EVCLUS: Evidential Clustering of Proximity Data

Thierry Denœux and Marie-Hélène Masson

UMR CNRS 6599 Heudiasyc
Université de Technologie de Compiègne
BP 20529 - F-60205 Compiègne cedex - France

Abstract

A new relational clustering method is introduced, based on the Dempster-Shafer theory of belief functions (or Evidence theory). Given a matrix of dissimilarities between n objects, this method (called EVCLUS) assigns a basic belief assignment (or mass function) to each object, in such a way that the degree of conflict between the masses given to any two objects, reflects their dissimilarity. A notion of credal partition is introduced, which subsumes those of hard, fuzzy and possibilistic partitions, allowing to gain deeper insight into the structure of the data. Experiments with several sets of real data demonstrate the good performances of the proposed method as compared to several state-of-the-art relational clustering techniques.
Keywords: Relational Data, Clustering, Unsupervised Learning, Dempster-Shafer theory, Evidence theory, Belief Functions, Multi-dimensional Scaling.

I. Introduction

Cluster analysis is concerned with the formal study of algorithms and methods for finding groups in data, groups (or classes) being defined as subsets of more or less "similar" objects [20]. Such a grouping may be expressed, e.g., as a *partition* (i.e., a set of subsets of objects, such that each object belongs to one and only one subset), a *hierarchy*, defined as a sequence of nested partitions, or a *fuzzy partition*, in which each object has a grade of membership to each group [2]. Another important distinction lies in the nature of the available data. The two most frequent data types are *object data*, in which each object is described explicitly by a list of attributes, and *proximity* (or *relational*) data, in which only pairwise similarities, or dissimilarities are given. Object data can always be transformed into proximity data using a suitable distance, dissimilarity or similarity function. In particular, this is a useful strategy in the case of mixed data types [21]. Consequently, clustering methods which can handle proximity data are, in some way, more general than methods applicable to object data only.

A quite extensive review of crisp and fuzzy relational clustering models can by found in [2, chapter 3]. These methods can be classified into three broad categories: hierarchical methods, methods based on the decomposition of fuzzy relations, and methods based on the optimization of an objective function. Given n objects to be classified in c classes, methods in the latter category aim at finding a fuzzy partition matrix, i.e. a matrix $U = (u_{ik})$ of size $n \times c$ such that

$$\sum_{k=1}^{c} u_{ik} = 1 \quad \forall\, i \in \{1, \ldots, n\} \tag{1}$$

and

$$\sum_{i=1}^{n} u_{ik} > 0 \quad \forall\, k \in \{1, \ldots, c\}\ . \tag{2}$$

Corresponding author. Email: Thierry.Denoeux@hds.utc.fr. Fax: (33) 3 44 23 44 77.

Each number $u_{ik} \in [0,1]$ is interpreted as a *degree of membership* of object i to cluster k. The "best" fuzzy partition matrix is found by optimizing a suitable objective function, using a grouped coordinate descent scheme. Methods in this category include Roubens' fuzzy non metric (FNM) model [28], the assignment-prototype (AP) model [38] and the relational fuzzy c-means (RFCM) model [16] (a similar approach may be found in [21]). The latter approach was later extended by Hathaway and Bezdek [15] to cope with non-Euclidean dissimilarity data, leading to the non-Euclidean relational fuzzy c-means (NERFCM) model. Finally, robust versions of the FNM and RFCM algorithms were proposed by Davé [5].

In this paper, a novel approach to clustering proximity data is presented, based on the Dempster-Shafer (DS) theory of belief functions, also referred to as "Evidence theory". In this approach, called EVCLUS (Evidential Clustering), the allocation of objects to classes is performed using the concept of basic belief assignment (bba), whereby a "mass of belief" is assigned to each possible subset of classes. Having assigned a bba to each object, it is possible to compute, for each two objects, the plausibility that they belong to the same class. It is then required that these plausibilities be, in some sense, compatible with the observed pairwise dissimilarities between objects.

Whereas evidence theory has been applied to supervised classification problems for a long time (see, e.g., [1], [39], [22], [7], [8], [9]), this is, to our knowledge, the first incursion of belief functions into the cluster analysis domain. In the same way as the concept of fuzzy partition subsumes that of crisp partition, resulting in greater expressive power of fuzzy clustering procedures as compared to hard ones, the concept of *credal partition* introduced in this paper is even more general, which allows in some cases (as will be shown) to gain deeper insight into the structure of the data. Additionally, evidential clustering provides the possibility to *combine* in a meaningful way credal partitions obtained from dissimilarity matrices provided, e.g., by several experts, or computed from different sets of measurements.

The rest of this paper is organized as follows. The necessary background on belief function will be recalled in Section 2. Our method will then be exposed in Section 3, and experimental results with three data sets are presented in Section 4. The problem of combining several credal partition is addressed in Section 5, and a first approach to selecting the number of clusters is proposed in Section 6. Section 7 concludes the paper.

II. Evidence theory

Let us consider a variable x taking values in a finite and unordered set Ω called the frame of discernment. Partial knowledge regarding the actual value taken by x can be represented by a *basic belief assignment* (bba) [32], [37], defined as a function m from 2^Ω to $[0,1]$, verifying:

$$\sum_{A \subseteq \Omega} m(A) = 1. \tag{3}$$

The subsets A of Ω such that $m(A) > 0$ are the *focal sets* of m. Each focal set A is a set of possible values for x, and the number $m(A)$ can be interpreted as a fraction of a unit mass of belief, which is allocated to A on the basis of a given evidential corpus. Complete ignorance corresponds to $m(\Omega) = 1$, and perfect knowledge of the value of x is represented by the allocation of the whole mass of belief to a unique singleton of Ω (m is then called a *certain* bba). Another particular case is that where all focal sets of m are singletons: m is then equivalent to a probability function, and is called a *Bayesian* bba.

A bba m such that $m(\emptyset) = 0$ is said to be normal. This condition was originally imposed by Shafer [32], but it may be relaxed if one accepts the *open-world assumption* stating that the set Ω might not be complete, and x might take its value outside Ω [34]. The quantity $m(\emptyset)$ is then interpreted as a mass of belief given to the hypothesis that x might not lie

in Ω.

A bba m can be equivalently represented by any of two non additive fuzzy measures: a belief function (BF) $\text{bel} : 2^\Omega \mapsto [0,1]$, defined as

$$\text{bel}(A) \triangleq \sum_{\emptyset \neq B \subseteq A} m(B) \quad \forall A \subseteq \Omega \ , \tag{4}$$

and a plausibility function $\text{pl} : 2^\Omega \mapsto [0,1]$, defined as

$$\text{pl}(A) \triangleq \text{bel}(\Omega) - \text{bel}(\overline{A}) \quad \forall A \subseteq \Omega \ , \tag{5}$$

where $\overline{A}$ denotes the complement of A. Function bel may be shown to have the property of complete monotonicity [32]. Its use for representing degrees of belief was justified by Smets on an axiomatic basis [36]. Whereas $\text{bel}(A)$ represents the amount of support given to A, the *potential* amount of support that *could be* given to A is measured by $\text{pl}(A)$. Note that both bel and pl boil down to a unique probability measure when m is a Bayesian bba.

Let us now assume that we have two bba's m_1 and m_2 representing distinct items of evidence concerning the value of x. The standard way of combining them is through the conjunctive sum operation $\textcircled{\cap}$ defined as:

$$(m_1 \textcircled{\cap} m_2)(A) \triangleq \sum_{B \cap C = A} m_1(B) m_2(C) \ , \tag{6}$$

for all $A \subseteq \Omega$. The quantity

$$K \triangleq (m_1 \textcircled{\cap} m_2)(\emptyset) = \sum_{B \cap C = \emptyset} m_1(B) m_2(C) \tag{7}$$

is called the *degree of conflict* between m_1 and m_2. It may be seen as a degree of disagreement between the two information sources. If necessary, the normality condition $m(\emptyset) = 0$ may be recovered by dividing each mass $(m_1 \textcircled{\cap} m_2)(A)$ by $1 - K$ (this operation is called Dempster's normalization). The resulting operation is then noted $\oplus$ and is referred to as Dempster's rule of combination [32]:

$$(m_1 \oplus m_2)(A) \triangleq \frac{1}{1-K} \sum_{B \cap C = A} m_1(B) m_2(C) \ . \tag{8}$$

Both $\textcircled{\cap}$ and $\oplus$ are commutative and associative, two desirable properties for combination operators. Note, however, that the use of these rules requires the two sources of information to be distinct and fully reliable, a condition rarely met in practice. A more cautious combination rule is the disjunctive sum $\textcircled{\cup}$ defined as

$$(m_1 \textcircled{\cup} m_2)(A) \triangleq \sum_{B \cup C = A} m_1(B) m_2(C) \ , \tag{9}$$

which is justified, in particular, when it is only known that at least one of the two sources is reliable [35].

Consider now a bba m^Ω defined on the Cartesian product $\Omega = \Omega_1 \times \Omega_2$ (from now on, the domain of a bba will be indicated as superscript when necessary). The marginal bba $m^{\Omega \downarrow \Omega_1}$ on Ω_1 is defined, for all $A \subseteq \Omega_1$, as

$$m^{\Omega \downarrow \Omega_1}(A) \triangleq \sum_{\{B \subseteq \Omega \ | \ \text{Proj}(B \downarrow \Omega_1) = A\}} m^\Omega(B) \ , \tag{10}$$

where $\text{Proj}(B \downarrow \Omega_1)$ denotes the projection of B onto Ω_1, defined as

$$\text{Proj}(B \downarrow \Omega_1) \triangleq \{\omega_1 \in \Omega_1 \mid \exists \omega_2 \in \Omega_2, (\omega_1, \omega_2) \in B\} \ . \tag{11}$$

Whereas marginalization goes from a larger frame to a smaller one, extension goes from a smaller frame to a larger one. Given a bba m^{Ω_1} on Ω_1, its *vacuous extension* [32], [35] on $\Omega = \Omega_1 \times \Omega_2$ is defined for all $B \subseteq \Omega$ as:

$$m^{\Omega_1 \uparrow \Omega}(B) \triangleq \begin{cases} m^{\Omega_1}(A) & \text{if } B = A \times \Omega_2 \text{ for some } A \subseteq \Omega_1 \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

This definition of the vacuous extension results from the Principle of Minimal Commitment [35], which formalizes the idea that one should never give more support than justified to any proposition.

Given to bba's m^{Ω_1} and m^{Ω_2}, their conjunctive sum on $\Omega = \Omega_1 \times \Omega_2$ may be obtained by combining their vacuous extensions on Ω, using (6). We thus obtain:

$$(m^{\Omega_1} \textcircled{\cap} m^{\Omega_2})(A \times B) = m^{\Omega_1}(A) m^{\Omega_2}(B) \ , \tag{13}$$

for all non empty subsets $A \subseteq \Omega_1$ and $B \subseteq \Omega_2$.

III. The method

A. Credal partition of a set of n objects

Let us consider a collection $O = \{o_1, \ldots, o_n\}$ of n objects, and a set $\Omega = \{\omega_1, \ldots, \omega_c\}$ of c classes forming a partition of O. Let us assume that we have only partial knowledge concerning the class membership of each object o_i, and that this knowledge is represented by a bba m_i on the set Ω. We recall that $m_i(\Omega) = 1$ stands for complete ignorance of the class of object i, whereas $m_i(\{\omega_k\}) = 1$ corresponds to full certainty that object i belongs to class k. All other situations correspond to partial knowledge of the class of o_i. For instance, the following bba:

$$\begin{aligned} m_i(\{\omega_k, \omega_\ell\}) &= 0.7 \\ m_i(\Omega) &= 0.3 \end{aligned}$$

means that we have some belief that object i belongs either to class ω_k or to class ω_ℓ, and the weight of this belief is equal to 0.7.

Let $M = (m_1, \ldots, m_n)$ denote the n-tuple of bba's related to the n objects. We shall call M a *credal partition* of O. Two particular cases are of interest:

- when each m_i is a *certain* bba, then M defines a conventional, crisp partition of O; this corresponds to a situation of complete knowledge;
- when each m_i is a *Bayesian* bba, then M specifies a fuzzy partition of O, as defined by Bezdek [2].

A credal c-partition (or partition of size c) will be defined as a credal partition $M = (m_1, \ldots, m_n)$ such that:

- each bba m_i, $i = 1, \ldots, n$ is defined on a frame Ω of c elements, and
- each class as a strictly positive degree of plausibility for at least one object, i.e., for all $\omega \in \Omega$, we have $\text{pl}_i(\{\omega\}) > 0$ for some $i \in \{1, \ldots, n\}$, pl_i being the plausibility function associated to m_i. Note that this condition is the equivalent of (2) in the definition of a fuzzy c-partition.

Example 1 Let us consider a collection O of $n = 5$ objects and $c = 3$ classes. A credal partition M of O is given in Table I. The class of object o_2 is known with certainty, whereas the class of o_5 is completely unknown. The other three cases correspond to situations of

partial knowledge (m_4 is Bayesian). The plausibilities $\mathrm{pl}_i(\{\omega\})$ of each singleton are given in Table II. Since each class is plausible for at least one object, M is a credal 3-partition of O. Note that the matrix given in Table II corresponds to a possibilistic partition as defined in [2].

INSERT TABLES I and II

B. Compatibility of an evidential partition with a dissimilarity matrix

As explained in the previous section, the notion of credal partition constitutes a reasonable framework for representing partial knowledge about the class membership of a set of objects. However, in an unsupervised learning situation, no prior knowledge about class labels is available, and one has to extract statistical knowledge from the data. In this section, we propose a principle that will provide the basis for inferring a credal partition from proximity data.

Without loss of generality, let us assume the available data to consist of a $n \times n$ dissimilarity matrix $D = (d_{ij})$, where $d_{ij} \geq 0$ measures the degree of dissimilarity between objects o_i and o_j. Matrix D will be supposed to be symmetric, with null diagonal elements. The dissimilarity values d_{ij} may have been assigned by experts (this is commonly the case in sensory analysis applications, see for example [4]), or they may be computed from attributes, possibly of different types [21]. Note that similarity values may be converted into dissimilarities by a transformation $d_{ij} = g(s_{ij})$, where g is a decreasing function.

As already mentioned in Section 1, objects within a group are usually expected to be similar to one another, and somewhat dissimilar from objects in other groups. As a consequence, observed dissimilarities can be interpreted as evidence concerning the group membership of objects, two similar objects being more likely to be in the same class than two dissimilar ones. The more similar, the more *plausible* it is that two objects belong to the same group. To formalize this idea, we need to express the plausibility, based on a credal partition, that two objects o_i and o_j are in the same group. This will then allow us to formulate a criterion of compatibility between a dissimilarity matrix D and a credal partition M.

Consider two objects o_i and o_j, and two bba's m_i and m_j quantifying one's beliefs regarding the class of objects i and j. To compute the plausibility that these two objects belong to the same class, we have to place ourselves in the Cartesian product $\Omega^2 = \Omega \times \Omega$, and combine the vacuous extensions of m_i and m_j using (13). Let $m_{i\times j}$ denote the result of this combination. We have:

$$m_{i\times j}(A \times B) = m_i(A) \cdot m_j(B), \qquad \forall A, B \subseteq \Omega, A \neq \emptyset, B \neq \emptyset \; . \tag{14}$$

It is a bba on Ω^2 that describes one's beliefs concerning the class membership of both objects. In Ω^2, the event "Objects o_i and o_j belong to the same class" corresponds to the following subset of Ω^2:

$$S = \{(\omega_1, \omega_1), (\omega_2, \omega_2), \ldots, (\omega_c, \omega_c)\}$$

Let $\mathrm{pl}_{i\times j}$ be the plausibility function associated to $m_{i\times j}$. We have

$$\begin{aligned}
\mathrm{pl}_{i\times j}(S) &= \sum_{\{A\times B\subseteq\Omega^2 \;|\; (A\times B)\cap S\neq\emptyset\}} m_{i\times j}(A \times B) && (15)\\
&= \sum_{A\cap B\neq\emptyset} m_i(A) \cdot m_j(B) && (16)\\
&= 1 - \sum_{A\cap B=\emptyset} m_i(A) \cdot m_j(B) && (17)\\
&= 1 - K_{ij} \; , && (18)
\end{aligned}$$

where K_{ij} is the degree of conflict between m_i and m_j, as defined by (7).

The plausibility that objects o_i and o_j belong to the same class is thus simply equal to one minus the degree of conflict between the bba's m_i and m_j associated to the two objects.

Given any two pairs of objects (o_i, o_j) and $(o_{i'}, o_{j'})$, it is natural to impose the following condition:

$$d_{ij} > d_{i'j'} \Rightarrow \mathrm{pl}_{i\times j}(S) \leq \mathrm{pl}_{i'\times j'}(S) \tag{19}$$

or, equivalently:

$$d_{ij} > d_{i'j'} \Rightarrow K_{ij} \geq K_{i'j'} \ , \tag{20}$$

i.e., the more dissimilar the objects, the less plausible it is that they belong to the same class, and the higher the conflict between the bba's. A credal partition M verifying this condition will be said to be *compatible* with D.

EXAMPLE 2 Let us consider the dissimilarity matrix D given in Table III. Is the credal partition M of Example 1 compatible with D ? The degrees of conflict K_{ij} $(i \neq j)$ among the bba's in M are shown in Table IV and plotted against the d_{ij} in Figure 1. As we can see, the rank order of dissimilarities is preserved in the degrees of conflict. Hence, M is compatible with D.

INSERT TABLES III and IV and Fig. 1

REMARK 1 Similarly, one could compute the degree of belief $\mathrm{bel}_{i\times j}(S)$ that o_i and o_j belong to the same class as

$$\begin{aligned} \mathrm{bel}_{i\times j}(S) &= \sum_{\{A\times B\subseteq\Omega^2 \ | \ \emptyset\neq(A\times B)\subseteq S\}} m_{i\times j}(A\times B) && (21)\\ &= \sum_{k=1}^{c} m_{i\times j}(\{(\omega_k,\omega_k)\}) && (22)\\ &= \sum_{k=1}^{c} m_i(\{\omega_k\})\cdot m_j(\{\omega_k\}) \ . && (23) \end{aligned}$$

However, this quantity seems to be less satisfactory than $\mathrm{pl}_{i\times j}(S)$ as a measure of "agreement" between m_i and m_j, as it depends only on the mass given to the singletons.

C. Learning a credal partition from data

The concept of compatible credal partition introduced in the previous section provides a basis for unsupervised learning procedures allowing to extract a credal partition from dissimilarity data. If we accept the argument developed in Section III-B, we need a method that, given a dissimilarity matrix D, generates a credal partition M that is either compatible with D, or at least "almost compatible" (in a sense to be defined).

This problem happens to be quite similar to the one addressed by multidimensional scaling (MDS) methods [4]. The purpose of MDS is, given a dissimilarity matrix D, to find a configuration of points in a p-dimensional space, such that the distances between points approximate the dissimilarities. There is a large literature on MDS methods, which are used extensively in sensory data analysis for interpreting subjectively assessed dissimilarities, and more generally in exploratory analysis for visualizing proximity data as well as high dimensional attribute data (in this case, the dissimilarities are computed as distances in the original feature space).

In our problem, each object is represented as a bba, which can be seen (unless some restrictions are imposed on the belief masses) as a point in a 2^c-dimensional space. Our concept of "credal partition" thus parallels that of "configuration" in MDS. The degree of

conflict K_{ij} between two bba's m_i and m_j may be seen as a form of "distance" between the representations of objects o_i and o_j (although it does not satisfy all the axioms of a distance). This close connection allows us to transpose MDS algorithms to our problem, by optimizing the credal partition M, so that the degrees of conflict K_{ij} between objects reflect the corresponding dissimilarities d_{ij}.

MDS algorithms generally consist in the iterative minimization of a *stress function* measuring the discrepancies between observed dissimilarities and reconstructed distances in the configuration space. The various methods available differ by the choice of the stress function, and the optimization algorithm used.

The MDS approach that is the closest to our original goal is *non metric* or *ordinal* MDS. It minimizes a stress function such as (with our notations):

$$I_{nm}(M, f) \triangleq \frac{\sum_{i<j}[K_{ij} - f(d_{ij})]^2}{\sum_{i<j}[K_{ij} - \overline{K}]^2} \ , \tag{24}$$

where $\overline{K}$ is the average degree of conflict, and f is any increasing function. Typically, $I_{nm}(M, f)$ is minimized, alternatively, with respect to M using a gradient-based iterative procedure, and with respect to f using isotonic regression [4]. This is a very powerful approach, since it allows to consider only the rank order of dissimilarities. However, it may be quite heavy computationally.

A more restrictive, but simpler approach is obtained by imposing a parametric form to the relationship between "distances" (i.e., degrees of conflict in our case) and dissimilarities, which is referred to as *metric* MDS. A simple stress function used in this approach is the normalized Sammon's stress function [29] defined a

$$I(M, a, b) \triangleq \frac{1}{C} \sum_{i<j} \frac{(aK_{ij} + b - d_{ij})^2}{d_{ij}} \ , \tag{25}$$

where a and b are two coefficients, and C a normalizing constant defined as:

$$C \triangleq \sum_{i<j} d_{ij} \ . \tag{26}$$

This criterion can be minimized iteratively with respect to M, a and b using a gradient-based procedure such as described in Appendix II. Note that the Sammon's stress function weights more heavily the errors made on small dissimilarities, which proves to be an effective strategy in a clustering context. Also, I is invariant under any affine transformation of the dissimilarities. For these reasons, this criterion has been used in the experiments described in Section IV.

REMARK 2 Each bba m_i must take values in $[0, 1]$ and satisfy Eq. (3). Hence, the optimization of I with respect to M is a constrained optimization problem. However, the constraints vanish if one uses the following parameterization:

$$m_i(A_k) = \frac{\exp(\alpha_{ik})}{\sum_{l=1}^{f} \exp(\alpha_{il})}, \tag{27}$$

where $A_k, k = 1, \ldots, f$ are the f focal sets ($f = 2^c$ in the general case), and the α_{ik} for $i = 1, \ldots, n$ and $k = 1, \ldots, f$ are nf real parameters representing the credal partition. In the following, these parameters will be identified with the corresponding credal partition: $M \equiv (\alpha_{ik})$.

D. Controlling the complexity

An important issue that deserves special attention in our approach is the dimension of the non linear optimization problem to be solved. The number of parameters to be optimized is linear in the number of objects but exponential in the number of clusters. Hence, computational problems may quickly arise when the number of classes increases. More fundamentally, the number of parameters may, for moderate n and large c, be commensurate with, or even greater than the number of observed dissimilarities, so that the problem can become severely ill-posed. Consequently, the number of degrees of freedom has to be controlled. This can be achieved in two ways:

First, the number of parameters may be drastically decreased by considering only a subclass of bba's with a limited number of focal sets. For example, we may constrain the focal sets to be either Ω, the empty set, or a singleton. In this way, the total number of parameters is reduced to $n(c+2)$, without sacrificing too much of the flexibility of belief functions. Note that, if one further restricts the bba's to be Bayesian (i.e., probability functions), our approach boils down to Sato's additive fuzzy clustering model [30]. However, the possibility to assign a mass to Ω and to the empty set endows our method with greater flexibility and interpretability, as will be shown in the sequel.

Another very efficient means of controlling the model complexity is to add a penalization term to the stress function. A similar approach is used, for instance, in neural network training to favor smooth input-output mappings (see, e.g., Ref. [3]). In a supervised learning context, this is achieved by considering a model with many adjustable parameters, and then adding to the error function a term that penalizes highly complex models. Searching for the minimum of the resulting criterion is then a way to balance error on the training set with complexity. In our case, the "error" term is the stress function, and we have to find a penalty term that favors "simple" bba's.

As we would like to extract as much information as possible from the data, it is reasonable to require the bba's to be as "informative" as possible. The definition of the "quantity of information" contained in a belief function has been the subject of a lot of research in the past few years [25], [23], and it is still, to some extent, an open question. However, several entropy measures have been proposed. Although the debate as to which one is the best is still going on, the total uncertainty introduced by Pal et al. [26] satisfies natural requirements and has interesting properties. It is defined, for a normal bba m, as:

$$H(m) \triangleq \sum_{A \in \mathcal{F}(m)} m(A) \log_2 \left(\frac{|A|}{m(A)} \right) \ , \tag{28}$$

where $\mathcal{F}(m)$ denotes the set of focal sets of m. This entropy measure can be decomposed as the sum of two terms:

$$H(m) = \sum_{A \in \mathcal{F}(m)} m(A) \log_2 |A| - \sum_{A \in \mathcal{F}(m)} m(A) \log_2 m(A) \ . \tag{29}$$

The first term is the nonspecificity measure [23], which reflects the degree of imprecision of m, while the second term gauges the inconsistency in m, and can be seen as a measure of conflict. Hence, $H(m)$ tends to be small when the mass is assigned to few focal sets, with small cardinality (it is proved in [26] that $H(m) = 0$ iff $m(\{\omega\}) = 1$ for some $\omega \in \Omega$).

The entropy measure defined by (28) was originally proposed for normal bba's (i.e., bba's m such that $m(\emptyset) = 0$), and some generalization to subnormal bba's has to be performed for our purpose. In [23, page 51], Klir and Wierman define the nonspecificity of a subnormal bba as:

$$N(m) \triangleq \sum_{A \in \mathcal{F}(m) \setminus \emptyset} m(A) \log_2 |A| + m(\emptyset) \log_2 |\Omega| \ . \tag{30}$$

This extension is justified by the fact that the mass given to the empty set corresponds to a situation of maximal uncertainty, just as the mass given to Ω. Replacing the first term in (29) by (30) leads to the following definition for the entropy of a normal or subnormal bba:

$$H(m) \triangleq \sum_{A \in \mathcal{F}(m) \setminus \{\emptyset\}} m(A) \log_2 \left(\frac{|A|}{m(A)} \right) + m(\emptyset) \log_2 \left(\frac{|\Omega|}{m(\emptyset)} \right). \tag{31}$$

Finally, the objective function to be minimized is:

$$J(M, f) \triangleq I(M, f) + \lambda \sum_{i=1}^{n} H(m_i), \tag{32}$$

where λ is the penalization coefficient that controls the extent to which the entropy term influences the form of the solution. Increasing λ will result in "simpler" bba's with a smaller number of focal sets.

E. Summary and complexity analysis

The overall EVCLUS method is summarized in Table V. Details concerning the computation of the stress function J and its the gradient are reported in Appendix I. The gradient-based optimization procedure used in our implementation of the method is described in Appendix II.

INSERT TABLE V

Concerning the algorithmic complexity of the method, equations (38) to (43) show that the computation of the stress criterion can be performed in $O(f^2 n^2)$ operations, where f is the number of focal sets ($f = c+2$ if they are restricted to Ω, $\emptyset$, and the c classes), whereas the complexity of the gradient calculation is $O(f^3 n^2)$ (see equations (44) to (51)). Hence, one iteration of the optimization procedure necessitates $O(f^3 n^2)$ operations. This confirms the practical interest of reducing the number of focal sets. For a few hundred objects and a limited number of classes (say, less than 20), which is the case in most clustering studies, calculations are easily tractable. For instance, for the cat cortex example described in Section IV-B (63 samples, 4 classes), each run of the algorithm implemented in Matlab takes about 30 seconds on a PC equipped with a Pentium II processor. For the protein dataset (Section IV-C), which involves 213 samples and 4 classes, each run takes about 3 minutes.

F. From credal clustering to fuzzy or hard clustering

A credal partition contains a lot of information, and it may be useful to summarize it in the form of a fuzzy or crisp partition. This may be achieved using the pignistic transformation introduced by Smets [37], which provides a convenient and justified mechanism for converting a normal bba m into a probability function p_m, defined as

$$p_m(\omega) \triangleq \sum_{\{A \subseteq \Omega \ / \ \omega \in A\}} \frac{m(A)}{|A|}, \tag{33}$$

for all $\omega \in \Omega$. In Smets' Transferable Belief Model [37], pignistic probabilities are used for decision making, in contrast to bba's which are used to represent and update beliefs. For a subnormal bba, a preliminary normalization step has to be performed. Two methods are Dempster's normalization, in which the masses given to non empty sets are divided by $1 - m(\emptyset)$, and Yager's normalization, in which the mass $m(\emptyset)$ is transfered to Ω [40].

A fuzzy partition may be obtained by calculating the pignistic probability function p_i induced by each bba m_i, and interpreting $p_i(\omega_k)$ as the degree of membership of object i to group k. In the particular case where the singletons, Ω and the empty set are the only

focal sets of m, and the Yager's normalization is used, the pignistic probabilities have the following expression:

$$p_i(\omega_k) = m_i(\{\omega_k\}) + \frac{m_i(\Omega) + m_i(\emptyset)}{c} \quad k = 1, c \tag{34}$$

A hard partition can then be easily obtained by assigning each object to the group with highest pignistic probability. In this sense, a credal partition may be viewed as a rich and general model of partitioning, from which fuzzy and hard partitions can be computed as by-products.

INSERT TABLE VI

IV. Results

A. Synthetic data set

This first example is adapted from a classical data set [38]. It is composed of a matrix of dissimilarities between 13 objects (Table VII). The dissimilarities between objects 2 to 13 were computed as the squared Euclidean distances between 12 points in a two-dimensional attribute space, as shown in Figure 2. Eleven of these points are part of Windham's data, whereas object 13 is an outlier and has been added to test the robustness of the method.

Additionally, the dissimilarities to a 13-th object (number 1) have been added in the data matrix. Object 1 does not have a representation in the attribute space and is quite similar to all other objects (see the first row and the first column of Table VII). Such a situation cannot occur when the dissimilarities are Euclidean distances, but it can be found when analyzing proximity data: for instance, in a sensory evaluation experiment, an object may be found similar to all other objects by an assessor (this may also be due to measurement or transcription errors). In contrast to object 13, which is an outlier, object 1 might be called an "inlier", i.e., an object which is atypical because of its closeness to all other objects. The task is to find a reasonable 2-partition of object 2 to 12 and to detect the particularity of objects 1 and 13.

INSERT TABLE VII

We compared the results obtained with our method and five classical clustering methods for relational data: the assignment-prototype algorithm (AP) [38], the Fuzzy Non Metric algorithm (FNM) [28], the Robust Fuzzy c-means algorithm (RFCM) [16], and its "Noise" version (NRFCM) [5], and the non-Euclidean RFCM algorithm (NERF) [15]. NRFCM, by using a "noise" cluster, is well-adapted to data sets containing noise and outliers, whereas NERF was designed to cope with non-Euclidean dissimilarities.

Among these five algorithms, three algorithms (FNM, RFCM and its noise version NRFCM) have a "fuzzification constant" h that controls the degree of "hardness" of the resulting fuzzy partition. A value $h = 2$ seems to be commonly adopted in most of the clustering studies [1]. NRFCM has another parameter δ, defined as the distance of a prototypical member of the "noise cluster" to all other objects.

Figure 3 shows the resulting fuzzy membership functions for the five classical algorithms (with h=2 and $\delta = 50$), and the bba's obtained using the EVCLUS algorithm with $\lambda = 0.05$ ($m_i(\{\omega_1\})$, $m_i(\{\omega_2\})$, $m_i(\Omega)$ and $m_i(\emptyset)$ are plotted against i). Object 13 (the outlier) is assigned a high degree of membership to class 2 by all methods except NRFCM and EVCLUS, which correctly detect this object as an outlier (by classifying it to the noise class for NRFCM, and by assigning almost all the mass to the empty set for EVCLUS). The main difference between NRFCM and EVCLUS concerns object 1 (the "inlier"), which is

[1]The Matlab code of NERF available at `http://www2.gasou.edu/facstaff/hathaway` uses exclusively this value.

arbitrarily given a high degree of membership to class 1 by all methods except EVCLUS, which for that object allocates almost all the mass to Ω.

Another view of the credal partition obtained by EVCLUS is given in Figure 4, where the plausibilities of class 1 and class 2 are represented for the 13 objects. We can see that both classes are completely plausible for object 1 (with $\mathrm{pl}_1(\{\omega_1\}) = \mathrm{pl}_1(\{\omega_2\}) \approx 1$), whereas none of the two classes is considered plausible for the outlier, since we have: $\mathrm{pl}_{13}(\{\omega_1\}) = \mathrm{pl}_{13}(\{\omega_2\}) \approx 0$. Note that assigning a large fraction of the unit mass to $\emptyset$ in the case of an outlier is consistent with the interpretation of $m(\emptyset)$ under the open-world assumption [34], an outlier being an object which does not appear to belong to any of the groups in presence.

Figure 5 shows the degrees of conflict K_{ij} plotted against the dissimilarities d_{ij} (such a plot is sometimes called a "Shepard diagram" in the MDS literature). This display allows to check visually the quality of the evidential partition (i.e., the rank order agreement between degrees of conflict and dissimilarities).

INSERT FIGURES 2, 3, 4 and 5

Remark 3 A situation as the one considered in this section, in which an object is very similar to objects that are quite dissimilar betwen themselves, is by no means unrealistic, as inconsistencies cannot always be avoided in real data sets. Our experience with sensory evaluation data shows us that large unexpected errors are not uncommon, even in carefully prepared and supervised experiments. For instance, an assessor may confuse an object with another, click on the wrong button, invert an intensity scale, etc. In large amounts of data, such inconsistences may be very difficult to detect merely by visual inspection of the data. Hence, it is important to have data analysis tools that allow to systematically bring such anomalies to the attention of the data analyst. EVCLUS is such a method: a large mass assigned to the empty set or to Ω is an indication of the peculiarity of an object, which has then to be explained by the analyst. Even more fundamentally, we have only assumed the d_{ij} to be dissimilarities, which is a very weak assumption. Typically, this notion is simply defined by two axioms [21]: commutativity ($d_{ij} = d_{ji}$), and

$$d_{ij} \geq d_{ii} \quad \forall j \neq i.$$

The triangular inequality, which is a "consistency principle" in the definition of a distance, is not postulated. Hence, it is not necessarily incoherent, for certain interpretations of the dissimilarities, to consider that an object may be similar to two objects, which are themselves dissimilar. For instance, assume that there exists a distance measure between objects, but the distances are only partially known. Let d_{ij} be the *smallest possible distance* between objects i and j. Then, if the true distances to object k are unknown, we may have $d_{ik} \approx 0$ and $d_{jk} \approx 0$ (because the corresponding distances *may be* small), even if d_{ij} is large.

B. "Cat cortex" data set

This real data set consists of a matrix of connection strengths between 65 cortical areas of the cat. It was collected by Scannell [31] and used by several authors to test visualization, discrimination or clustering algorithms for proximity data [12], [13], [17]. The proximity values are measured on an ordinal scale and range from 0 (self-connection), to 4 (absent or unreported connection) with intermediate values: 1 (dense connection), 2 (intermediate connection) and 3 (weak connection). From functional considerations, the cortex can be divided into four regions: auditory (A), visual (V), somatosensory (S), and frontolimbic (F). The clustering task is to find a four-class partition of the 65 cortical areas, based on the dissimilarity data, which is consistent with the functional regions.

We ran EVCLUS 50 times with different random initializations (with $\lambda = 0.01$, and 6 focal elements: $\{\omega_i\}, i = 1, \ldots, 4$, Ω and $\emptyset$), and the solution with the minimum value of the objective function was retained. A two-dimensional map of the data was obtained by means of a MDS algorithm. Figure 6 shows the fuzzy and hard partitions derived from the credal partition using the pignistic transformation defined by Eq. (33). In this representation, a different symbol is used for each cluster, and the symbol size is proportional to the maximum of the pignistic probabilities. It can be seen that the four clusters are almost perfectly recovered, with an error rate of 4.6 % (only three points among 65 are misclassified). This is consistent with the leave-one-out error rates obtained by Graepel et al. [13], [14]) using various supervised learning algorithms (which, consequently, use more information).

INSERT FIGURE 6

We also applied the five algorithms described in the previous section using the same experimental settings: for different values of h between 1.05 and 2, each algorithm was run 50 times, and the solution with minimum objective value is retained. All methods converged towards a useless solution with equal membership of $1/c$ (or $1/(c+1)$ in the case of NRFCM) for all points when the fuzzification constant h is set to 2. FNM, RFCM, NRFCM are able to give a reasonable solution for small values of the parameter h ($h < 1.4$). The best solution, exhibiting a misclassification rate of 4.6 %, is obtained using RFCM with $h = 1.2$. EVCLUS thus perform as well as RFCM when applied to this kind of data.

C. "Protein" data set

This real data set consists of a proximity matrix derived from the structural comparison of 213 protein sequences (see [18] and [13] for previous experiments with these data). Each of these proteins is known to belong to one of four classes of globins: hemoglobin-α (HA), hemoglobin-β (HB), myoglobin (M) and heterogeneous globins (G). It is thus of interest to see whether these four classes can be recovered in an unsupervised way from the dissimilarities only.

EVCLUS provides an interesting 4-class solution, as shown in Figure 7. This solution was obtained with the same experimental settings as for the cat cortex dataset (50 random trials, $\lambda = 0.005$, and 6 focal elements: $\{\omega_i\}, i = 1, \ldots, 4$, Ω and $\emptyset$). A two-dimensional MDS configuration together with the pignistic and hard partitions are shown in Figure 7. The hard partition is almost identical to the known classification of the proteins in 4 classes, only one point out of 213 is misclassified.

For the same data set,with a fuzzification constant $h = 2$, the five algorithms mentioned in the previous section (AP, FNM, RFCM, NRFCM and NERF) converge towards a trivial solution with equal membership of $1/c$ for all points (or $1/(c+1)$ in the case of NRFCM), and fail to find a meaningful partition into 4 clusters. The best results are obtained with RFCM (or equivalently NRFCM) with a fuzzification constant set to 1.05: 5 proteins out of 213 are misclassified.

EVCLUS provides some additional information via the analysis of the mass allocated to the empty set, as shown in Figure 8, where the size of the symbols is proportional to the mass of the empty set. Surprisingly, the highest masses are attributed to the members of the G-class, although these points cannot be considered at first sight as outliers (one should be aware, however, that the two-dimensional MDS configuration shown in Figures 7 and 8 does not necessarily constitute a faithful representation of the data set). Figure 9 provides a reasonable explanation. It shows, for each cluster, a box-plot of the dissimilarities between, on the one hand, proteins belonging to the same class, (within-class dissimilarities) and, on the other hand, proteins belonging to different classes (between-class dissimilarities). One can see that the G-class is characterized by high within-class dissimilarity values, and small differences between within and between-class dissimilarities.

The mass allocated to the empty set thus acts as a detector of the peculiarity of this class. Furthermore, it contributes to the robustness of the algorithm, by allowing to find a credal partition such that the bba's for class G objects have high degree of conflict with all other bba's, inside and outside their class.

The G-class can also be detected as a special cluster using NRFCM for certain values of parameter δ: depending on this parameter, the elements of the G-class are either well-clustered, or rejected in the noise class. However, the elements of that class cannot be considered strictly speaking as outliers. EVCLUS avoid this difficulty by operating at two levels: a credal level where masses are allocated to the focal elements, and a decision level where a partition is computed from the pignistic probabilities. The peculiarity of the G-class is represented at the credal level, whereas the correct partition is found out at the pignistic level.

INSERT FIGURES 7, 8, 9

V. Combining several credal partitions

A. Basic approach

Let us assume that we have p dissimilarity matrices $D^{(j)}$, $j = 1, ..., p$ obtained from p experts or information sources. Examples of such situations are sensory evaluation experiments (see Section V-B), in which different products are compared by several assessors, and time-dependent data, in which measurements of dissimilarity between objects are available at successive time steps (see, e.g. [30]). Combining the dissimilarities by, e.g., averaging is not necessarily a good approach as this may hide conflict between information sources. A better approach may be to induce credal partitions from each dissimilarity matrix, and to combine these partitions using suitable operators from evidence theory. However, when performing such a combination, some caution must be exercised because similar clusters may appear in different partitions with different numbers. Consequently, all permutations of the indices $\mathbb{N}_c = \{1, \ldots, c\}$ must be considered before combination to find the "best match" between the partitions.

More precisely, let us assume that we have two evidential partitions $M^{(1)} = (m_1^{(1)}, \ldots, m_n^{(1)})$ and $M^{(2)} = (m_1^{(2)}, \ldots, m_n^{(2)})$ (the generalization to p partitions is straightforward). These credal partitions may be considered as equivalent ($M^{(1)} \equiv M^{(2)}$) if there exists a permutation σ such that, for all $\forall i \in \{1, \ldots, n\}$, we have

$$m_i^{(1)}(\{\omega_{i_1}, \ldots, \omega_{i_r}\}) = m_i^{(2)}(\{\omega_{\sigma(i_1)}, \ldots, \omega_{\sigma(i_r)}\}) \quad \forall \{i_1, \ldots, i_r\} \subseteq \mathbb{N}_c \ . \tag{35}$$

We shall denote by $\mathcal{C}(M^{(1)})$ the equivalence class of $M^{(1)}$ (i.e., the set of all credal partitions M such that $M \equiv M^{(1)}$).

More generally, we may define the discrepancy between two evidential partitions $M^{(1)}$ and $M^{(2)}$ by:

$$\delta(M^{(1)}, M^{(2)}) = \sum_{i=1}^{n} K(m_i^{(1)}, m_i^{(2)}),$$

where $K(m_i^{(1)}, m_i^{(2)})$ is the degree of conflict between $m_i^{(1)}$ and $m_i^{(2)}$. This measure allows to find, in the equivalence class of $M^{(2)}$, the "best match" $M^{(2)*}$ with $M^{(1)}$, defined by:

$$\delta(M^{(1)}, M^{(2)*}) = \min_{M \in \mathcal{C}(M^{(2)})} \delta(M^{(1)}, M) \ . \tag{36}$$

The conjunctive sum of $M^{(1)}$ and $M^{(2)}$ can then be defined as the credal partition obtained by combining $m_i^{(1)}$ with $m_i^{(2)*}$ conjunctively, for all $i \in \{1, \ldots, n\}$:

$$M^{(1)} \textcircled{\cap} M^{(2)} = (m_1^{(1)} \textcircled{\cap} m_1^{(2)*}, \ldots, m_n^{(1)} \textcircled{\cap} m_n^{(2)*}) \ . \tag{37}$$

An intermediate step that may be carried out before the combination is to expand the frame of discernment by adding an additional class ω_{c+1}, and to transfer each mass $m_i(\emptyset)$ to $\{\omega_{c+1}\}$. This operation results from the interpretation of $m_i(\emptyset)$ as a mass of belief supporting the hypothesis that object i might not belong to any class in Ω (open-world assumption). Defining an additional cluster (similar to the "noise cluster" in NRFCM) is then a way to restore the close-world assumption.

To summarize, the fusion of two credal partitions actually involves three steps: matching, expansion of the frame, and combination. This approach is demonstrated on a real example in the following section.

B. Sensory data set

The considered data set comes from a sensory profiling experiment in which assessors give a score to a number of products for several sensory attributes. The collected data thus consists of several products-attributes matrices (one for each assessor), from which dissimilarity matrices can be derived using, e.g., the Euclidean distance measure. The goal of this particular study was to understand the overall perceived association or relationship between 13 products described using four attributes. For that purpose, it is useful to search for a consensual partition of the products in the attribute space.

The EVCLUS algorithm was applied to each expert's dissimilarity matrix, to find a two-class credal partition of the products. The expert partitions were then merged in a conjunctive manner using the Dempter's rule of combination. Tables VIII and IX show the credal partitions $M^{(1)}$ and $M^{(2)}$ obtained by applying EVCLUS to the dissimilarity matrices of two experts with the following settings: $c = 2$, four focal elements $\{\omega_1\}$, $\{\omega_2\}$, Ω, and $\emptyset$, $\lambda = 0.005$. In the hard partition induced by $M^{(1)}$, cluster 1 is composed of objects 1, 3, 9, 11 and 13, whereas in the hard partition induced by $M^{(2)}$, cluster 2 is composed of the same objects, except object 3. This suggests the existence of a good agreement between both experts, who disagree only about object 3. Table X shows the credal partition $M^{(2)*}$ obtained by permuting clusters 1 and 2, and transferring the masses $m_i^{(2)}(\emptyset)$ to a third class ω_3.

Finally, the result of the conjunctive combination of $M^{(1)}$ (after expansion of the frame) and $M^{(2)*}$ is shown in Table XI. The same results are displayed in different form in Figures 10 and 11. As can be seen, the two partitions are in a good agreement except for product 3. The fused partition confirms this consensus and reveals the conflict about the third point through the important mass allocated to the empty set.

INSERT TABLES VIII, IX, X, XI

INSERT FIGURES 10 and 11

In the same way, it is possible to fuse any number of partitions to derive a consensual partition among several experts. However, the Dempster's rule of combination is known to suffer from a major drawback: as new sources are added, the degree of conflict increases, and the mass allocated to the empty set becomes more and more important. Until the development of more robust combination rules, the presented approach remains practically limited to a small number of experts.

VI. Choosing the number of classes

One of the fundamental issue in clustering is the determination of the proper number of classes of the partition. This is a difficult problem which has been intensively studied in the past and numerous procedures have been proposed (see, for example, [6], [24], [19], [27]). Most of the methods involve plotting the value of some validity criterion against the number of groups and looking either for a large change in the criterion ("elbow" solution) or a minimum or maximum of the criterion. We propose to use the value of the stress J as

an internal index of validity of our method and illustrate its application using a synthetic data set. The data set consists of Euclidean distances between 100 points (shown in Figure 12) simulated from four bivariate normal distributions (25 points in each cluster) with the following mean vectors:

$$m_1 = [-2.5; 2.5]^t \quad m_2 = [2.5; 2.5]^t \quad m_3 = [2.5; -2.5]^t \quad m_4 = [-2.5; -2.5]^t,$$

and common covariance matrix equal to the identity matrix. There is no significant overlap between the clusters which exhibit reasonable properties of "external" isolation and "internal" cohesion.

INSERT FIGURES 12 and 13

The following experimental setup was applied: for different values of λ, (namely 0.01, 0.05, 0.1 and 0.2), the number c of classes was varied from 2 to 6. For each value of c, the EVCLUS algorithm was run 10 times and the minimum value of J was retained. Results are displayed in Figure 13. It can be seen that the minimum value of the criterion is always obtained for $c = 4$ clusters, this minimum being more and more pronounced as λ increases. This simple example suggests that the stress value can be considered as a valuable validity index to indicate the cluster structure in a data set.

Concerning the determination of λ, we can remark that the purpose of this hyper-parameter is twofold: it controls both the hardness of the resulting partition, and the number of free parameters of the method, thus improving the optimization tractability. The first point is user and data dependent. The greater λ, the harder is the partition. For the second point, some general guidelines can be given to the practitioner. First it can be noticed that the range of variations of λ is not very large, independently of the stress function chosen. The regularization term in the stress function should not take too much importance in the optimization process, as compared to the term which measures the discrepancy between the dissimilarities and the degrees of conflict. Typically, in each experiment, values ranging between 0.005 and 0.1 have been used. Optimization problems such as the convergence to a spurious local minimum can indicate that a too large value has been used, so that the number of free parameters has been overreduced and the model is unable to fit the data. On the contrary, a bad stability of the results among several runs of the algorithm indicates that the problem is not constrained enough and that λ should be increased. Finally an examination of the Shepard's diagram (plot of K_{ij} vs. d_{ij}) can also give very useful indications regarding the hardness of the partition and the quality of the fit.

VII. Conclusions

We have suggested a new method for clustering relational data based on the theory of belief functions. In this approach, called EVCLUS, each object is assigned a bba over a given set of classes, in such a way that the degree of conflict between two bba's reflects the dissimilarity between the corresponding objects. The set of bba's forms a "credal partition", which extends the existing concepts of hard, fuzzy (probabilistic) and possibilistic partition. This additional flexibility results both in greater expressive power (allowing, e.g., to detect the dissimilarity of an object from all clusters or, conversely, the similarity to all or several clusters), and in improved robustness with respect to atypical data. The method was shown to be capable of discovering meaningful clusters in several non-Euclidean data sets, and its performances compared favorably with those of state-of-the-art techniques such as the NRFCM algorithm. Furthermore, it allows to combine clustering results obtained from several dissimilarity matrices provided by different experts or measurement devices.

The theory of belief function constitutes a rich framework for representing partial knowledge, and formalizing the clustering problem in this context offers several perspectives,

many of which remain to be explored. For example, an interesting possibility in some applications might be to combine data with prior knowledge regarding plausible groupings of the objects. This, and other extensions of the present work are left for further study.

Acknowledgments

The authors would like to thank Thore Graepel for providing the protein and the cat cortex data sets and PSA Peugeot Citröen for supporting this study and providing the sensory data set.

Appendix

I. Mathematical appendix

A. Calculation of the stress function

Let $\mathcal{F} = \{A_1, \ldots, A_f\} \subseteq 2^\Omega$ be the set of focal elements considered for the bba in the credal partition $M = (m_1, \ldots, m_n)$. A typical choice is $\mathcal{F} = \{\emptyset, \{\omega_1\}, \ldots, \{\omega_c\}, \Omega\}$. By convention, $\mathcal{F}$ will be assumed to contain the empty set, with $A_1 = \emptyset$.

Let m_{ik} denote the mass assigned to A_k ($k = 1, \ldots, f$) by the bba m_i. It is expressed as a function of parameters α_{il}, $l = 1, \ldots, f$ using the following re-parameterization:

$$m_{ik} = \frac{\exp(\alpha_{ik})}{\sum_{l=1}^{f} \exp(\alpha_{il})} \ . \tag{38}$$

The stress function minimized in EVCLUS is

$$J(M, a, b) = I(M, a, b) + \lambda \sum_{i=1}^{n} H_i, \tag{39}$$

with:

$$I = \frac{1}{C} \sum_i \sum_{j>i} \frac{(aK_{ij} + b - d_{ij})^2}{d_{ij}}, \tag{40}$$

where $C = \sum_i \sum_{j>i} d_{ij}$, and

$$H_i = \sum_{k=2}^{f} m_{ik} \log_2 \left(\frac{|A|}{m_{ik}} \right) + m_{i1} \log_2 \left(\frac{|\Omega|}{m_{i1}} \right) \tag{41}$$

Note that J is actually of function of $nf+2$ parameters: α_{ik} ($i = 1, \ldots, n$, $k = 1, \ldots, f$), a and b. The degrees of conflict in (40) may be expressed as:

$$K_{ij} = \sum_{k=1}^{f} \sum_{k'=1}^{f} m_{ik} m_{jk'} \xi_{kk'}, \tag{42}$$

with:

$$\xi_{kk'} = \begin{cases} 1 & \text{if } A_k \cap A_{k'} = \emptyset \\ 0 & \text{otherwise} \end{cases} \ . \tag{43}$$

B. Gradient calculation

This section gives some details concerning the computation of the partial derivatives of J with respect to the learning parameters (a, b and the α_{ik}).

We have

$$\frac{\partial J}{\partial \alpha_{il}} = \frac{\partial I}{\partial \alpha_{il}} + \lambda \frac{\partial H_i}{\partial \alpha_{il}}, \qquad \frac{\partial J}{\partial a} = \frac{\partial I}{\partial a}, \qquad \frac{\partial J}{\partial b} = \frac{\partial I}{\partial b}. \tag{44}$$

The following partial derivatives can then be easily obtained:

$$\frac{\partial I}{\partial a} = \frac{2}{C} \sum_i \sum_{j>i} \frac{K_{ij}(aK_{ij} + b - d_{ij})}{d_{ij}} \tag{45}$$

$$\frac{\partial I}{\partial b} = \frac{2}{C} \sum_i \sum_{j>i} \frac{(aK_{ij} + b - d_{ij})}{d_{ij}} \tag{46}$$

$$\frac{\partial I}{\partial \alpha_{il}} = \frac{2a}{C} \sum_{j>i} \frac{(aK_{ij} + b - d_{ij})}{d_{ij}} \frac{\partial K_{ij}}{\partial \alpha_{il}} \tag{47}$$

$$\frac{\partial K_{ij}}{\partial \alpha_{il}} = \sum_{k,k'} \frac{\partial m_{ik}}{\partial \alpha_{il}} m_{jk'} \xi_{kk'} \quad i,j \in \{1,\ldots,n\},\ l \in \{1,\ldots,f\} \tag{48}$$

$$\frac{\partial m_{ik}}{\partial \alpha_{il}} = \begin{cases} m_{ik}(1 - m_{ik}) & l = k \\ -m_{ik} m_{il} & l \neq k \end{cases} \quad i \in \{1,\ldots,n\},\ k,l \in \{1,\ldots,f\} \tag{49}$$

The partial derivatives of the regularization term H_i can be obtained in a similar fashion, starting from:

$$H_i = \sum_{k=1}^{f} m_{ik} \log_2(|A_k|) - \sum_{k=1}^{f} m_{ik} \log_2 m_{ik} \tag{50}$$

with, by abuse of notation, $|A_1| = |\Omega|$. Then,

$$\frac{\partial H_i}{\partial \alpha_{il}} = \sum_{k=1}^{f} \frac{\partial m_{ik}}{\partial \alpha_{il}} \left[\log_2 \left(\frac{|A_k|}{m_{ik}} \right) - 1 \right] . \tag{51}$$

II. Optimization algorithm

The EVCLUS method is fundamentally based on the minimization of the stress function $J(M, a, b)$, which can be performed using any unconstrained nonlinear programming algorithm. In the following, we sketch the particular algorithm used in the experiments reported in Section IV, which is quite similar to the method described in [33] for neural network training.

Let $\mathbf{w}$ be the vector of parameters, and $J(\mathbf{w})$ the objective function to be minimized. The algorithm is a variant of gradient descent in which each parameter w_i has its own step size η_j, and the step sizes are adapted during the optimization process, depending on the evolution of the objective function and on the sign of the derivatives at successive iterations. Let t be the iteration counter. Let us first assume that the objective function has decreased between iterations $t-1$ and t. Then the following rule is applied to update each step size η_j:

$$\eta_j(t) = \begin{cases} \beta\, \eta_j(t-1) & \text{if } \dfrac{\partial J}{\partial w_j}(t-1) \cdot \dfrac{\partial J}{\partial w_j}(t) > 0 \\ \gamma\, \eta_j(t-1) & \text{otherwise,} \end{cases} \tag{52}$$

where $\beta > 1$ and $\gamma < 1$ are two coefficients. Hence, the step size is increased if the derivatives have kept the same sign during two iterations, and it is increased if the sign of the derivative has changed, which indicates that we have "jumped over" a minimum. The parameters are then updated by:

$$w_j(t+1) = w_j(t) - \eta_j(t) \frac{\partial J}{\partial w_j}(t). \tag{53}$$

If now the objective function has increased between iterations $t-1$ and t, all step sizes are decreased simultaneously:

$$\eta_j(t) = \delta\ \eta_j(t-1) \quad \forall j \tag{54}$$

with $\delta < 1$, and the parameters are updated starting from where they were at the previous iteration:

$$w_j(t+1) = w_j(t-1) - \eta_j(t)\frac{\partial J}{\partial w_j}(t-1). \tag{55}$$

We have used the following numerical values of the coefficients:

$$\beta = 1.2 \quad \gamma = 0.8 \quad \delta = 0.5$$

References

[1] A. Appriou. Probabilités et incertitude en fusion de données multi-senseurs. *Revue Scientifique et Technique de la Défense*, (11):27–40, 1991.

[2] J. C. Bezdek, J. Keller, R. Krishnapuram, and N. R. Pal. *Fuzzy models and algorithms for pattern recognition and image processing*. Kluwer Academic Publishers, Boston, 1999.

[3] C. M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, Oxford, 1995.

[4] I. Borg and P. Groenen. *Modern multidimensional scaling*. Springer, New-York, 1997.

[5] R.N. Davé. Clustering relational data containing noise and outliers. In *FUZZ'IEEE 98*, pages 1411–1416, 1998.

[6] D. Davies, and D. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,224–227, 1979.

[7] T. Denœux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. on Systems, Man and Cybernetics*, 25(05):804–813, 1995.

[8] T. Denœux. Analysis of evidence-theoretic decision rules for pattern classification. *Pattern Recognition*, 30(7):1095–1107, 1997.

[9] T. Denœux. A neural network classifier based on Dempster-Shafer theory. *IEEE Trans. on Systems, Man and Cybernetics A*, 30(2):131–150, 2000.

[10] D. Dubois and H. Prade. Representation and combination of uncertainty with belief functions and possibility measures. *Computational Intelligence*, 4:244–264, 1988.

[11] J.C. Gower. Genralized Procustes Analysis. *Psychometrika*, 40:33–51, 1975.

[12] T. Graepel and K. Obermayer. A stochastic self-organizing map for proximity data. *Neural Computation*, 11:139–155, 1999.

[13] T. Graepel, R. Herbrich, P. Bollmann-Sdorra, and K. Obermayer. Classification on pairwise proximity data. in *Advances in Neural Information Processing Systems 11*, M. Kearns, S. Solla, and D. Kohn, eds., MIT Press, Cambridge, MA, 438–444, 1999.

[14] Thore Graepel, Ralf Herbrich, Bernhard Schölkopf, Alex Smola, Peter Bartlett, Klaus Robert-Maller, Klaus Obermayer, and Robert Williamson. Classification on Proximity Data with LP–Machines. in *Proceedings of the Ninth International Conference on Artificial Neural Networks*,304–309, 1999.

[15] R.J. Hathaway and J.C. Bezdek. Nerf c-means : Non-euclidean relational fuzzy clustering. *Pattern Recognition*, 27:429–437, 1994.

[16] R.J. Hathaway, J.W. Davenport, and J.C. Bezdek. Relational duals of the c-means clustering algorithms. *Pattern recognition*, 22(2):205–211, 1989.

[17] T. Hofmann, and J. Buhmann. Multidimensional scaling and data clustering. in *Advances in Neural Information Processing Systems 7*, G. Tesauro, D. Touretzky and T. Leen, eds., MIT Press, Cambridge, MA, 459–466, 1995.

[18] T. Hofmann, and J. Buhmann. Pairwise data clustering by deterministic annealing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(1):1–14, 1997.

[19] A. Jain, and J. Moreau. Bootstrap technique in cluster analysis. *Pattern Recognition*,20:547–569,1987.

[20] A. K. Jain and R. C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Englewood Cliffs, NJ., 1988.

[21] L. Kaufman and P. J. Rousseeuw. *Finding groups in data*. Wiley, New-York, 1990.

[22] H. Kim and P. H. Swain. Evidential reasoning approach to multisource-data classification in remote sensing. *IEEE Transactions on Systems, Man and Cybernetics*, 25(8):1257–1265, 1995.

[23] G. J. Klir and M. J. Wierman. *Uncertainty-Based Information. Elements of Generalized Information Theory*. Springer-Verlag, New-York, 1998.

[24] G.W. Milligan, and M.C. Cooper. An examination of procedures for determining the number of clusters in a data-set. *Psychometrika*,50:159–179,1985.

[25] N. R. Pal, J. C. Bezdek, and R. Hemasinha. Uncertainty measures for evidential reasoning I: A review. *International Journal of Approximate Reasoning*, 7:165–183, 1992.

[26] N. R. Pal, J. C. Bezdek, and R. Hemasinha. Uncertainty measures for evidential reasoning II: New measure of total uncertainty. *International Journal of Approximate Reasoning*, 8:1–16, 1993.

[27] N. Pal, and J. Bezdek. On cluster validity for the fuzzy c-means model. *IEEE Transactions on Fuzzy Systems*,3:370–376,1995.

[28] M. Roubens. Pattern classification problems and fuzzy sets. *Fuzzy sets and systems*, 1:239–253, 1978.

[29] J. W. Sammon. A non-linear mapping for data structure analysis. *IEEE Transactions on Computers*, 18:401–409, 1969.

[30] M. Sato, Y. Sato, and L. C. Jain. *Fuzzy clustering models and applications*. Physica-Verlag, Heidelberg, 1997.

[31] J.W. Scannell, C. Blakemore, and M.P. Young. Analysis of connectivity in the cat cerebral cortex. *Journal of Neuroscience*, 15(2):1463–1483, 1995.

[32] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, N.J., 1976.

[33] F. M. Silva and L. B. Almeida. Speeding up backpropagation. In *Advanced Neural Computers*, R. Eckmiller, ed., Elsevier-North-Holland, New-York, 151–158, 1990.

[34] P. Smets. The combination of evidence in the Transferable Belief Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):447–458, 1990.

[35] P. Smets. Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. *International Journal of Approximate Reasoning*, 9:1–35, 1993.

[36] P. Smets. The normative representation of quantified beliefs by belief functions. *Artificial Intelligence*, 92(1–2):229–242, 1997.

[37] P. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66:191–243, 1994.

[38] M.P. Windham. Numerical classification of proximity data with assignment measures. *Journal of classification*, 2:157–172, 1985.

[39] L. Xu, A. Krzyzak, and C. Y. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 22(3):418–435, 1992.

[40] R. R. Yager. On the normalization of fuzzy belief structure. *International Journal of Approximate Reasoning*, 14:127–153,1996.

Figure captions

- Figure 1: Degrees of conflict vs. dissimilarities (Example 2)
- Figure 2: Synthetic data set.
- Figure 3: Synthetic data set. Results of the six algorithms.
- Figure 4: Synthetic data set. Plausibilities of classes 1 and 2 for the 13 objects.
- Figure 5: Synthetic data set. Degrees of conflict K_{ij} vs. dissimilarities d_{ij}.
- Figure 6: MDS configuration for the Cat cortex data set. A different symbol is used for each group found by the EVCLUS algorithm, the symbol size being proportional to the pignistic probability of the corresponding group. The first letter of each label (S, V, A, F) indicates the true class memberships.
- Figure 7: MDS configuration for the Protein data set. A different symbol is used for each group found by the EVCLUS algorithm, the symbol size being proportional to the pignistic probability of the corresponding group. The labels (HA, HB, M, G) indicate the true class memberships.
- Figure 8: Protein data set. MDS configuration with mass $m_i(\emptyset)$ for each of the 213 objects.
- Figure 9: Protein data set. Box-plots of within and between-class dissimilarities.
- Figure 10: Sensory data set. Degree of belief $\mathrm{bel}_i(\{\omega_1\})$ and $\mathrm{bel}_i(\{\omega_2\})$ for the individual partitions (expert 1 and 2) and for the fused partition.
- Figure 11: Sensory data set. Mass of the empty set of the fused partition.
- Figure 12: Four-class data set.
- Figure 13: Four-class data set. Plot of the stress value against the number of clusters for different values of λ.

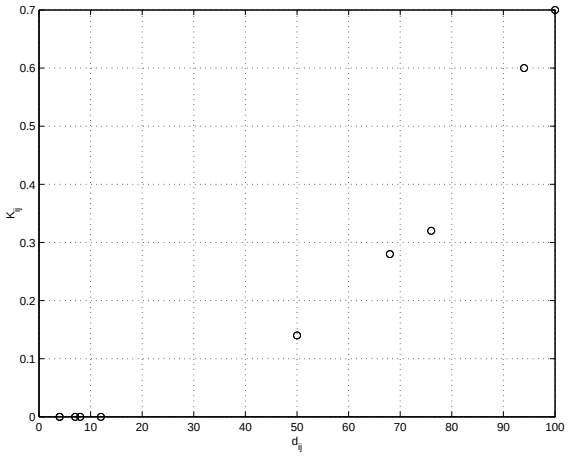
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
K_{ij}
0
10
20
30
40
50
60
70
80
90
100
d_{ij}

Fig. 1.

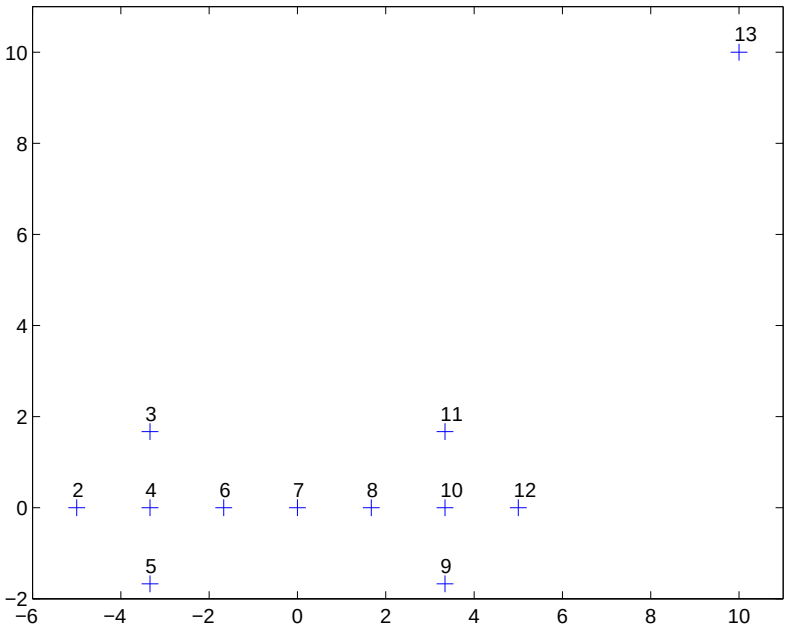
13
10
8
6
4
2
0
−2
3
11
2
4
6
7
8
10
12
5
9
−6
−4
−2
0
2
4
6
8
10

Fig. 2.

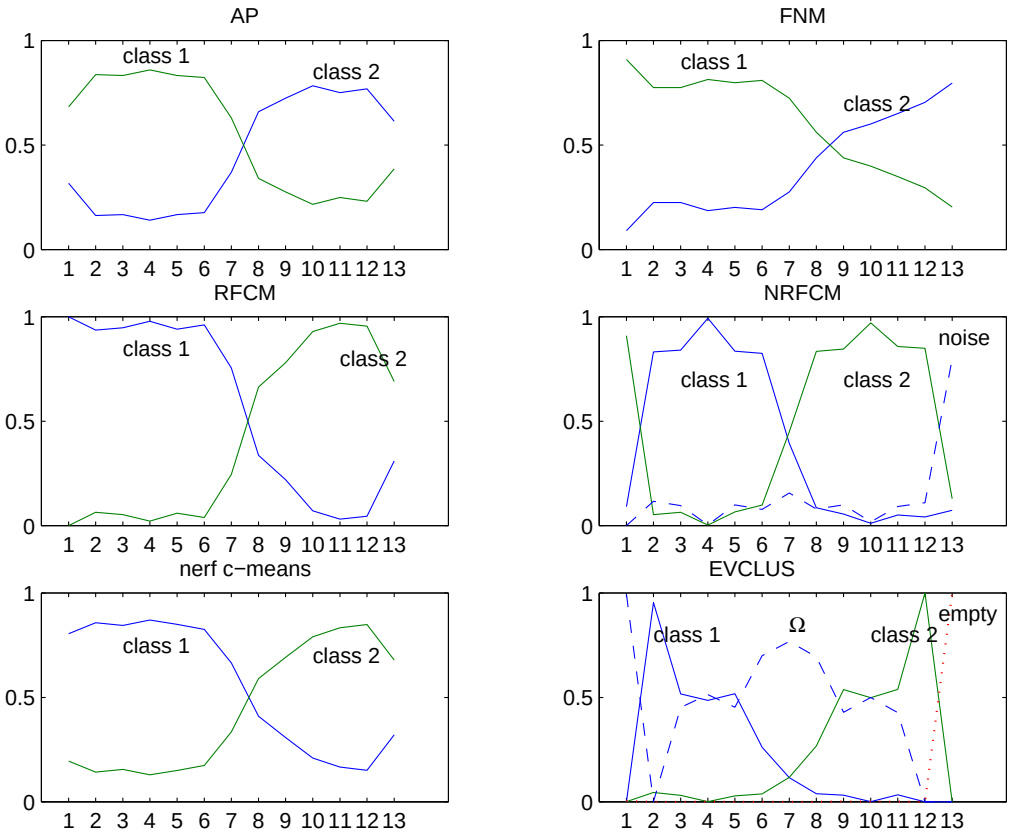
AP
class 1
class 2
FNM
class 1
class 2
RFCM
class 1
class 2
NRFCM
noise
class 1
class 2
nerf c−means
class 1
class 2
EVCLUS
class 1
Ω
class 2
empty

Fig. 3.

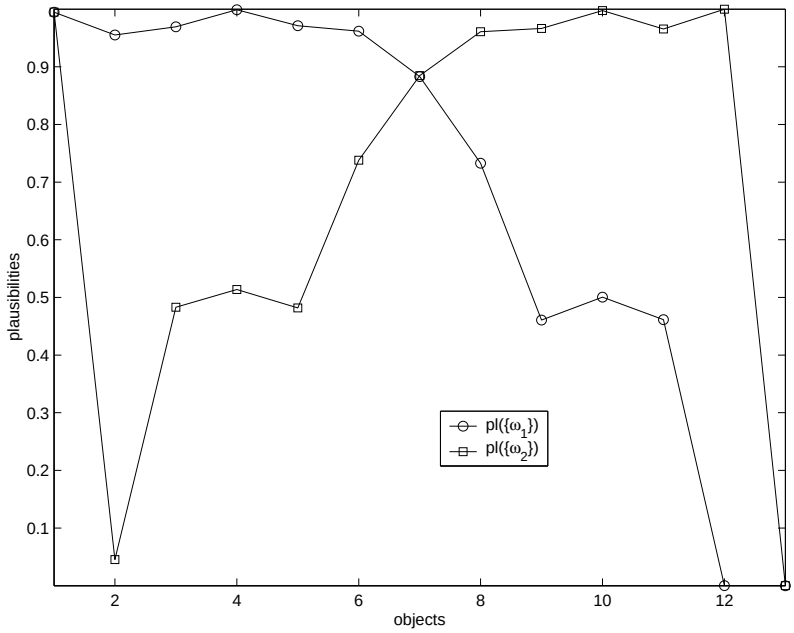
plausibilities
objects
pl({ω1})
pl({ω2})

Fig. 4.

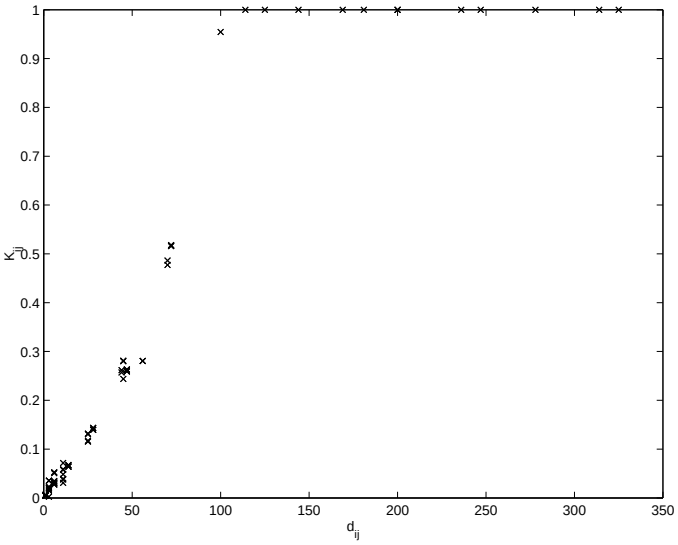
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
50
100
150
200
250
300
350
κ_{ij}
d_{ij}

Fig. 5.

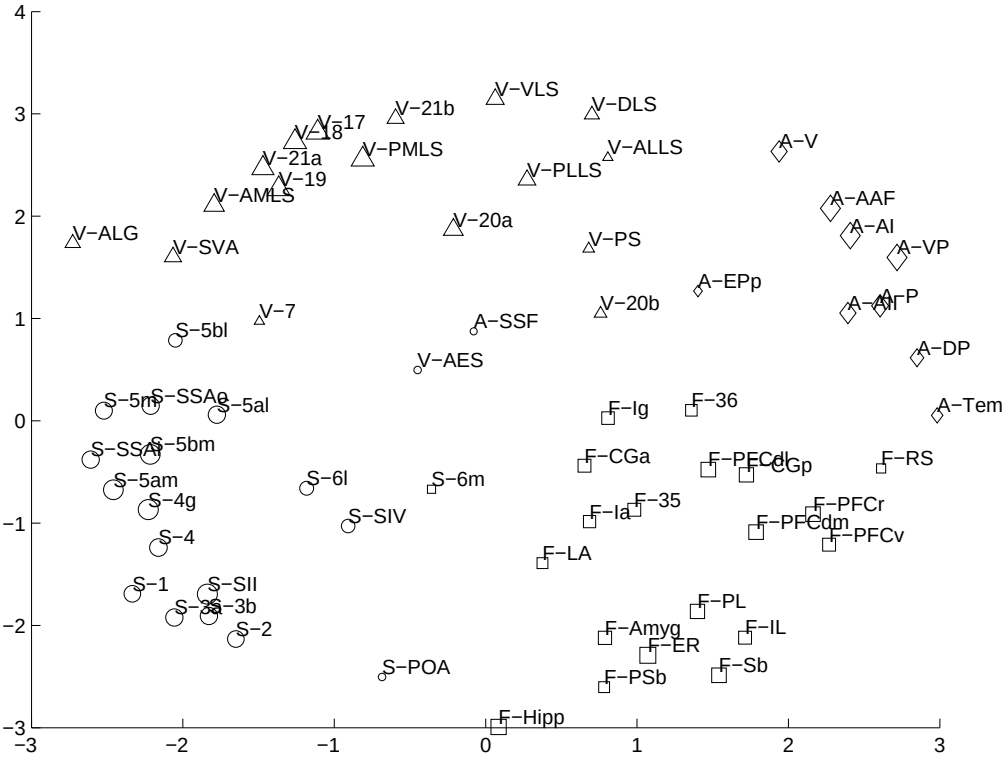
V−VLS
V−DLS
V−21b
V−17
V−18
V−21a
V−PMLS
V−ALLS
A−V
V−19
V−PLLS
V−AMLS
A−AAF
V−20a
A−AI
V−ALG
V−PS
V−SVA
A−VP
A−EPp
V−20b
V−7
A−SSF
S−5bl
A−DP
V−AES
S−5m
S−SSAe
S−5al
F−Ig
F−36
A−Tem
S−SSAi
S−5bm
F−CGa
F−PFCdl
F−CGp
F−RS
S−5am
S−6l
S−6m
S−4g
F−35
F−PFCr
S−SIV
F−Ia
F−PFCdm
S−4
F−PFCv
F−LA
S−1
S−SII
F−PL
S−3a
S−3b
F−Amyg
F−IL
S−2
F−ER
S−POA
F−Sb
F−PSb
F−Hipp

Fig. 6.

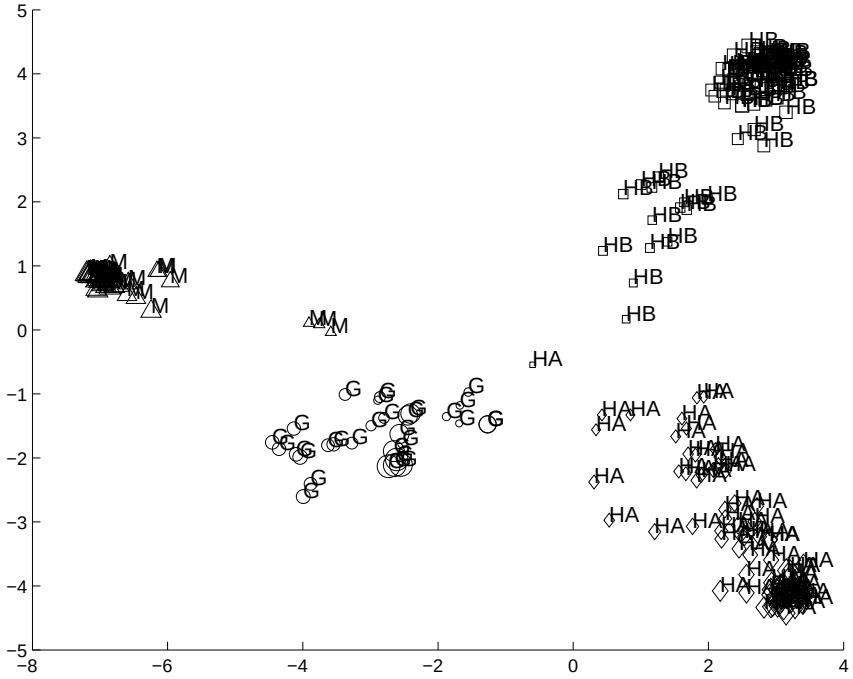

Fig. 7.

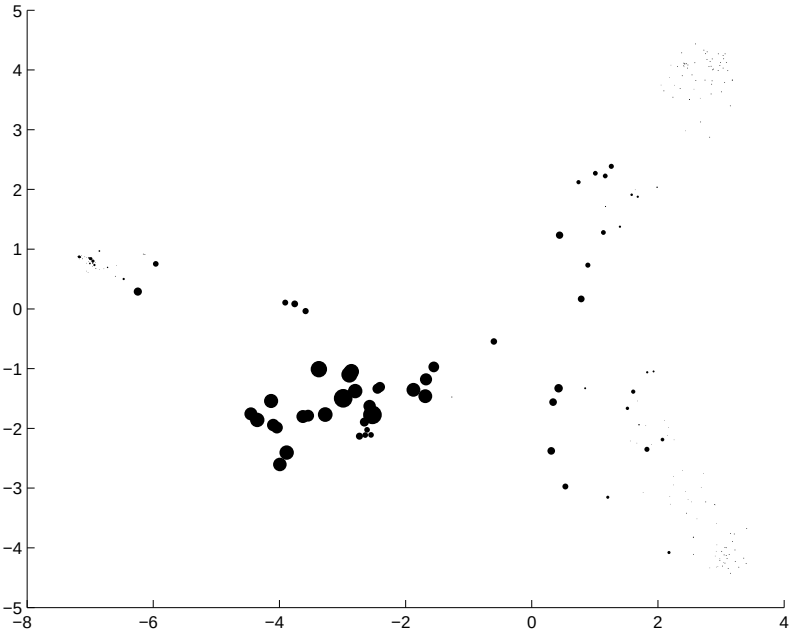

Fig. 8.

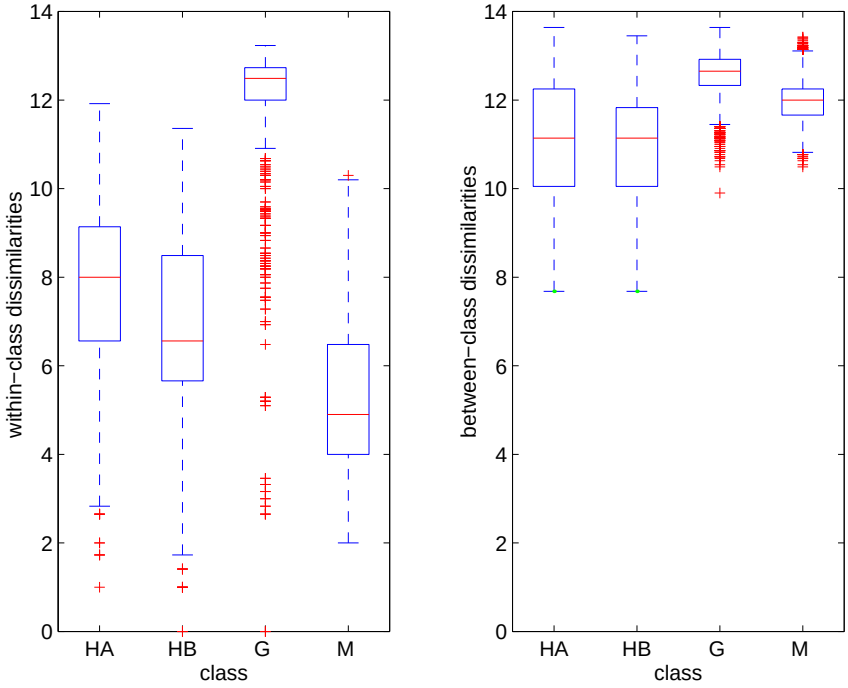
within−class dissimilarities
between−class dissimilarities
HA
HB
G
M
class
0
2
4
6
8
10
12
14

Fig. 9.

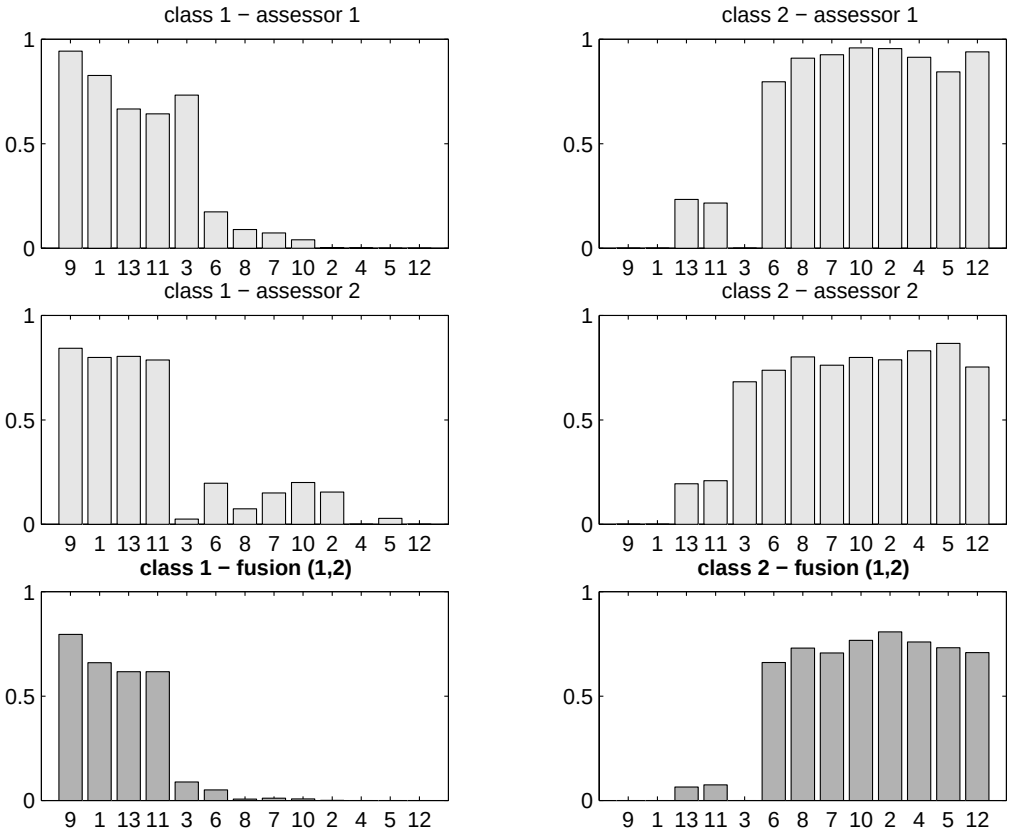
class 1 − assessor 1
class 2 − assessor 1
class 1 − assessor 2
class 2 − assessor 2
class 1 − fusion (1,2)
class 2 − fusion (1,2)
9 1 13 11 3 6 8 7 10 2 4 5 12
0
0.5
1

Fig. 10.

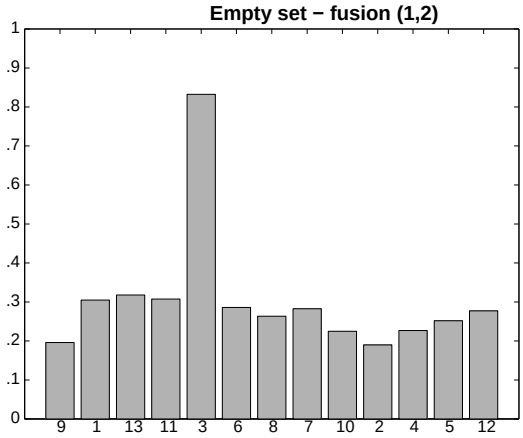
Empty set − fusion (1,2)
1
.9
.8
.7
.6
.5
.4
.3
.2
.1
0
9 1 13 11 3 6 8 7 10 2 4 5 12

Fig. 11.

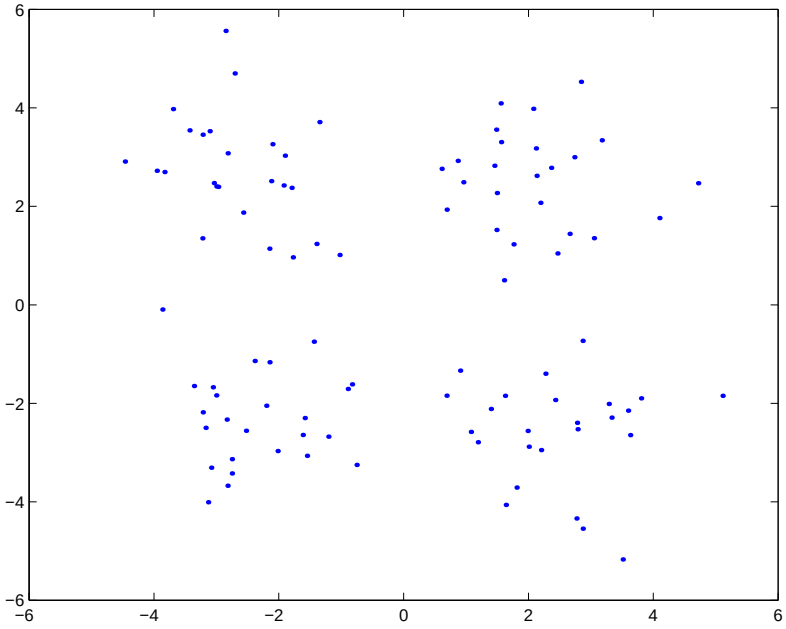
6
4
2
0
−2
−4
−6
−6 −4 −2 0 2 4 6

Fig. 12.

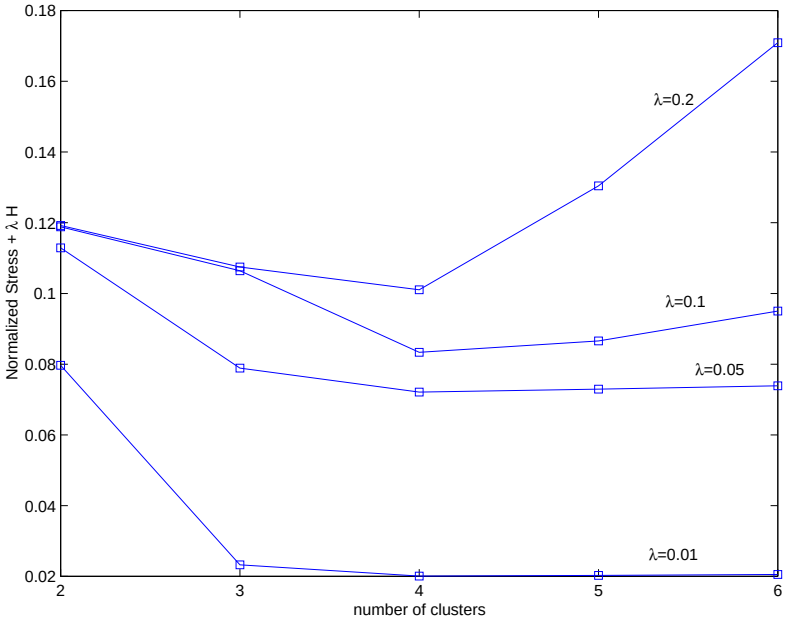
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
2
3
4
5
6
Normalized Stress + λ H
number of clusters
λ=0.2
λ=0.1
λ=0.05
λ=0.01

Fig. 13.

Table captions

- Table I: Credal partition of Example 1
- Table II: Plausibilities of the singletons for the credal partition of Example 1
- Table III: Dissimilarity matrix of Example 2
- Table IV: Degrees of conflict between the bba's of Example 1
- Table V: Summary of the EVCLUS method
- Table VI: Pignistic probabilities for the credal partition of Example 1
- Table VII: Synthetic data set
- Table VII: Bba's obtained from judge 1
- Table IX: Bba's obtained from judge 2
- Table X: Bba's obtained from judge 2 after permuting clusters 1 and 2, and expanding the frame $\Omega = \{\omega_1, \omega_2\}$ to $\Theta = \{\omega_1, \omega_2, \omega_3\}$
- Table XI: Bba's obtained after conjunctive fusion

TABLE I

F	$m_1(F)$	$m_2(F)$	$m_3(F)$	$m_4(F)$	$m_5(F)$
$\emptyset$	0	0	0	0	0
$\{\omega_1\}$	0	0	0	0.2	0
$\{\omega_2\}$	0	1	0	0.4	0
$\{\omega_1, \omega_2\}$	0.7	0	0	0	0
$\{\omega_3\}$	0	0	0.2	0.4	0
$\{\omega_1, \omega_3\}$	0	0	0.5	0	0
$\{\omega_2, \omega_3\}$	0	0	0	0	0
Ω	0.3	0	0.3	0	1

TABLE II

k	$\mathrm{pl}_1(\{\omega_k\})$	$\mathrm{pl}_2(\{\omega_k\})$	$\mathrm{pl}_3(\{\omega_k\})$	$\mathrm{pl}_4(\{\omega_k\})$	$\mathrm{pl}_5(\{\omega_k\})$
1	1	0	0.8	0.2	1
2	1	1	0.3	0.4	1
3	0.3	0	1	0.4	1

TABLE III

d_{ij}	1	2	3	4	5
1	0	7	50	68	12
2	7	0	100	94	8
3	50	100	0	76	4
4	68	94	76	0	4
5	12	8	4	4	0

TABLE IV

K_{ij}	1	2	3	4	5
1	-	0	0.14	0.28	0
2	0	-	0.70	0.60	0
3	0.14	0.70	-	0.32	0
4	0.28	0.60	0.32	-	0
5	0	0	0	0	-

TABLE V

Store	dissimilarity matrix D
Pick	number of clusters c penalization coefficient λ set of focal elements $\mathcal{F}$ ($f = \lvert\mathcal{F}\rvert$)
Initialize	credal partition $M \equiv (\alpha_{ik}), i = 1, \ldots, n,\ k = 1, \ldots, f$ and coefficients a, b: $\alpha_{ik} \sim \mathcal{N}(0, \sigma^2)$ ($\sigma = 0.1$) $a \leftarrow \dfrac{\sum_{i<j} d_{ij}}{\sum_{i<j} K_{ij}}$ $b \leftarrow 0$
Minimize	stress function $J(M, a, b)$ defined by eq. (38)-(43) with respect to a, b and the α_{ik}, using the procedure described in Appendix II

TABLE VI

k	$p_1(\omega_k)$	$p_2(\omega_k)$	$p_3(\omega_k)$	$p_4(\omega_k)$	$p_5(\omega_k)$
1	0.45	0	0.35	0.2	1/3
2	0.45	1	0.1	0.4	1/3
3	0.1	0	0.55	0.4	1/3

TABLE VII

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0	1	1	1	1	1	1	1	1	1	1	1	200
2	1	0	6	3	6	11	25	44	72	70	72	100	325
3	1	6	0	3	11	6	14	28	56	47	45	72	247
4	1	3	3	0	3	3	11	25	47	45	47	70	278
5	1	6	11	3	0	6	14	28	45	47	56	72	314
6	1	11	6	3	6	0	3	11	28	25	28	44	236
7	1	25	14	11	14	3	0	3	14	11	14	25	200
8	1	44	28	25	28	11	3	0	6	3	6	11	169
9	1	72	56	47	45	28	14	6	0	3	11	6	181
10	1	70	47	45	47	25	11	3	3	0	3	3	144
11	1	72	45	47	56	28	14	6	11	3	0	6	114
12	1	100	72	70	72	44	25	11	6	3	6	0	125
13	200	325	247	278	314	236	200	169	181	144	114	125	0

TABLE VIII

i	$m_i^{(1)}(\{\omega_1\})$	$m_i^{(1)}(\{\omega_2\})$	$m_i^{(1)}(\Omega)$	$m_i(\emptyset)$
1	0.83	0.00	0.00	0.17
2	0.00	0.96	0.00	0.04
3	0.73	0.00	0.00	0.26
4	0.00	0.91	0.00	0.08
5	0.00	0.84	0.00	0.16
6	0.17	0.80	0.03	0.00
7	0.07	0.93	0.00	0.00
8	0.090	0.91	0.00	0.00
9	0.94	0.00	0.00	0.06
10	0.04	0.96	0.00	0.00
11	0.64	0.22	0.14	0.00
12	0.00	0.94	0.00	0.06
13	0.67	0.23	0.10	0.00

TABLE IX

i	$m_i^{(2)}(\{\omega_1\})$	$m_i^{(2)}(\{\omega_2\})$	$m_i^{(2)}(\Omega)$	$m_i^{(2)}(\emptyset)$
1	0.00	0.80	0.00	0.20
2	0.79	0.15	0.06	0.00
3	0.68	0.02	0.10	0.20
4	0.83	0.00	0.00	0.17
5	0.87	0.03	0.00	0.10
6	0.74	0.20	0.07	0.00
7	0.76	0.15	0.00	0.09
8	0.80	0.07	0.00	0.12
9	0.00	0.84	0.00	0.16
10	0.80	0.20	0.00	0.00
11	0.21	0.79	0.00	0.00
12	0.75	0.00	0.00	0.25
13	0.19	0.80	0.00	0.00

TABLE X

i	$m_i^{(2)*}(\{\omega_1\})$	$m_i^{(2)*}(\{\omega_2\})$	$m_i^{(2)*}(\{\omega_3\})$	$m_i^{(2)*}(\Theta)$	$m_i^{(2*)}(\emptyset)$
1	0.80	0.00	0.20	0.00	0
2	0.15	0.79	0.00	0.06	0
3	0.02	0.68	0.20	0.10	0
4	0.00	0.83	0.17	0.00	0
5	0.03	0.87	0.10	0.00	0
6	0.20	0.74	0.00	0.07	0
7	0.15	0.76	0.09	0.00	0
8	0.07	0.80	0.12	0.00	0
9	0.84	0.00	0.16	0.00	0
10	0.20	0.80	0.00	0.00	0
11	0.21	0.79	0.00	0.00	0
12	0.00	0.75	0.25	0.00	0
13	0.80	0.19	0.00	0.00	0

TABLE XI

i	$m^{(1,2)}(\{\omega_1\})$	$m^{(1,2)}(\{\omega_2\})$	$m^{(1,2)}(\{\omega_3\})$	$m^{(1,2)}(\Theta)$	$m^{(1,2)}(\emptyset)$
1	0.66	0.00	0.03	0.00	0.30
2	0.00	0.81	0.00	0.00	0.19
3	0.09	0.00	0.08	0.00	0.83
4	0.00	0.76	0.01	0.00	0.23
5	0.00	0.73	0.02	0.00	0.25
6	0.05	0.66	0.00	0.00	0.29
7	0.01	0.71	0.00	0.00	0.28
8	0.01	0.73	0.00	0.00	0.26
9	0.80	0.00	0.01	0.00	0.20
10	0.09	0.77	0.00	0.00	0.22
11	0.62	0.07	0.00	0.00	0.31
12	0.00	0.71	0.01	0.00	0.28
13	0.62	0.07	0.00	0.00	0.32

Vitae

T. Denœux graduated in 1985 as an engineer from the Ecole Nationale des Ponts et Chaussées in Paris, and received a doctorate from the same institution in 1989. He is currently a Full Professor with the Department of Information Processing Engineering at the Université de Technologie de Compiègne, France. His current research interests concern fuzzy data analysis, belief functions theory and, more generally, the management of imprecision and uncertainty in data analysis, pattern recognition and information fusion.

Marie-Hélène Masson received the Engineer degree in Computer Science and a PhD from the Université de Technologie de Compiègne. She has been an assistant professor at the Université de Picardie Jules Verne (IUT de l'Oise) and a member of the Heudiasyc laboratory (UMR CNRS 6599) since 1993. Her research interests include statistical pattern recognition and data analysis.