On evidential clustering with partial supervision

Violaine Antoine[1], Kévin Gravouil[1,2], and Nicolas Labroche[3]

[1] Clermont Auvergne University, UMR 6158, LIMOS, F-63000,
[2] Clermont Auvergne University, INRA, MEDIS, LMGE, F-63000,
Clermont-Ferrand, France
{violaine.antoine,kevin.gravouil}@uca.fr
[3] University of Tours, LIFAT , EA 6300, Blois, France
nicolas.labroche@univ-tours.fr

Abstract. This paper introduces a new semi-supervised evidential clustering algorithm. It considers label constraints and exploits the evidence theory to create a credal partition coherent with the background knowledge. The main characteristics of the new method is its ability to express the uncertainties of partial prior information by assigning each constrained object to a set of labels. It enriches previous existing algorithm that allows the preservation of the uncertainty in the constraint by adding the possibility to favor crisp decision following the inherent structure of the dataset. The advantages of the proposed approach are illustrated using both a synthetic dataset and a real genomics dataset.

Keywords: evidential clustering, partial labels, semi-supervised clustering, belief function.

1 Introduction

Evidential clustering algorithms, such as ECM [1], rely on the theoretical foundation of belief functions and evidence theory [2] and allow to express many types of uncertainty about the assignment of an object to a cluster. It enables to handle crisp single cluster assignment, as well as cluster membership degrees, total ignorance and outliers detection. The credal partition, which is formed with the assignments of all the objects, generalizes other soft partitions such as fuzzy, possibilistic or rough partitions [3].

Clustering is a complex unsupervised task that often requires additional assumptions to determine relevant solutions. The performances of a clustering algorithm can be highly improved by using background knowledge [4]. To this end, several semi-supervised evidential clustering approaches have been proposed [5–7]. In [7], the SECM-pl algorithm integrates prior information in the form of labeled data instances. The particularity of SECM-pl is its ability to handle partial knowledge, which corresponds to the uncertainty about the assignment of an object to several classes. This partial knowledge is controlled by the algorithm in such a way that the uncertainty can be preserved.

In this paper, we propose an approach that generalizes SECM-pl, which maintains a high flexibility on the constraints, by favoring a decision making on the

constraints. The paper is organized as follows: section 2 recalls the basics concerning the evidence theory and its application in clustering. Section 3 details the novel SECM algorithm and focuses on how labels constraints are expressed and incorporated in ECM. Section 4 presents experimental settings and results. Finally a discussion and future work are presented in section 5.

2 Preliminaries

2.1 Belief functions

The evidence theory (or belief functions theory) [2, 8] is a mathematical framework that enables to reflect the state of partial and unreliable knowledge. Let $\Omega = \{\omega_1, \ldots, \omega_c\}$ be the frame of discernment where ω_i is the true state of the system which will be defined below. The mass function $m : 2^\Omega \rightarrow [0,\ 1]$, also called basic belief assignment (bba), measures the degree of belief that ω_i belongs to a subset $A \subseteq \Omega$. It satisfies $\sum_{A \subseteq \Omega} m(A) = 1$. Any subset A such that $m(A) > 0$ is named a focal set of m. Given a mass function m, the plausibility function $pl : 2^\Omega \rightarrow [0,\ 1]$ is defined by:

$$pl(A) = \sum_{B \cap A \neq \emptyset} m(B), \qquad \forall A \subseteq \Omega. \tag{1}$$

The quantity $pl(A)$ corresponds to the maximal degree of belief that could be given to A. To make a decision, a mass function can be transformed into a pignistic probability distribution $BetP$ [8].

2.2 Evidential c-Means

Evidential clustering algorithms generate for each object $\mathbf{x}_1, \ldots \mathbf{x}_i, \ldots \mathbf{x}_n \in \mathbb{R}^p$ a mass function $\mathbf{m}_i$ on the set $\Omega = \{\omega_1, \ldots, \omega_c\}$ denoting the clusters. The collection $\mathbf{M} = (\mathbf{m}_1, \ldots, \mathbf{m}_n)$ forms the credal partition and allows to represent the uncertainties and imprecisions regarding the class membership of each object. ECM [1] is the credibilistic version of Fuzzy C-Means [9]. It considers for each subset $A_j \subseteq \Omega$ a representation of the subset with a prototype vector $\mathbf{v}_j$ in $\mathbb{R}^p$. The objective function is:

$$J_{ECM}(\mathbf{M}, \mathbf{V}) = \sum_{i=1}^{n} \sum_{A_j \subseteq \Omega, A_j \neq \emptyset} |A_j|^\alpha m_{ij}^\beta d_{ij}^2 + \sum_{i=1}^{n} \rho^2 m_{i\emptyset}^\beta, \tag{2}$$

where $\mathbf{V}$ is the collection of prototypes, $m_{ij} = m_i(A_j)$ corresponds to the bba of the object $\mathbf{x}_i$ for the subset A_j, $m_{i\emptyset}$ denotes the mass of $\mathbf{x}_i$ allocated to the empty set and d_{ij}^2 represents the squared Euclidean distance between $\mathbf{x}_i$ and the prototype $\mathbf{v}_j$. The last term of the objective function enables to handle the empty set which can be interpreted as a cluster for outliers. The ρ parameter is a fixed coefficient representing the distance between any object and the empty set. The two parameters α and $\beta > 1$ are introduced to penalize the degree of

belief assigned to subsets with a high cardinality and to control the fuzziness of the partition. The objective function is subject to

$$\sum_{j/A_j \subseteq \Omega, A_j \neq \emptyset} m_{ik} + m_{i\emptyset} = 1; \quad m_{ij} \geq 0 \quad \forall i = \{1, \dots n\}, \forall j / A_j \subseteq \Omega. \tag{3}$$

2.3 SECM-pl

The main idea of the algorithm [7] is to add a penalty term in the objective function of ECM, in order to take into account a set of already labeled objects. Any mass function which partially or fully respects a constraint on a label ω_k has a high plausibility $pl(\omega_k)$ given to the label. Similarly, an object constrained in several classes, i.e. on the set $A_j \subset \Omega$ is respected with mass functions given a high plausibility $pl(A_j)$. Thus, the following penalty term has been proposed:

$$J_S = \sum_{i=1}^{n} \sum_{A_j \subset \Omega, A_j \neq \emptyset} b_{ij}(1 - Pl_i(A_j)), \tag{4}$$

where $b_{ij} = 1$ if $\mathbf{x}_i$ is constrained on A_j and 0 otherwise.

3 New ECM algorithm with partial supervision

3.1 Modeling the constraints

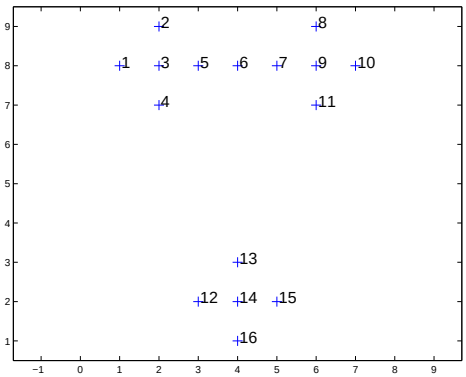
Let us consider a set of partially labeled constraints, i.e. a collection of objects $\mathbf{x}_i$ such that $\mathbf{x}_i \in A_j$, $\forall A_j \neq \emptyset$. If A_j is a singleton, then the object i belongs to a class with certainty. Otherwise, $\mathbf{x}_i$ belongs to a class listed in A_j without knowing which one more precisely. Notice that $\mathbf{x}_i \in \Omega$ corresponds to complete ignorance concerning the class of the object i. Degrees of belief containing the set of clusters A_j or a part of it should be favored as well as mass functions of subsets with a low cardinality. Thus, we define the measure $1 \geq T_{ij} \geq 0$ by the following formula:

$$T_{ij} = T_i(A_j) = \sum_{A_j \cap A_l \neq \emptyset} \frac{|A_j \cap A_l|^{\frac{r}{2}}}{|A_l|^r} m_{il}, \quad \forall i \in \{1 \dots n\}, A_j \subseteq \Omega, \tag{5}$$

where $r \geq 0$ controls a degree of penalization of the subsets. The coefficient $|A_l|^r$ is used to penalize subsets with a high cardinality and $|A_j \cap A_l|^{\frac{r}{2}}$ allows to concentrate efforts on subsets containing mostly elements of A_j. Notice that when $r = 0$, T_{ij} corresponds to the plausibility that the object $\mathbf{x}_i$ belongs to A_j. For the rest of the paper, we set $r = 1$.

3.2 Illustration

The behavior of the new measure T_{ij} is illustrated with the DiamondK3 dataset presented Figure 1(a). This dataset is composed of 15 objects that should be separated into 3 groups. As it can be observed, points 13 to 16 are well isolated, whereas objects 1 to 11 seem to correspond to two natural clusters connected by the object 6. Let us suppose that some partial knowledge is available: e.g. object 6 is in the cluster ω_1 and object 13 belongs either to ω_1 or to ω_3, but not to ω_2. Thus, we obtain the two following constraints: $\mathbf{x}_6 \in \{\omega_1\}$ and $\mathbf{x}_{13} \in \{\omega_1, \omega_3\}$.



(a)

$m_i(A_j)$	$\emptyset$	0	0	0	0	0
	ω_1	**1**	0	0	0	0
	ω_2	0	0	0	0	**1**
	$\{\omega_1, \omega_2\}$	0	0	**1**	0	0
	ω_3	0	0	0	0	0
	$\{\omega_1, \omega_3\}$	0	**1**	0	0	0
	$\{\omega_2, \omega_3\}$	0	0	0	0	0
	Ω	0	0	0	**1**	0
$T_i(A_j)$	ω_1	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{3}$	0
	$\{\omega_1, \omega_3\}$	1	$\frac{\sqrt{2}}{2}$	$\frac{1}{2}$	$\frac{1}{3}$	0
$Pl(A_j)$	ω_1	1	1	1	1	0
	$\{\omega_1, \omega_3\}$	1	1	1	1	0

(b)

Fig. 1. DiamondK3 dataset (a) and illustration of the proposed penalty term $T_i(A_j)$ when considering several possible mass functions and compared to penalty term based on plausibility $Pl(A_j)$ for previous SECM-pl [7] (b).

Figure 1(b) presents in each column a set of possible mass functions for an object $\mathbf{x}_i$ coming from the DiamondK3 dataset. First, let us consider that $\mathbf{x}_i = \mathbf{x}_6$ and let us assume that $m_6(\omega_1) = 1$ as shown in the first column of Figure 1(b). Thus, the constraint is respected and it can be observed that $T_6(\omega_1) = 1$. Inversely, if $m_6(\omega_2) = 1$ as presented in the last column of Figure 1(b), then the constraint is totally neglected and $T_6(\omega_1) = 0$. Other columns illustrate partial respect of the constraint, since the bba is allocated to subsets containing the label ω_1. The larger the cardinality of the subset, the lower the value of T_{ij}.

Let us assume that $\mathbf{x}_i = \mathbf{x}_{13}$ and let us focus on the value obtained by $T_i(\{\omega_1, \omega_3\})$ for the set of possible mass functions. As it can be observed, $T_{ij} = 0$ when no focal sets contain ω_1 and/or ω_3. Conversely, if there exists a degree of belief not null on a subset including at least one of the classes included in the constraint, then $T_{ij} > 0$. As previously, the larger the cardinality of the subset, the lower the value of T_{ij}. For the same amount of subsets, for example columns 2 and 3 in Figure 1(b), a higher value is given to subsets containing the most of classes in the constraint, i.e. $\{\omega_1, \omega_3\}$. This is a significant difference with the plausibility measure for which all subsets intersecting with the constraints contribute equally to the final value.

3.3 Objective function and optimization

Based on the mass function m_i of an object i, we can quantify the degree to which a partial constraint $\mathbf{x}_i \in A_j$ is respected by computing T_{ij} in equation (5). $T_{ij} = 1$ when the belief is given to a cluster in A_j and is 0 when the belief is assigned to none of the clusters included in A_j, i.e. when the constraint is not respected. If we consider now that the bbas have to be found, a natural requirement is to obtain a value of T_{ij} as high as possible if there exists a constraint such that $\mathbf{x}_i \in A_j$. This goal is achieved by minimizing the following objective function:

$$J_{SECM}(M,V) = (1-\gamma)\frac{1}{2^c n} J_{ECM}(M,V) + \gamma \frac{1}{s} \sum_{i=1}^{n} \sum_{A_j \subset \Omega, A_j \neq \emptyset} b_{ij}(1 - T_{ij}), \quad (6)$$

such that constraints (3) are respected, s corresponds to the number of constraints, and $b_{ij} = 1$ if $\mathbf{x}_i \in A_j$, i.e if the object i is constrained with A_j and 0 otherwise.

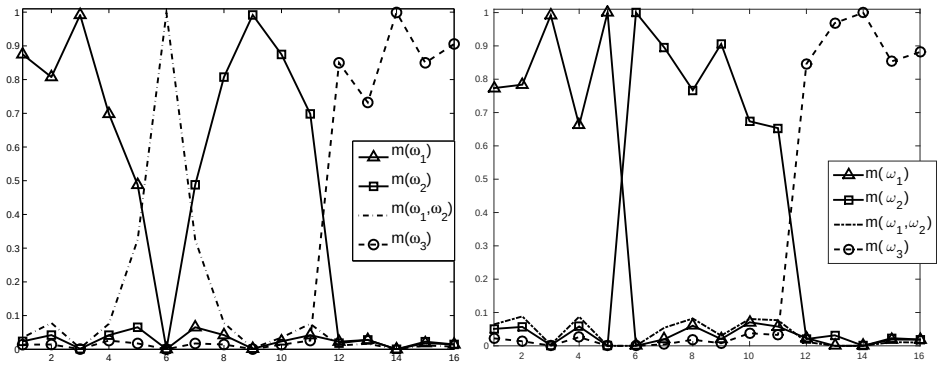
The coefficient γ controls the tradeoff between the objective function of ECM and the constraints. Notice that if $r = 0$ for the computation of T_{ij}, then J_{SECM} is identical to the objective function proposed in [7]. Such setting allows the penalty term to give equal importance to any subset intersecting with the constraints, whereas $r > 0$ favors subsets with low cardinality. As ECM, the credal partitioning is carried out through an iterative optimization of the objective function, with the update of the mass functions and the prototypes. If β is set to 2, then the problem becomes quadratic with linear constraints and can be resolved with classical methods, for instance [10].

4 Experimentations

4.1 Toy example

To illustrate the behavior of the SECM algorithm, we used the DiamondK3 dataset. First, an execution of ECM is performed with $\alpha = 1$, $\beta = 2$, $\rho^2 = 10^3$ and the final mass functions for the most representative subsets varying with the objects number are presented Figure 2(a). It can be seen that ECM identifies the 3 clusters by assigning the belief to the 3 singletons. The object 6, which is located between the cluster ω_1 and ω_2, is ambiguous as it can belong to either ω_1 or ω_2. Thus, ECM assigns for $\mathbf{x}_6$ a high mass for the subset $\{\omega_1, \omega_2\}$.

Let us consider now that the following set of constraints are available: $\mathbf{x}_5 \in \{\omega_1\}$, $\mathbf{x}_6 \in \{\omega_2\}$ and $\mathbf{x}_{13} \in \{\omega_1, \omega_2\}$. The SECM algorithm is executed with $\gamma = 0.5$ and the credal partition obtained is presented Figure 2(b). As it can be observed, constraints are well respected. The object 6, previously ambiguous with the ECM algorithm, is now assigned with certainty to ω_2. Similarly, the object 5 had with ECM its belief divided into $\{\omega_1, \omega_2\}$ and ω_1, whereas now all its belief is given to $\{\omega_1\}$. Finally, the mass function $m_{13}(\omega_3)$ for the object



(a) (b)

Fig. 2. Credal partitions obtained for DiamondK3 with (a) ECM and (b) SECM such that $\mathbf{x}_6 \in \{\omega_1\}$ and $\mathbf{x}_5 \in \{\omega_2\}$ and $\mathbf{x}_{13} \in \{\omega_1, \omega_3\}$.

13, which is already high with ECM, has increased with SECM. It shows that SECM is able to constrained $\mathbf{x}_{13}$ more specifically on ω_3 following the inherent structure of the dataset.

4.2 Genomics application

Dataset: Dozens of thousands microorganism's genomes are available in public databases. We selected three known genomes from the RefSeq database [11], namely *Clostridium acetoburylicum*, *Bacillus cereus* and *Brachyspira hyodysenteriae*, to simulate a small microbial community. DNA sequences were extracted from these genomes then embedded in numerical vectors using normalized tetranucleotide frequencies with a CONCOCT-inspired approach [12]. The final dataset, called tetragen, is composed of 22 attributes and 1188 objects corresponding to DNA sequences. Classes, i.e. the genomes *B. hyodysenteriae*, *C. acetoburylicum* and *B. cereus* contain respectively 288, 383 and 517 instances. In order to obtain the tetragen dataset, the largest DNA sequences were divided in several objects. We took benefit of this process to create label constraints: we assigned two DNA sequences composed of 13 and 21 objects in the subsets $\{B.\ cereus\}$ and $\{B.\ cereus, B.\ hyodysenteriae\}$ respectively. As a consequence, we obtained a dataset composed of 2.9% of constrained objects. Figure 3 presents the class and prior information used for the tetragen dataset.

Experimental protocol: For both ECM and SECM, we performed 10 executions with random initialization of the centroids and kept the credal partition giving the minimum value for the objective function. To synthesize the information provided by the partitions, we transformed them into hard credal partitions by assigning each object to the subset of classes with the highest mass. Figures 4(a) and (b) illustrates the obtained results. As it can be observed, constraints helped SECM to impact the boundary of ω_3.

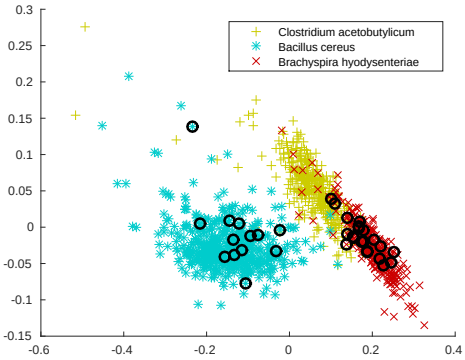


Fig. 3. Real classes (color) and constrained objects (encircled) for the tetragen data set.

In order to compare the methods, partitions obtained with ECM and SECM were transformed into hard partitions by selecting the cluster with the maximal pignistic probability. Then, their agreement with the real partition were measured with the Adjusted Rand Index (ARI) [13] and the Normalized Mutual Information (NMI). Both of them provide a 1 value when the partitions totally match.With ECM, we obtained ARI=0.75 and NMI=0.71 whereas SECM gives an ARI=0.78 and a NMI=0.73. It shows that a few number of constrained objects, even partially labeled, can lead our clustering algorithm to a better result than ECM.

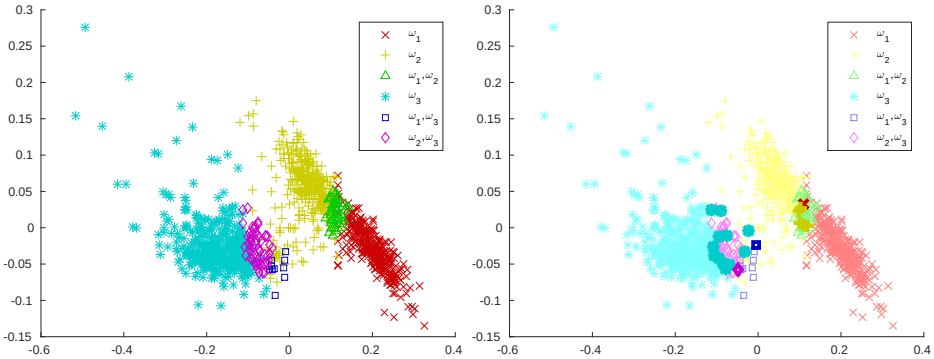


Fig. 4. Hard credal partition obtained with (a) ECM and (b) SECM for tetragen. Colors are lightened in (b) for objects for which the assignment has not changed between the two algorithms.

5 Conclusion

In this paper, a new semi-supervised clustering algorithm called SECM is proposed. It generalizes previous approach [7] based on partial label constraints. The new penalty term can be parameterized to favor either any credal partition

for which constraints are still plausible or only credal partitions for which constrained objects have belief on subsets with low cardinalities. A proof of concept is provided and shows the benefits of the new algorithm. Finally, a real test is performed on genomics data set and shows the necessity of such expressive approaches in real use case.

In the future, extensive tests on real and synthetic datasets should be conducted in order to show the influence of the parameter r and to compare various semi-supervised clustering algorithms. The genomics use case should also be developed as it offers a relevant testbed for partial user knowledge integration. A further work is to scale SECM for larger datasets, in order to apply the algorithm in a real genomics application.

References

1. Masson, M.H., Denœux, T.: ECM: An evidential version of the fuzzy c-means algorithm. Pattern Recognition **41**(4) (2008) 1384–1397
2. Shafer, G.: A Mathematical Theory of Evidence. Princeton Univ. Press. Princeton, NJ (1976)
3. Denœux, T., Kanjanatarakul, O.: Beyond fuzzy, possibilistic and rough: an investigation of belief functions in clustering. In: Soft Methods for Data Science. Springer (2017) 157–164
4. Basu, S., Davidson, I., Wagstaff, K.: Constrained clustering: Advances in algorithms, theory, and applications. CRC Press (2008)
5. Antoine, V., Quost, B., Masson, M.H., Denœux, T.: CECM: Constrained Evidential C-Means algorithm. Computational Statistics and Data Analysis **56** (2012) 894–914
6. Antoine, V., Quost, B., Masson, M.H., Denœux, T.: Evidential clustering with instance-level constraints for proximity data. Soft Computing **18**(7) (July 2014) 1321–1335
7. Antoine, V., Labroche, N., Vu, V.V.: Evidential seed-based semi-supervised clustering. In: Soft Computing and Intelligent Systems (SCIS), Kitakyushu, Japan, IEEE (December 2014) 706–711
8. Smets, P., Kennes, R.: The transferable belief model. Artificial Intelligence **66** (1994) 191–234
9. Bezdek, J.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenium Press, New-York (1981)
10. Ye, Y., Tse, E.: An extension of Karmarkar's projective algorithm for convex quadratic programming. Mathematical Programming **44**(1) (1989) 157–179
11. Pruitt, K., Tatusova, T., Maglott, D.: Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research **35**(suppl_1) (2006) D61–D65
12. Alneberg, J., Bjarnason, B., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U., Lahti, L., Loman, N., Andersson, A., Quince, C.: Binning metagenomic contigs by coverage and composition. Nature methods **11**(11) (2014) 1144
13. Hubert, L., Arabie, P.: Comparing partitions. Journal of classification **2**(1) (1985) 193–218