Clustering of proximity data using belief functions

Thierry Denœux
HEUDIASYC – UMR CNRS 6599
Univ. de Technologie de Compiègne
BP 20529 - F-60205 Compiègne cedex
Thierry.Denoeux@hds.utc.fr

Mylène Masson
HEUDIASYC – UMR CNRS 6599
Univ. de Picardie Jules Verne
IUT de l'Oise
BP 20529 - F-60205 Compiègne cedex
Mylene.Masson@hds.utc.fr

Abstract

A new clustering method for relational data is proposed, based on Evidence theory. In this approach, masses of belief assigned to subsets of classes are used to compute the plausibility that two objects belong to the same class. It is then required that these plausibilities be compatible with the observed dissimilarities between objects. Experiments illustrate the ability of the method to handle noisy or non Euclidean data.

Keywords: Relational clustering, belief functions, outliers, non Euclidean data.

1 Introduction

Whereas evidence theory has been applied to supervised classification problems for a long time (see, e.g., [4]), the work presented in this paper is, to our knowledge, the first incursion of belief functions into the cluster analysis domain. Cluster analysis is concerned with methods for finding groups in data, groups (or classes) being defined as subsets of more or less "similar" objects [11]. 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. A quite extensive review of crisp and fuzzy relational clustering models can by found in [1, 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 $U = (u_{ik})$ of size $n \times c$ such that:

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

and

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

Each number $u_{ik} \in [0,1]$ is interpreted as a *degree of membership* of object i to cluster k.

Examples of such methods are the fuzzy non metric (FNM) model [16], the assignment-prototype (AP) model [20] and the relational fuzzy c-means (RFCM) model [9] (a similar approach may be found in [12]). The latter approach was later extended by Hathaway and Bezdek [8] 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 [3].

In this paper, a novel approach to clustering proximity data is presented, based on Dempster-Shafer (DS) theory of belief functions, also referred to as "Evidence theory". In this approach, the allocation of objects to classes is performed using the concept of basic belief asignment (bba), whereby a "mass of belief" is assigned to each possible subset of classes. Using a suitable noninteractivity assumption, 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. The rest of this paper is organized as follows. The necessary background on belief functions will be recalled in Section 2. Our method will then be exposed in Section 3, and experimental results will be presented in Section 4. Section 5 will conclude the paper.

2 Evidence theory

Let us consider a variable x taking values in a finite and unordered set Ω. Partial knowledge regarding the actual value taken by x can be represented by a *basic belief assignment* (bba) [18, 19], defined as a function m from 2^Ω to $[0,1]$, verifying:

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

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 [18], 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 [19]. 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) $\mathrm{bel} : 2^\Omega \mapsto [0,1]$, defined as

$$\mathrm{bel}(A) \triangleq \sum_{\emptyset \neq B \subseteq A} m(B) \quad \forall A \subseteq \Omega\ , \qquad (2)$$

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

$$\mathrm{pl}(A) \triangleq \mathrm{bel}(\Omega) - \mathrm{bel}(\overline{A}) \quad \forall A \subseteq \Omega\ , \qquad (3)$$

where $\overline{A}$ denotes the complement of A. Whereas $\mathrm{bel}(A)$ represents the amount of support given to A, the *potential* amount of support that *could be* given to A is measured by $\mathrm{pl}(A)$. 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 $\cap$ defined as:

$$(m_1 \cap m_2)(A) \triangleq \sum_{B \cap C = A} m_1(B) m_2(C)\ , \qquad (4)$$

for all $A \subseteq \Omega$. The quantity $K = (m_1 \cap m_2)(\emptyset)$ 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 $(m_1 \cap m_2)(A)$ by $1 - K$. The resulting operation is noted $\oplus$ and is called Dempster's rule of combination [18]:

$$(m_1 \oplus m_2)(A) \triangleq \frac{1}{1-K} \sum_{B \cap C = A} m_1(B) m_2(C)\ . \qquad (5)$$

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^{Ω_1} on Ω_1 is defined for all $A \subseteq \Omega_1$ as

$$m^{\Omega_1}(A) \triangleq \sum_{\{B \subseteq \Omega \,\mid\, \mathrm{Proj}(B \downarrow \Omega_1) = A\}} m^\Omega(B), \qquad (6)$$

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

$$\mathrm{Proj}(B \downarrow \Omega_1) \triangleq \{\omega_1 \in \Omega_1 \mid \exists \omega_2 \in \Omega_2, (\omega_1, \omega_2) \in B\} . \quad (7)$$

The two marginal bba's m^{Ω_1} and m^{Ω_2} are said to be noninteractive iff for all $A \subseteq \Omega_1$ and for all $B \subseteq \Omega_2$

$$m^{\Omega}(A \times B) = m^{\Omega_1}(A) m^{\Omega_2}(B) . \quad (8)$$

These definitions can be easily extended to bba's defined over the Cartesian product of n sets $\Omega_1, \ldots, \Omega_n$.

3 The method

3.1 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 $m_i(\Omega)$ 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 Ω; this corresponds to a situation of complete knowledge;
- when each m_i is a *Bayesian* bba, then M specifies a fuzzy partition of Ω, as defined by Bezdek [1].

A credal c-partition (or partition of size c) will be defined as a credal partition $M = (m_1, \ldots, m_n)$ such that, for all $\omega \in \Omega$, we have

$$\mathrm{pl}_i(\{\omega\}) > 0$$

for some $i \in \{1, \ldots, n\}$, pl_i being the plausibility function associated to m_i.

Example 1 Let us consider a collection O of $n = 4$ objects and $c = 3$ classes. A credal partition M of O is given in Table 1. The class of object o_2 is known with certainty, whereas the class of o_4 is completely unknown. The two other cases correspond to situations of partial knowledge. The plausibilities $\mathrm{pl}_i(\{\omega\})$ of each singleton are given in Table 2. 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 2 defines a possibilistic partition as defined in [1].

Table 1: Credal partition of Example 1

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

Table 2: Plausibilities of the singletons for the credal partition of Example 1

i	$\mathrm{pl}_1(\{\omega_i\})$	$\mathrm{pl}_2(\{\omega_i\})$	$\mathrm{pl}_3(\{\omega_i\})$	$\mathrm{pl}_4(\{\omega_i\})$
1	1	0	0.8	1
2	1	1	0.3	1
3	0.3	0	1	1

3.2 Compatibility of an evidential partition with a dissimilarity matrix

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.

It is reasonable to assume that two similar objects are more likely to be in the same class, than two dissimilar ones. The more similar, the more *plausible* it is that they belong to the same group. To formalize this idea, we need to calculate 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 dissimililarity 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 to consider the joint bba $m_{i\times j}$ on Ω^2 related to the vector variable (y_i, y_j). If m_i and m_j are assumed to be noninteractive, then $m_{i\times j}$ is completely determined by m_i and m_j, and we have $\forall A, B \subseteq \Omega$:

$$m_{i\times j}(A \times B) = m_i(A) m_j(B). \qquad (9)$$

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

$$\begin{aligned} \mathrm{pl}_{i\times j}(S) &= \sum_{(A\times B)\cap S \neq \emptyset} m_{i\times j}(A \times B) \\ &= \sum_{A\cap B \neq \emptyset} m_i(A) m_j(B) \\ &= 1 - \sum_{A\cap B = \emptyset} m_i(A) m_j(B) \\ &= 1 - K_{ij} \ , \end{aligned} \qquad (10)$$

where K_{ij} is the degree of conflict between m_i and m_j.

Hence, the plausibility that objects o_i and o_j belong to the same class is 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:

$$d_{ij} > d_{i'j'} \Rightarrow \mathrm{pl}_{i\times j}(S) \leq \mathrm{pl}_{i'\times j'}(S) \qquad (11)$$

or, equivalently:

$$d_{ij} > d_{i'j'} \Rightarrow K_{ij} \geq K_{i'j'} \ , \qquad (12)$$

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.

3.3 Learning a credal partition from data

To extract a credal partition from dissimilarity data, 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 [2]. The purpose of MDS methods 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 as a point in a 2^c-dimensional space. Hence, the concept of "credal partition" 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. This close connection allows us to transpose MDS algorithms to our problem.

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 simplest one is obtained by imposing a linear relationship between "distances" (i.e., degrees of conflict in our case) and dissimilarities, which is referred to as *metric* MDS. The stress function used in our case is:

$$\sigma(M, a, b) \triangleq \frac{\sum_{i<j}(aK_{ij} + b - d_{ij})^2}{\sum_{i<j} d_{ij}^2} , \quad (13)$$

where a and b are two coefficients, and the denominator is a normalizing constant. This stress function can be minimized iteratively with respect to M, a and b using a gradient-based procedure. Note that this method is invariant under any affine transformation of the dissimilarities.

Remark 1 Each bba m_i must satisfy Eq. (1). Hence, the optimization of σ with respect to M is a constrained optimization problem. However, the contraints vanish if one uses the following parameterization:

$$m_i(A_l) = \frac{\exp(\alpha_{il})}{\sum_{k=1}^{2^c} \exp(\alpha_{ik})}, \quad (14)$$

where $A_l, l = 1, \ldots, 2^c$ are the subsets of Ω, and the α_{il} for $i = 1, \ldots, n$ and $l = 1, \ldots, 2^c$ are $n2^c$ real parameters.

3.4 Controlling the number of parameters

An important issue 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. If c is large, the number of free parameters 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.

Another very efficient means of reducing the number of free parameters is to add a penalization term to the stress function. This approach does not reduce the number of parameters to be optimized but limits the *effective* number of parameters of the method. It is thus a way to control the complexity of the classification model. In our case, we would like to extract as much information as possible from the data, so that 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 [14, 13], and it is still, to some extent, an open question. However, several entropy measures have been proposed. The total uncertainty introduced by Pal et al. [15] satisfies natural requirements and has interesting properties. It is defined, for a normal bba m, as:

$$H(m) \triangleq \sum_{A \in \mathcal{F}(m)} m(A) \log_2 \left(\frac{|A|}{m(A)} \right) , \quad (15)$$

where $\mathcal{F}(m)$ denotes the set of focal sets of m. $H(m)$ is minimized when the mass is assigned to few focal sets, with small cardinality (it is proved in [15] that $H(m) = 0$ iff $m(\{\omega\}) = 1$ for some $\omega \in \Omega$).

To apply (15) to a subnormal bba m (i.e., such that $m(\emptyset) > 0$), some normalization has to be performed. Two common normalization procedures are Dempster's normalization (in which the mass given to $\emptyset$ is deleted and all other belief masses are divided by $1 - m(\emptyset)$ [18], and Yager's normalization, in which the mass $m(\emptyset)$ is transferred to Ω [21]. The latter approach has been preferred in our approach, because it allows to penalize subnormal bba's more efficiently. The expression of total uncertainty for a subnormal bba m then becomes:

$$\begin{aligned} H(m) &= \sum_{A \in \mathcal{F}(m) \setminus \{\emptyset\}} m(A) \log_2 \left(\frac{|A|}{m(A)} \right) \\ &+ m(\emptyset) \log_2 \left(\frac{|\Omega|}{m(\emptyset)} \right). \end{aligned} \quad (16)$$

Finally, the objective function to be minimized is:

$$J(M, a, b) \triangleq \sigma(M, a, b) + \lambda \sum_{i=1}^{n} H(m_i). \quad (17)$$

3.5 From credal clustering to fuzzy or hard clustering

Although we believe that a lot of information may be gained in analyzing a credal partition, it is always possible to transform it into a fuzzy or hard partition. This conversion is based on the concept of *pignistic* probability [19] defined, for a normalized bba m, by:

$$BetP(A) \triangleq \sum_{\emptyset \neq B \subseteq \Omega} m(B) \frac{|A \cap B|}{|B|} \quad (18)$$

To obtain a fuzzy partition, one calculates the pignistic probability of each singleton ω_k. In the case where these singletons, Ω and the empty set are the only focal sets of the bba, the expression of the pignistic probabilities is given by:

$$BetP(\{\omega_k\}) = m(\{\omega_k\}) + \frac{m(\Omega) + m(\emptyset)}{c}, \quad (19)$$

for all $k = 1, c$ (we assume that Yager's normalization is used). A hard partition can then be easily obtained from the values of pignistic probabilities. In this sense, a credal partition may be viewed as a general model of partionning, including fuzzy and hard partitions.

4 Results

4.1 Synthetic dataset

This first example is inspired from a classical dataset [20]. A (13×13) dissimilarity matrix was generated by computing the squared Euclidean distances of a two dimensional object dataset represented in figure 1. The $13th$ object, an outlier, is useful to study the robustness of the method. The first object is assumed to be close to all other objets, and is not represented in this figure. The dissimilarity between this point and objects 2 to 12 is arbitrarily set to 1 and to 200 with the 13th object. This object is intended to reflect either noisy, unreliable data, or imprecise evaluations coming from subjective assessments. We compare the results obtained with our method and five classical clustering methods based on relationnal data: Windham's assignment-prototype algorithm (AP) [20], the Fuzzy Non Metric algorithm (FNM) [16], the Relational Fuzzy c-means algorithm (RFCM) [9], and its "Noise" version (NRFCM) [3], and the non-Euclidean RFCM algorithm (NERF) [8]. NRFCM, by using a "noise" cluster, is well-adapted to datasets containing noise and outliers, whereas NERF is intended to cope with non-Euclidean dissimilarities. The task is to find a reasonable 2-partition of object 2 to 12 and to detect the particularity of objects 1 and 13. The figure 2 shows the resulting fuzzy membership functions for the five classical algorithms, and the bba obtained with evidential clustering. Note that only 4 focal elements were considered: $\{\omega_1, \omega_2, \Omega, \emptyset\}$. As could be expected, among the five algorithms, only NRFCM is able to detect the outlier but the method fails with the first object (which is classified in class 2). The evidential clustering method (EVCLUS) provides a clear understanding of the data by allocating an important mass to the empty set for the outlier and to Ω for the first point.

4.2 "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 [17] and

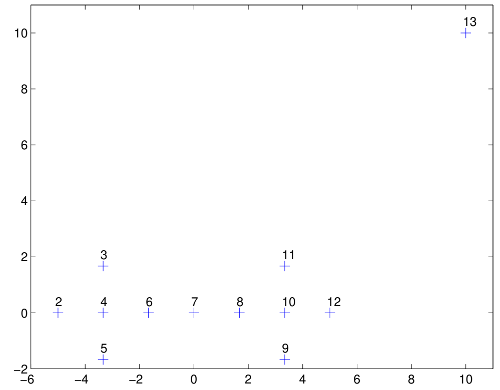

Figure 1: Synthetic dataset.

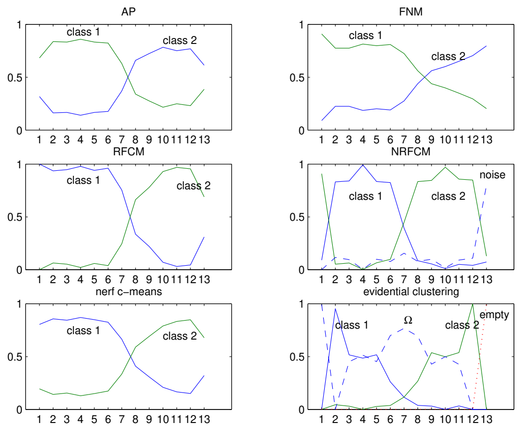

Figure 2: Synthetic dataset. Results of the six algorithms.

used by several authors to test visualization, discrimination or clustering algorithms based on proximity data [6, 7, 10]. The proximity values range from 0 (self-connection), to 4 (absent or unreported connection) with intermediate values : 1 (dense connection), 2 (intermediate connection) and 3 (weak connection). The cortex has been divided into four functional areas: auditory (A), visual (V), somatosensory (S), and frontolimbic (F). The clustering task is to partition the data into four clusters. Six focal elements were considered for applying the evidential clustering method: 4 singletons $\{\omega_i\}$ $(i = 1, 4)$, Ω and $\emptyset$. The five algorithms (AP, FNM, RFCM, NRFCM, NERF) converge toward a bad solution with equal membership of $1/c$ for all points. In order to provide a simple display of the results with EVCLUS, a two dimensional representation of the cortical areas has been

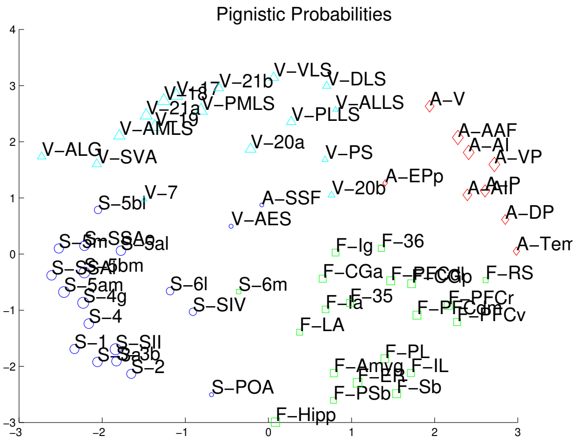

Figure 3: Cat cortex data set. Pignistic probabilities and hard assignments.

obtained from the proximity matrix using a classical MDS algorithm. The classification displayed on figure 3 is done according to the maximum of the pignistic probabilities. The clusters are represented by different symbols and the size of the symbols is proportional to the maximum of the pignistic probabilities. It can be seen that the four functional areas of the cortex are well-recovered. The error rate (only three points among 65 are misclassified), competes honourably with those reported in discrimination studies [6, 7].

5 Conclusion

In this paper we have suggested a new way of classifying relational data based on the theory of evidence. The classification task is performed in a very natural way, by only imposing that, the more two objects are similar, the more likely they belong to the same cluster. The concept of credal partition can be considered as a generalization of a probabilistic or possibilistic partition and offers a very flexible framework to handle noisy, imprecise or non-Euclidean data. Experiments on various datasets, which are not all reported here, have shown the efficiency of this approach.

Acknowledgements

The authors would like to thank Thore Graepel for providing the cat cortex dataset.

References

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

[2] I. Borg and P. Groenen. *Modern multidimensional scaling.* Springer, 1997.

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

[4] 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.

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

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

[7] T. Graepel, R. Herbrich, B. Schölkopf, A. Smola, P. Bartlett, K. Robert-Müller, K. Obermayer, and R. Williamson. Classification on Proximity Data with LP–Machines. in *Proceedings of ICANN'99*, 304–309, 1999.

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

[9] 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.

[10] T. Hofmann, and J. Buhmann. Multidimensional scaling and data clustering. in *Advances in Neural Information Processing Systems 7*, MIT Press, Cambridge, MA, 459–466, 1995.

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

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

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

[14] 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.

[15] 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.

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

[17] 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.

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

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

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

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