Belief Hierarchical Clustering

Wiem Maalel, Kuang Zhou, Arnaud Martin and Zied Elouedi

Abstract In the data mining field many clustering methods have been proposed, yet standard versions do not take base on uncertain databases. This paper deals with a new approach to cluster uncertain data by using a hierarchical clustering defined within the belief function framework. The main objective of the belief hierarchical clustering is to allow an object to belong to one or several clusters. To each belonging, a degree of belief is associated, and clusters are combined based on the pignistic properties. Experiments with real uncertain data show that our proposed method can be considered as a propitious tool.

1 Introduction

Due to the increase of imperfect data, the process of decision making is becoming harder. In order to face this, the data analysis is being applied in various fields.

Clustering is mostly used in data mining and aims at grouping a set of similar objects into clusters. In this context, many clustering algorithms exist and are categorized into two main families. The first family involves the partitioning methods based on density such as k-means algorithm [6] that is widely used thanks to its convergence speed. It partitions the data into k clusters represented by their centers. The second family includes the hierarchical clustering methods such as the top-down and

Wiem Maalel
LARODEC, Université de Tunis, ISG Tunis, e-mail: wiem.maalel@gmail.com

Kuang Zhou
School of Automation, Northwestern Polytechnical University, China; IRISA, Université de Rennes 1, France, e-mail: kzhoumath@163.com

Arnaud Martin
IRISA, Université de Rennes 1, France, e-mail: Arnaud.Martin@univ-rennes1.fr

Zied Elouedi
LARODEC, Université de Tunis, ISG Tunis, e-mail: zied.elouedi@gmx.fr

the Hierarchical Ascendant Clustering (HAC) [5]. This latter consists on constructing clusters recursively by partitioning the objects in a bottom-up way. This process leads to good result visualizations. Nevertheless, it has a non-linear complexity.

All these standard methods deal with certain and precise data. Thus, in order to facilitate the decision making, it would be more appropriate to handle uncertain data. Here, we need a soft clustering process that will take into account the possibility that objects belong to more then one cluster.

In such a case, several methods have been established. Among them, the Fuzzy C-Means [1] which consists on assigning a membership to each data point corresponding to the cluster center, and the weights minimizing the total weighted mean-square error. This method constantly converges. Patently, Evidential c-Means (ECM) [3, 7] is deemed to be a very fateful method. It enhances the FCM and generates a credal partition from attribute data. This method deals with the clustering of object data. Accordingly, the belief k-Modes method [4] is a popular method, which builds K groups characterized by uncertain attribute values and provides a classification of new instances. Schubert has also found a clustering algorithm [8] which uses the mass on the empty set to build a classifier.

Our objective in this paper is to develop a belief hierarchical clustering method, in order to ensure the membership of objects in several clusters, and to handle the uncertainty in data under the belief function framework.

This reminder is organized as follows: in the next section we review the ascendant hierarchical clustering, its concepts and its characteristics. In section 3, we recall some of the basic concepts of belief function theory. Our method is described in section 4 and we evaluate its performance on a real data set in section 5. Finally, Section 6 is a conclusion for the whole paper.

2 Ascendant hierarchical clustering

This method consists on agglomerating the close clusters in order to have finally one cluster containing all the objects x_j (where $j = 1,..,N$).
Let's consider $\mathscr{P}^K = \{C_1,...,C_i\}$ the set of clusters. If $K = N$, $C_1 = x_1,...,C_N = x_N$. Thereafter, throughout all the step of clustering we will move from a partition $\mathscr{P}^K$ to a partition $\mathscr{P}^{K-1}$. The result generated is described by a hierarchical clustering tree (dendrogram), where the nodes represent the successive fusions and the height of the nodes represents the value of the distance between two objects which gives a concrete meaning to the level of nodes conscripted as "indexed hierarchy". This latter is usually indexed by the values of the distances (or dissimilarity) for each aggregation step. The indexed hierarchy can be seen as a set with an ultrametric distance d which satisfies these properties: *i)* $x = y \iff d(x,y) = 0$, *ii)* $d(x,y) = d(y,x)$, *iii)* $d(x,y) \leq d(x,z) + d(y,z), \forall x,y,z \in \mathbb{R}$. The algorithm is as follows:

- Initialisation: the initial clusters are the N-singletons. We compute their dissimilarity matrix.
- Iterate these two steps until the aggregation turns into a single cluster:

– Combine the two most similar (closest) elements (clusters) from the selected groups according to some distance rules.
– Update the matrix distance by replacing the two grouped elements by the new one and calculate its distance from each of the other classes.

Once all these steps completed, we do not recover a partition of K clusters, but a partition of $K-1$ clusters. Hence, we had to point out the aggregation criterion (distance rules) between two points and between two clusters. We can use the Euclidian distance between N objects x defined in a space $\mathbb{R}$. Different distances can be considered between two clusters: we can consider the minimum as follows:

$$\mathbf{d}(C_j^i, C_{j'}^i) = \min_{x_k \in C_j^i, x_{k'} \in C_{j'}^i} d(x_k, x_{k'}) \tag{1}$$

with $j, j' = 1, ..., i$. The maximum can also be considered, however, the minimum and maximum distances create compact clusters but sensitive to "outliers". The average can also be used, but the most used method is Ward's method, using Huygens formula to compute this:

$$\Delta I_{inter(C_j^i, C_{j'}^i)} = \frac{m_{C_j} m_{C_{j'}}}{m_{C_j} + m_{C_{j'}}} d^2(\mathbf{C_j^i}, \mathbf{C_{j'}^i}) \tag{2}$$

where m_{C_j} and $m_{C_{j'}}$ are numbers of elements of C_j and $C_{j'}$ respectively and $\mathbf{C_j^i}, \mathbf{C_{j'}^i}$ the centers. Then, we had to find the couple of clusters minimizing the distance:

$$(C_l^k, C_{l'}^k) = d(C_l^i, C_{l'}^i) = \min_{C_j^i, C_{l'}^i \in \mathscr{C}^i} \mathbf{d}(C_j^i, C_{j'}^i) \tag{3}$$

3 Basis on the theory of belief functions

In this Section, we briefly review the main concepts that will be used in our method that underlies the theory of belief functions [9] as interpreted in the Transferable Belief Model (TBM) [10]. Let's suppose that the frame of discernment is $\Omega = \{\omega_1, \omega_2, ..., \omega_3\}$. Ω is a finite set that reflects a state of partial knowledge that can be represented by a basis belief assignment defined as:

$$m : 2^\Omega \to [0,1]$$
$$\sum_{A \subseteq \Omega} m(A) = 1 \tag{4}$$

The value $m(A)$ is named a basic belief mass (bbm) of A. The subset $A \in 2^\Omega$ is called focal element if $m(A) > 0$. One of the important rule in the belief theory is the conjunctive rule which consists on combining two basic belief assignments m_1 and m_2 induced from two distinct and reliable information sources defined as:

$$m_1 \copyright m_2(C) = \sum_{A \cap B = C} m_1(A) \cdot m_2(B), \quad \forall C \subseteq \Omega \tag{5}$$

The Dempster rule is the normalized conjunctive rule:

$$m_1 \oplus m_2(C) = \frac{m_1 \copyright m_2(C)}{1 - m_1 \copyright m_2(\emptyset)}, \quad \forall C \subseteq \Omega \tag{6}$$

In order to ensure the decision making, beliefs are transformed into probability measures recorded BetP, and defined as follows:

$$\text{BetP}(A) = \sum_{B \subseteq \Omega} \frac{|A \cap B|}{|B|} \frac{m(B)}{(1 - m(\emptyset))}, \forall A \in \Omega \tag{7}$$

4 Belief hierarchical clustering

In order to set down a way to develop a belief hierarchical clustering, we choose to work on different levels: on one hand, the object level, on the other hand, the cluster level.

At the beginning, for N objects we have, the frame of discernment is $\Omega = \{x_1, ..., x_N\}$ and for each object belonging to one cluster, a degree of belief is assigned. Let $\mathscr{P}^N$ be the partition of N objects. Hence, we define a mass function for each object x_i, inspired from the k-nearest neighbors [2] method which is defined as follows:

$$\begin{aligned} m_i^{\Omega_i}(x_j) &= \alpha e^{-\gamma d^2(x_i, x_j)} \\ m(\Omega) &= 1 - \alpha e^{-\gamma d^2(x_i, x_j)} \end{aligned} \tag{8}$$

where $i \neq j$, α and β are two parameters we can optimize [11], d can be considered as the Euclidean distance, and the frame of discernment is given by $\Omega_i = \{x_1, ..., x_N\} \setminus \{x_i\}$. In order to move from the partition of N objects to a partition of $N - 1$ objects we have to find both nearest objects (x_i, x_j) to form a cluster. Eventually, the partition of $N - 1$ clusters will be given by $\mathscr{P}^{N-1} = \{(x_i, x_j), x_k\}$ where $k = 1, ..., N \setminus \{i, j\}$. The nearest objects are found considering the pignistic probability, defined on the frame Ω_i, of each object x_i:

$$(x_i, x_j) = \underset{x_i, x_j \in \mathscr{P}^N}{\arg \max} \text{BetP}^{\Omega_i}(x_j) \tag{9}$$

Then, this first couple of objects is a cluster. Now consider that we have a partition $\mathscr{P}^K$ of K clusters $\{C_1, ..., C_K\}$. In order to find the best partition $\mathscr{P}^{K-1}$ of $K - 1$ clusters, we have to find the best couple of clusters to be merged. First, if we consider one of the classical distances **d**, presented in section 2, between the clusters, we delineate a mass function, defined within the frame Ω_i for each cluster $C_i \in \mathscr{P}^K$ with $C_i \neq C_j$ by:

$$m^{\Omega_i}(C_j) = \alpha e^{-\gamma d^2(C_i, C_j)} \qquad (10)$$

$$m^{\Omega_i}(\Omega_i) = 1 - \alpha e^{-\gamma d^2(C_i, C_j)} \qquad (11)$$

where $\Omega_i = \{C_1, \ldots, C_K\} \setminus \{C_i\}$. Then, both clusters to merge are given by:

$$(C_i, C_j) = \underset{C_i, C_j \in \mathscr{P}^K}{\arg\max} \, \text{BetP}^{\Omega_i}(C_j) * \text{BetP}^{\Omega_j}(C_i) \qquad (12)$$

and the partition $\mathscr{P}^{K-1}$ is made from the new cluster (C_i, C_j) and all the other clusters of $\mathscr{P}^K$. The point by doing so is to prove that if we maximize the degree of probability we will have the couple of clusters to combine. Of course, this approach will give exactly the same partitions than the classical ascendant hierarchical clustering, but the dendrogram can be built from BetP and the best partition (*i.e.* the number of clusters) can be preferred to find. The indexed hierarchy will be indexed by the sum of BetP which will lead to more precise and specific results according to the dissimilarity between objects and therefore will facilitate our process.

Hereafter, we define another way to build the partition $\mathscr{P}^{K-1}$. For each initial object x_i to classify, it exists a cluster of $\mathscr{P}^K$ such as $x_i \in C_k$. We consider the frame of discernment $\Omega_i = \{C_1, \ldots, C_K\} \setminus \{C_k\}$, m can be noted m^{Ω} and we define the mass function:

$$m^{\Omega_i}(C_{k_j}) = \prod_{x_j \in C_{k_j}} \alpha e^{-\gamma d^2(x_i, x_j)} \qquad (13)$$

$$m^{\Omega_i}(\Omega_i) = 1 - \prod_{x_j \in C_{k_j}} \alpha e^{-\gamma d^2(x_i, x_j)} \qquad (14)$$

In order to find a mass function for each cluster C_i of $\mathscr{P}^K$, we combine all the mass functions given by all objects of C_i by a combination rule such as the Dempster rule of combination given by equation (6). Then, to merge both clusters we use the equation (12) as before. The sum of the pignisitic probability will be the index of the dendrogram, called BetP index.

5 Experimentations

Experiments were first applied on diamond data set composed of twelve objects as describe in Figure 1.a and analyzed in [7]. The dendrograms for both classical and Belief Hierarchical Clustering (BHC) are represented by Figures 1.b and 1.c. The object 12 is well considered as an outlier with both approaches. With the belief hierarchical clustering, this object is clearly different, thanks to the pignistic probability. For HAC, the distance between object 12 and other objects is small, however, for BHC, there is a big gap between object 12 and others. This point out that our method is better for detecting outliers. If the objects 5 and 6 are associated to 1, 2, 3 and 4 with the classical hierarchical clustering, with BHC these points are more

identified as different. This synthetic data set is special because of the equidistance
of the points and there is no uncertainty.

a. Diamond data set

b. Hierarchical clustering
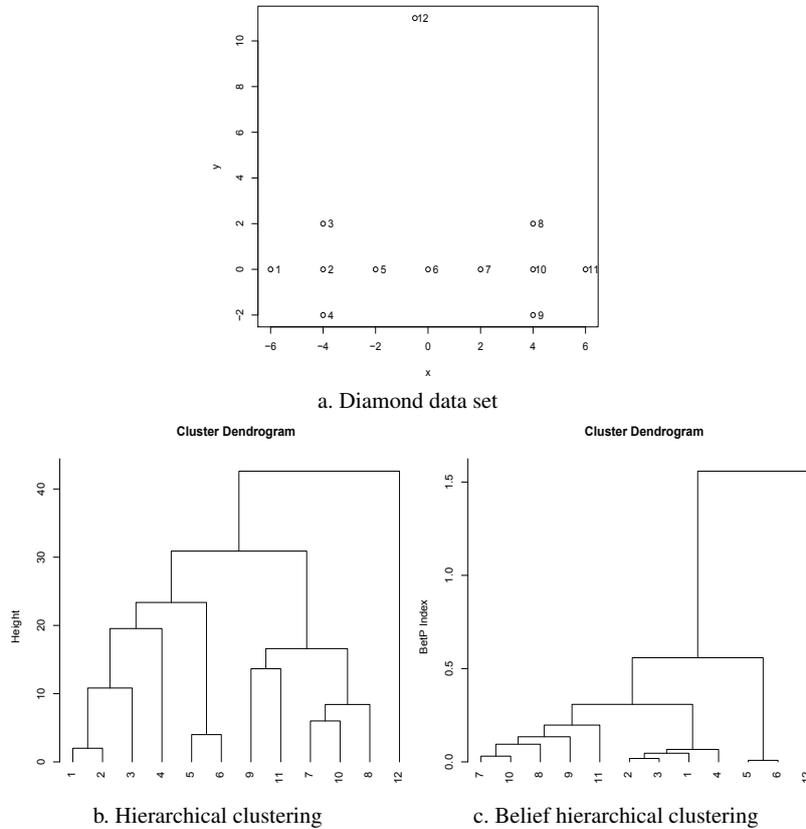
c. Belief hierarchical clustering

Fig. 1: Clustering results for Diamond data set.

We continue our experiments with a well-known data set, Iris data set, which
is composed of flowers from four types of species of Iris described by sepal length,
sepal width, petal length, and petal width. The data set contains three clusters known
to have a significant overlap. In order to reduce the complexity and present distinctly
the dendrogram, we first used the *k*-means method to get initial few clusters for our
algorithm. Several experiments have been used with several number of clusters. We
present in Figure 2 the obtained dendrograms for 10 and 13 clusters. We notice
different combinations between the nearest clusters for both classical and belief
hierarchical clustering. The best situation for BHC is obtained with the pignistic
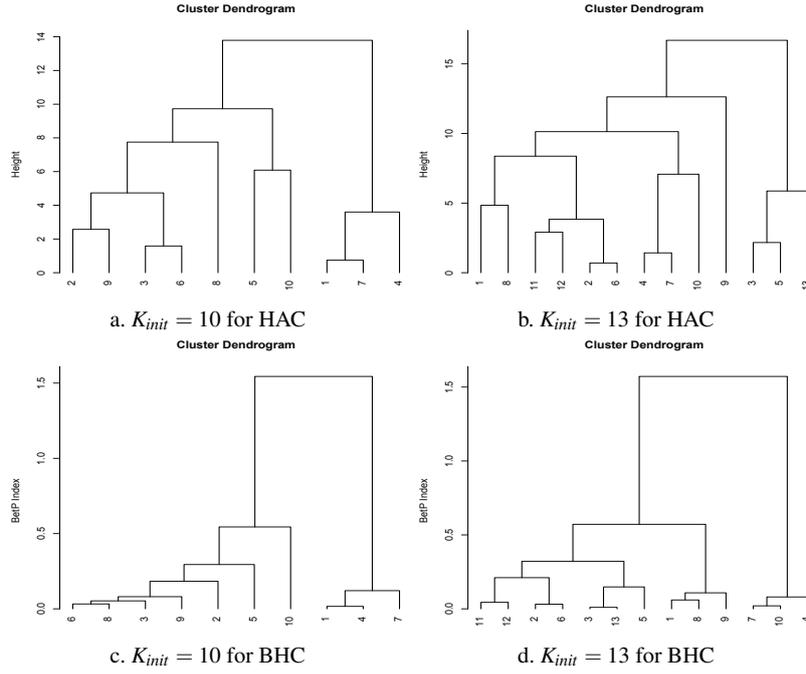
a. $K_{init} = 10$ for HAC   b. $K_{init} = 13$ for HAC

c. $K_{init} = 10$ for BHC   d. $K_{init} = 13$ for BHC

Fig. 2: Clustering results on IRIS data set for both hierarchical (HAC) (Fig. a and b) and belief hierarchical (BHC) (Fig. c and d) clustering (K_{init} is the cluster number by k-means first).

equal to 0.5 because it indicates that the data set is composed of three significant clusters which reflects the real situation. For the classical hierarchical clustering the results are not so obvious. Indeed, for HAC, it is difficult to decide for the optimum cluster number because the measure used to merge clusters is the euclidean distance and it is small as seen in Figure 2.c. However, for BHC, it is more easy to do this due to the use of the pignistic probability.

In order to evaluate the performance of our method, we use some of the most popular measures: precision, recall and Rand Index (RI). The results for both BHC and HAC are summarized in Table 1. The first three columns are for BHC, while the others are for HAC. In fact, we suppose that F_c represents the final number of clusters and we start with $F_c = 2$ until $F_c = 6$. We fixed the value of k_{init} at 13. We note that for $F_c = 2$ the precision is low while the recall is of high value, and that when we have a high cluster number ($F_c = 5$ or 6), the precision will be high but the recall will be relatively low. Thus, we note that for the same number of final clusters (*e.g.* $F_c = 4$), our method is better in terms of precision, recall and RI.

Table 1: Evaluation results

	BHC			HAC		
	Precision	Recall	RI	Precision	Recall	RI
$F_c = 2$	0.5951	1.0000	0.7763	0.5951	1.0000	0.7763
$F_c = 3$	0.8011	0.8438	0.8797	0.6079	0.9282	0.7795
$F_c = 4$	0.9506	0.8275	0.9291	0.8183	0.7230	0.8561
$F_c = 5$	0.8523	0.6063	0.8360	0.8523	0.6063	0.8360
$F_c = 6$	0.9433	0.5524	0.8419	0.8916	0.5818	0.8392

6 Conclusion

Ultimately, we have introduced a new clustering method using the hierarchical paradigm in order to implement uncertainty in the belief function framework. This method puts the emphasis on the fact that one object may belong to several clusters. It seeks to merge clusters based on its pignistic probability. Our method was proved on data sets and the corresponding results have clearly shown its efficiency. The algorithm complexity has revealed itself as the usual problem of the belief function theory. Our future work will be devoted to focus on this peculiar problem.

References

1. Bezdek, J.C., Ehrlich, R., Fulls, W.: Fcm: The fuzzy c-means clustering algorithm. Computers and Geosciences **10**(2-3), 191–203 (1984)
2. Denœux, T.: A *k*-Nearest Neighbor Classification Rule Based on Dempster-Shafer Theory. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans **25**(5), 804–813 (1995)
3. Denœux, T., Masson, M.: EVCLUS: Evidential Clustering of Proximity Data. IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics **34**(1), 95–109 (2004)
4. Hariz, S.B., Elouedi, Z., Mellouli, K.: Clustering approach using belief function theory. In: AIMSA, pp. 162–171 (2006)
5. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning; data mining, inference and prediction. Springer Verlag, New York (2001)
6. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability **11** (1967)
7. Masson, M., Denœux, T.: Clustering interval-valued proximity data using belief functions. Pattern Recognition Letters **25**, 163–171 (2004)
8. Schubert, J.: Clustering belief functions based on attracting and conflicting metalevel evidence. In: B. Bouchon-Meunier, L. Foulloy, R. Yager (eds.) Intelligent Systems for Information Processing: From Representation to Applications. Elsevier Science (2003)
9. Shafer, G.: Mathematical Theory of evidence. Princeton Univ (1976)
10. Smets, P., Kennes, R.: The Transferable Belief Model. Artificial Intelligent **66**, 191–234 (1994)
11. Zouhal, L.M., Denœux, T.: An Evidence-Theoric *k*-NN Rule with Parameter Optimization. IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews **28**(2), 263–271 (1998)