ for all $A, B \subseteq \Omega$, and Bel is the dual necessity measure; pl is then the corresponding possibility distribution, i.e., $Pl(A) = \max_{\omega\in A} pl(\omega)$ for all $A \subseteq \Omega$. A consonant mass function can be uniquely recovered from its contour function.

Let m_1 and m_2 be two mass functions defined on the same set Ω. Their *degree of conflict* [30] is defined as

$$\kappa = \sum_{A\cap B=\emptyset} m_1(A)m_2(B). \tag{3}$$

It is comprised between 0 and 1. When m_1 and m_2 are two mass functions representing two independent pieces of evidence about the same question, κ is interpreted as a measure of conflict between these two pieces of evidence. A different interpretation of κ was provided in [7], for the case where m_1 and m_2 represent independent pieces of evidence about two different questions Q_1 and Q_2, with the same set of possible answers Ω: in that case, $1 - \kappa$ is the plausibility that the true answers to Q_1 and Q_2 are identical.

Example 1. Let us assume that the questions of interest concern the nationalities of Ann and Henri. Let $\Omega = \{\text{Singapore}, \text{Thailand}, \text{France}, \text{Canada}\}$ be the set of possible answers to both questions. We receive some evidence that Ann comes from an Asian country, with probability 0.8, and independent evidence that Henri originates from a country where French is an official language, with probability 0.5. What is the plausibility that Ann and Henri have the same nationality? The two pieces of evidence translate into the following mass functions

$$m_1(\{\text{Singapore}, \text{Thailand}\}) = 0.8, \quad m_1(\Omega) = 0.2, \tag{4a}$$

$$m_2(\{\text{France}, \text{Canada}\}) = 0.5, \quad m_2(\Omega) = 0.5. \tag{4b}$$

The degree of conflict between m_1 and m_2 is

$$\kappa = m_1(\{\text{Singapore}, \text{Thailand}\})m_2(\{\text{France}, \text{Canada}\}) \tag{5a}$$

$$= 0.8 \times 0.5 = 0.4. \tag{5b}$$

The requested plausibility is thus $1 - 0.4 = 0.6$. □