Towards the efficient estimation of ECM parameters

Ekaterina Zagainaia

LIMOS
Under the supervision of:
Violaine Antoine, Engelbert Mephu Nguifo

Content

- Introduction
- The methods of clusterization
- Evidential c-means
- Intelligent ECM
- Criterias for determining k
- Conclusion

Introduction

- ECM (Evidential c-means) is a variant of *k*-means that generates a *credal partition*.

- Such partition, based on the theory of *belief function*, enables to handle ambiguous objects and outliers by assigning a degree of belief on subsets of clusters.

- It brings richer information about the class membership of an object than hard, fuzzy or possibilistic partition unfortunately at the cost of a higher complexity.

- Similarly to *k*-means, the number of clusters c has to be supplied by an expert.

Introduction

- This paper proposes a new method to efficiently determine c and the associated subsets.

- Promising results have been obtained on experimental toy data.

The methods of clusterization

- Data clustering is one of the most popular method in data analysis. It enables to assign objects into groups of similar objects.

- There exists a wide range of algorithms able to perform this task:
  - ***k*-means algorithm**
  - **Intelligent *k*-means** (IK-means) was proposed to select the correct number of clusters in *k*-means. IK-means is fast and deterministic, but it may drastically overestimate the number of clusters.

- The *k*-means algorithm is a method that generate a *crisp partition*. In practice, there always exists outliers and objects located between two or more classes. A crisp partition is not suitable in these type of situations.

Evidential c-means

Introduction

- The theory of *belief functions*, and particularly the notion of *credal partition* enables us to represent the partial knowledge about objects.

- It allows to represent a wide range of situations concerning the class membership of an object.

- Algorithms returning a credal partition are referred to as *evidential clustering algorithms* (for example, ECM: Evidential c-means).

- At the same time these algorithms imply a higher complexity with the respect to c.

Evidential c-means

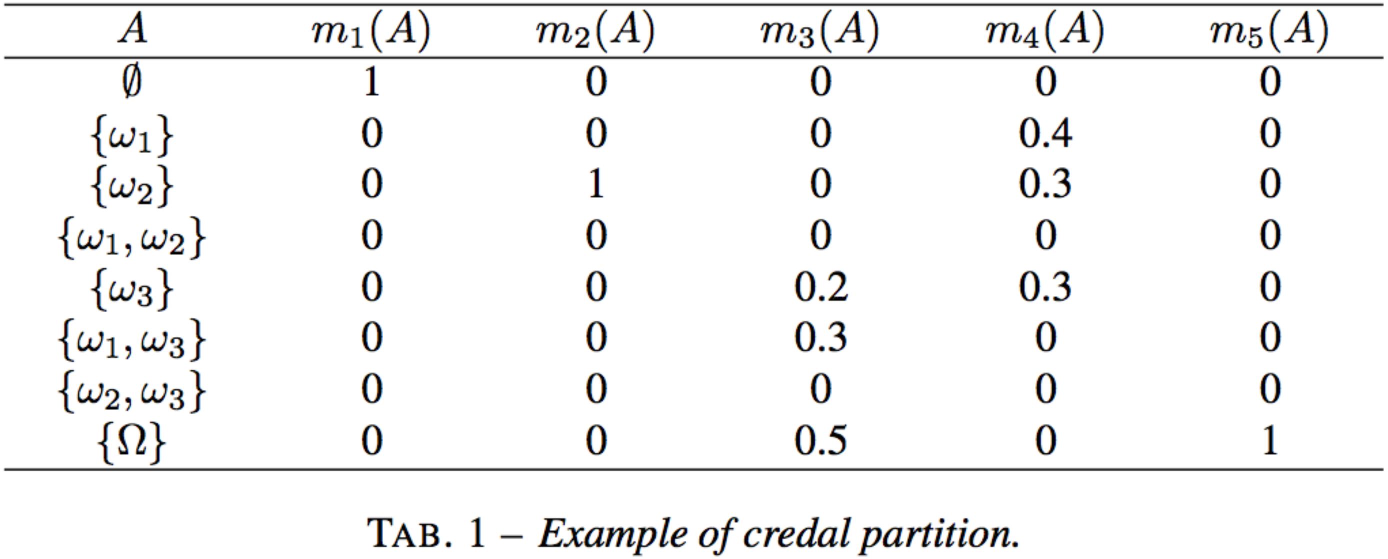
Introduction

- The theory of *belief functions* is a theoretical framework for dealing with unreliable and partial knowledge.
- Let Ω be a finite set called frame of discernment.
- A *belief assignment* (bba), defined as a mass function m : $2^{\Omega} \rightarrow [0, 1]$ represents partial knowledge regarding the actual value taken by a variable y.
- This mass function corresponds to: $\sum_{A \subseteq \Omega} m(A) = 1$
- $\Omega = \{\omega_1, . . . , \omega_c\}$ is the set of classes and y corresponds to the real class taken by an object.
- A credal partition is the concatenation of the bbas of each object.

Evidential c-means

Example

ECM makes possible to model all situations from full certainty to complete ignorance concerning the class of every object.



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

TAB. 1 – *Example of credal partition.*

The hard credal partition is obtained using the rule of maximum on the bbas.

Evidential c-means

Summary

- ECM is a variant of *k*-means that generates a credal partition instead of a crisp partition.
- It allows a better modeling and a more detailed description of complex data (for example, in the domain of medicine).
- The method has a *linear complexity* with the respect to the number of objects and the number of attributes.
- And it has an *exponential complexity* with the respect to the number of clusters.
- If c is the number of clusters, there exists 2^c subsets and as many values to find for a bba associated to an object.
- The computation time is then mainly slowed down by the number of subsets.

Evidential c-means

Previous development

Proposals:

- Masson and Denœux (2008) suggest to reduce the number of subsets to Ω and those having a cardinality less or equal to two.
- Also they propose to automatically find the number of clusters by computing a validity index for different values of c.

Shortcomings:

- The limitation to specific subsets is arbitrary.
- The above method to choose c is slow with the respect to the time, since it implies to run several times ECM.

Intelligent ECM

- In this work we propose to automatically find the number of clusters and the most important subsets before running ECM, in the manner of IK-means with *k*-means.

- *I* - the set of current objects
- A - the set of ambiguous objects
- The centroid of the all dataset is referred to as g and ω_g corresponds to its associated cluster.
- For a new cluster ω_t, we define S_t the set of objects in $\{\omega_t\}$ and A_t the set of objects in $\{\omega_g, \omega_t\}$.

Intelligent ECM

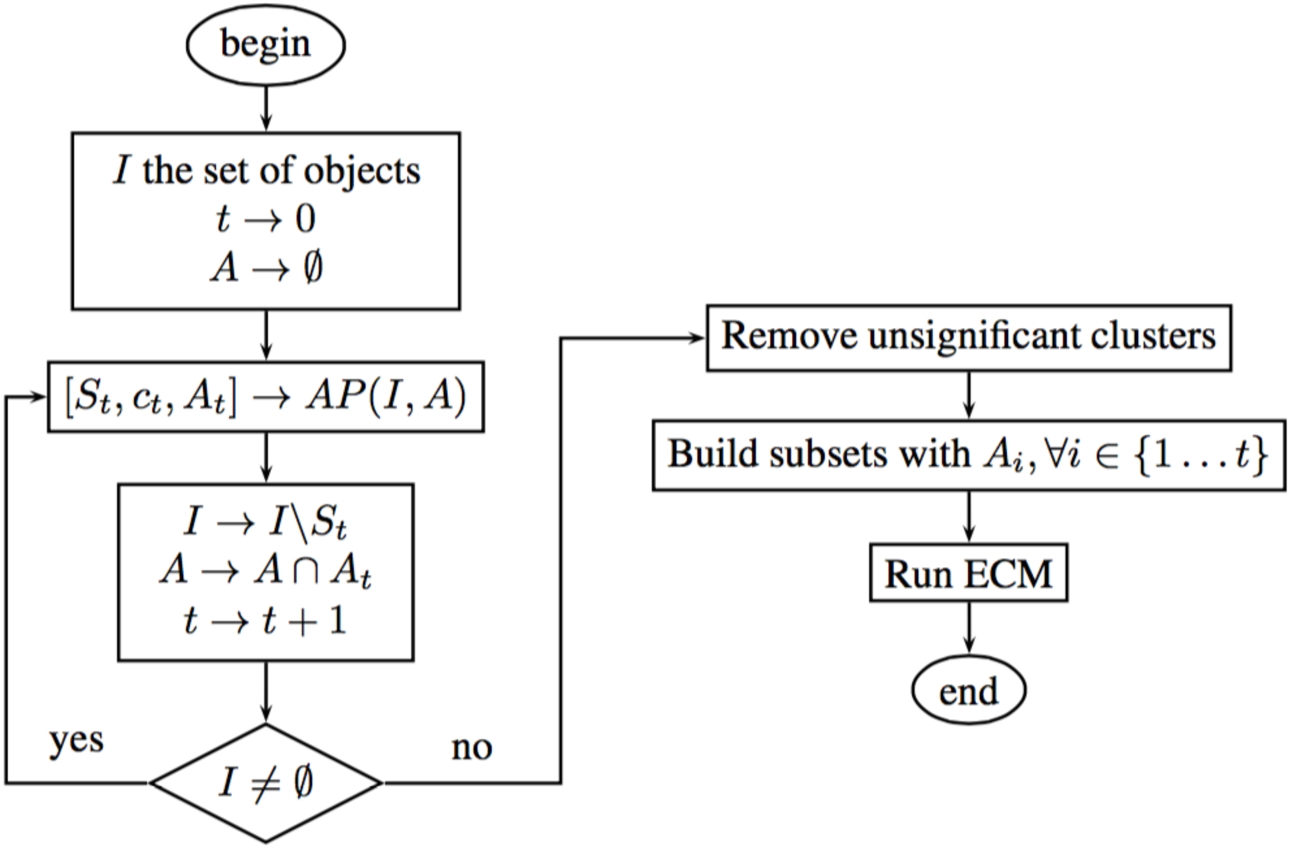
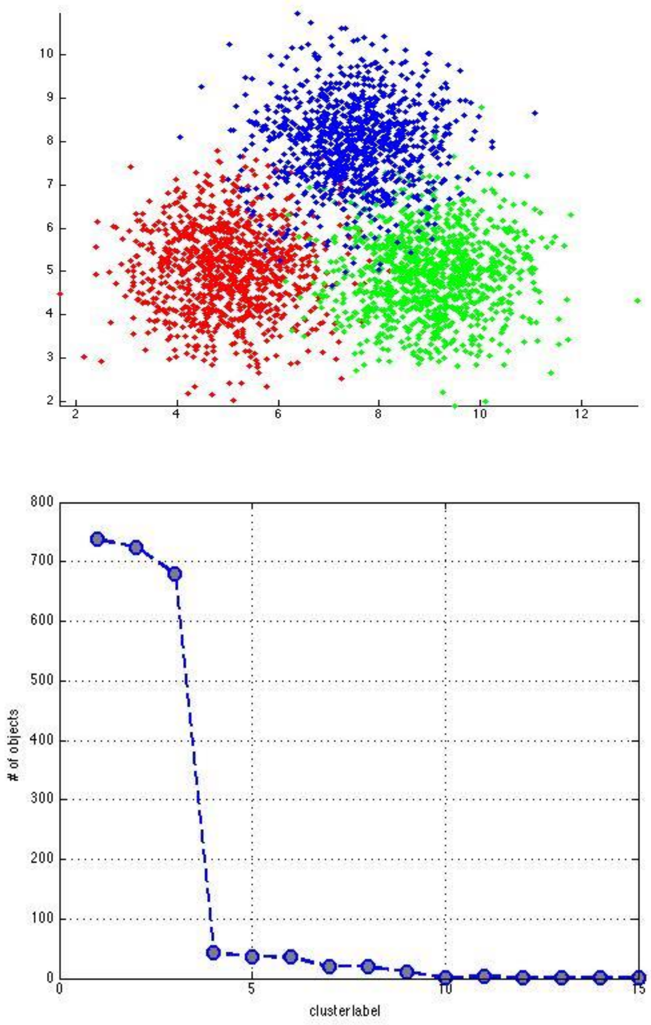

begin
I the set of objects
$t \rightarrow 0$
$A \rightarrow \emptyset$
$[S_t, c_t, A_t] \rightarrow AP(I, A)$
$I \rightarrow I \backslash S_t$
$A \rightarrow A \cap A_t$
$t \rightarrow t + 1$
$I \neq \emptyset$
yes
no
Remove unsignificant clusters
Build subsets with $A_i, \forall i \in \{1 \ldots t\}$
Run ECM
end

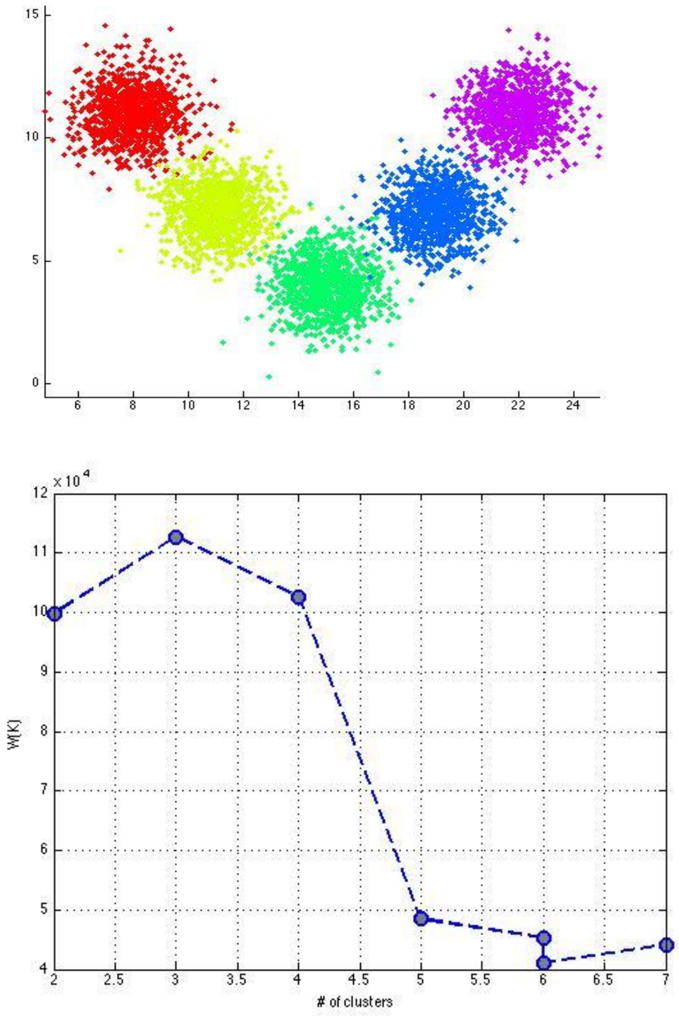
FIG. 1 – *Intelligent ECM.*

Criterias for determining k

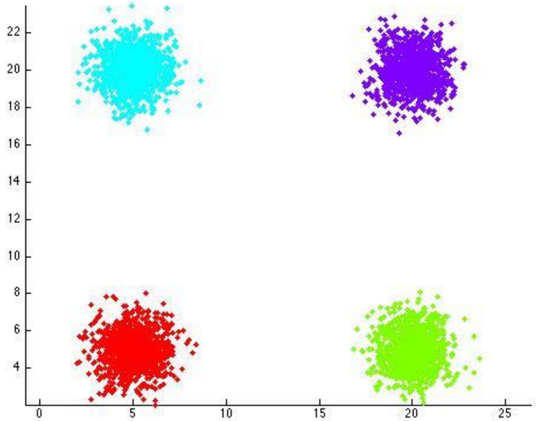
Removing insignificant clusters

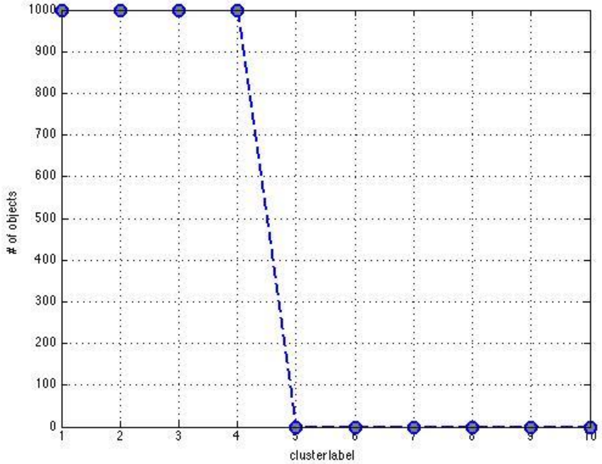


The within-cluster dispersion



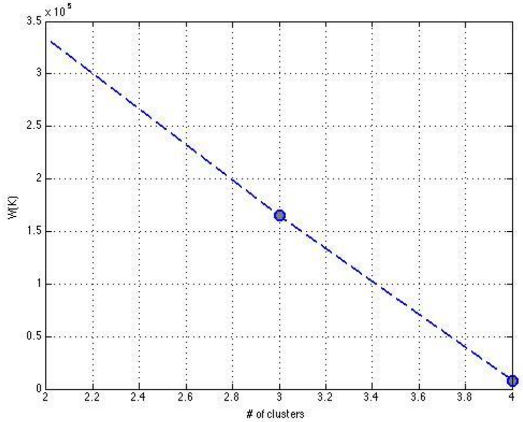
Criterias for determining k





Removing insignificant clusters

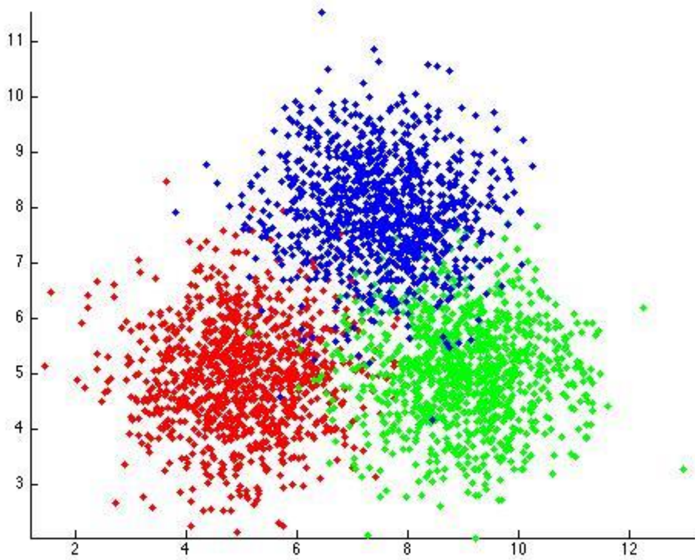
# of clusters	2	3	**4**	5 (4)	6 (4)	7 (4)…
W(K)	**1.0e+05 *** 3.3351	1.6478	0.0789	0.0789	0.0789	0.0789
H(K)	**1.0e+04 *** 0.4093	7.9483	**0**	0	0	-



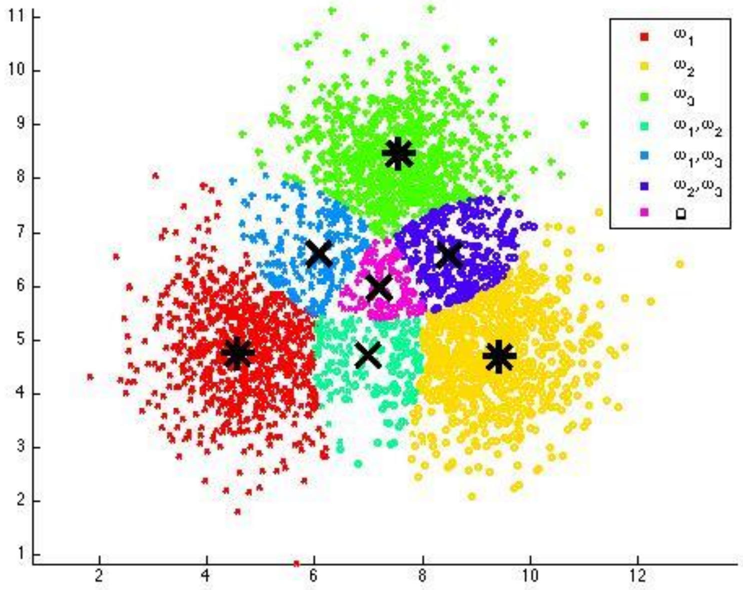
Criterias for determining k

- When the value of c has been defined, subsets associated to the remaining clusters are selected using the set of ambiguous objects A.
- Finally, a normal execution of ECM is carried out with c and the selected subsets.
- Result of experiment:

Data set



Result of clusterization



Conclusion

- We have developed a new method called Intelligent ECM to estimate the parameters needed for the ECM algorithm.

- Adopting such method makes it possible to avoid arbitrary choice of subsets.

- In addition, the new algorithm, choosing in a fast and smart way the optimal number of clusters for overlapping data sets, is proposed.

- The within-cluster dispersion can be used to determine the number of clusters for any type of the data, but with higher time cost.

- The proposed algorithm, Intelligent ECM can be applied on larger dataset than the classical ECM method.

Conclusion

- Future work consists in analyzing the behavior of Intelligent ECM on various datasets.

- In addition, several adjustments of the method have to be explored in order to make it more robust.

- For example, the determination of the objects belonging to a cluster or a subset can be modified.