sources are dependent, thus a source may be dependent towards another one by saying the same thing (positive dependence) or saying the opposite (negative dependence).

In the following, we introduce preliminaries of Dempster-Shafer theory as well as evidential databases in the second section. In the third section, a belief clustering method is presented and its classification result is used to estimate the sources degree of independence. If sources seem to be dependent, it is interesting to investigate whether this dependency is positive or negative in the fourth section. This method is tested on random mass functions in the fifth section. Finally, conclusions are drawn.

2 Theory of belief functions

The theory of belief functions [4, 11] is used to model imperfect data.

In the theory of belief functions, the *frame of discernment* also called *universe of discourse* $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ is a set of n elementary and mutually exclusive and exhaustive hypotheses. These hypotheses are all the possible and eventual solutions of the problem under study.

The *power set* 2^Ω is the set of all subsets made up of hypotheses and union of hypotheses from Ω.

The *basic belief function (bba)* also called *mass function* is a function defined on the power set 2^Ω and affects a value from $[0,1]$ to each subset. A mass function m is a function:

$$m : 2^\Omega \mapsto [0,1] \tag{1}$$

such that:

$$\sum_{A \subseteq \Omega} m(A) = 1 \tag{2}$$

One or many subsets may have a non null mass, this mass is the source's belief that the solution of the problem under study is in that subset.

The *belief function (bel)* is the minimal belief allocated to a subset A justified by available information on B ($B \subseteq A$):

$$\begin{aligned} bel : 2^\Omega &\to [0,1] \\ A &\mapsto \sum_{B \subseteq A, B \neq \emptyset} m(B) \end{aligned} \tag{3}$$

The implicability function b is proposed to simplify computations:

$$\begin{aligned} b : 2^\Omega &\to [0,1] \\ A &\mapsto \sum_{B \subseteq A} m(B) = bel(A) + m(\emptyset) \end{aligned} \tag{4}$$

The theory of belief functions is used to model uncertain information and also to combine them. A great number of combination rules are proposed such as *Dempster's rule of combination* [4] which is used to combine two different mass functions m_1 and m_2

provided by two different sources as follows:

$$m_{1\oplus 2}(A) = (m_1 \oplus m_2)(A) = \begin{cases} \dfrac{\sum\limits_{B\cap C=A} m_1(B)\times m_2(C)}{1-\sum\limits_{B\cap C=\emptyset} m_1(B)\times m_2(C)} & \forall A \subseteq \Omega,\ A \neq \emptyset \\ 0 & \text{if } A = \emptyset \end{cases} \tag{5}$$

The pignistic transformation is used to compute pignistic probabilities from masses in the purpose of making a decision. The pignistic probability of a singleton X is given by:

$$BetP(X) = \sum_{Y\in 2^{\Theta}, Y\neq\emptyset} \frac{|X\cap Y|}{|Y|} \frac{m(Y)}{1-m(\emptyset)}. \tag{6}$$

2.1 Conditioning

When handling a mass function, a new evidence can arise confirming that a proposition A is true. Therefore, the mass affected to each focal element C has to be reallocated in order to take consideration of this new evidence. This is achieved by the conditioning operator. Conditioning a mass function m over a subset A$\subseteq \Omega$ consists on restricting the frame of possible propositions 2^{Ω} to the set of subsets having a non empty intersection with A. Therefore the mass allocated to $C \subseteq \Omega$ is transferred to $\{C \cap A\}$. The obtained mass function, result of the conditioning, is noted $m_{[A]} : 2^{\Omega} \rightarrow [0,1]$ such that [10]:

$$m_{[A]}(C) = \begin{cases} 0 & \text{for } C \not\subseteq A \\ \sum\limits_{X\subseteq \bar{A}} m(C\cup X) & \text{for } C \subseteq A \end{cases} \tag{7}$$

where $\bar{A}$ is the complementary of A.

2.2 Generalized bayesian theorem and disjunctive rule of combination

The *generalized bayesian theorem* (*GBT*), proposed by Smets [9], is a generalization of the bayesian theorem where the joint belief function replaces the conditional probabilities. Let X and Y be two dependent variables defined on the frames of discernment Ω_X and Ω_Y. Suppose that the conditional belief function $bel_{[X]}(Y)$ represents the conditional belief on Y according to X.

The aim is to compute the belief on X conditioned on Y. Thus, the GBT is used to build $bel_{[Y]}(X)$:

$$\begin{aligned} bel_{[Y]}(X) &= b_{[Y]}(X) - b_{[Y]}(\emptyset) \\ bel_{[Y]}(X) &= \prod_{x_i\in\bar{X}} b_{[x_i]}(\bar{Y}) \end{aligned} \tag{8}$$

The conditional belief function $bel_{[X]}(Y)$ can be extended to the joint frame of discernment $\Omega_X \times \Omega_Y$, then conditioned on $y_i \subseteq \Omega_Y$ and the result is then marginalized on X, the corresponding operator is the disjunctive rule of combination:

$$\begin{aligned} bel_{[X]}(Y) &= b_{[X]}(Y) - b_{[X]}(\emptyset) \\ bel_{[X]}(Y) &= \prod_{x_i\in X} b_{[x_i]}(Y) \end{aligned} \tag{9}$$

2.3 Evidential database

Classic databases are used to store certain data, whereas data are not always certain but can sometimes be uncertain and even incomplete. The use of *evidential database* (*EDB*), also called *D-S database*, for storing data from different levels of uncertainty. Evidential databases proposed by [1] and [6] are databases containing both certain and/or uncertain data. Uncertainty and incompleteness in evidential databases are modeled with the theory of belief functions previously introduced.

An evidential database is a database having n records and p attributes such that every attribute a ($1 \leq a \leq p$) has an exhaustive domain Ω_a containing all its possible values: its *frame of discernment* [6].

An EDB has at least one *evidential attribute*. Values of this attribute can be uncertain, thus these values are mass functions and named *evidential values*. An *evidential value* V_{ia} for the i^{th} record and the a^{th} attribute is a mass function such that:

$$\begin{aligned} & m_{ia} : 2^{\Omega_a} \rightarrow [0,1] \text{ with:} \\ & m_{ia}(\emptyset) = 0 \text{ and } \sum_{X \subseteq \Omega_a} m_{ia}(X) = 1 \end{aligned} \tag{10}$$

Table 1 is the example of an evidential database having 2 evidential attributes namely $road\ \ condition$ and $weather$. Records of this evidential database are road condition and weather predictions for the five coming days according to one source. The domain $\Omega_{weather} = \{Sunny\ \ S,\ \ Rainy\ \ R,\ \ Windy\ \ W\}$ is the frame of discernment of the evidential attribute $weather$ and the domain $\Omega_{RC} = \{Safe\ \ S,\ \ Perilous\ \ P,\ \ Dangerous\ \ D\}$ is the frame of discernment of the evidential attribute *road condition*.

Table 1. Example of an EDB

Day	Road condition	Weather
d_1	$\{P \cup D\}(1)$	$S(0.3)\ \ R(0.7)$
d_2	$S(1)$	$S(0.2)\ \ \{S \cup W\}(0.6)\ \ \{S \cup R \cup W\}(0.2)$
d_3	$\{S \cup P \cup D\}(1)$	$\{S \cup R \cup W\}(1)$
d_4	$S(0.6)\ \ \{S \cup P\}(0.4)$	$S(0.4)\ \ \{S \cup R \cup W\}(0.6)$
d_5	$S(1)$	$S(1)$

3 Independence

Evidential databases previously described store a great number of records (objects). Similar objects may be stored in that type of databases meaning that similar situations can be redundant. Clustering techniques are used to group several similar objects into the same cluster. When having n objects, the most similar ones are affected to the same group. Applying a clustering technique to evidential database records (*i.e.* to mass functions) is useful in order to group redundant cases. Some evidential clustering techniques are already proposed such as [5, 2, 8]. A method of sources independence estimating is submitted in [3] and recalled in the following. In this paper we suggest to specify the type of dependence when sources are dependent and also to use this information for evidential database enrichment.

3.1 Clustering

We use here a clustering technique using a distance on belief functions given by [7] such as in [2]. The number of clusters C have to be known, a set T contains n objects $o_i : 1 \leq i \leq n$ which values m_{ij} are belief functions defined on the frame of discernment Ω_a. Ω_a is the frame of discernment of the evidential attribute.

This set T is a table of an evidential database having at least one evidential attribute and at most p evidential attributes. m_{ia} is a mass function value of the a^{th} attribute for the i^{th} object (record), this mass function is defined on the frame of discernment Ω_a (Ω_a is the domain of the a^{th} attribute). A dissimilarity measure is used to quantify the dissimilarity of an object o_i having $\{m_{i1}, \dots, m_{ij}, \dots, m_{ip}\}$ as its attributes values towards a cluster Cl_k containing n_k objects o_j. The dissimilarity D of the object o_i and the cluster Cl_k is as follows:

$$D(o_i, Cl_k) = \frac{1}{n_k} \sum_{j=1}^{n_k} \frac{1}{p} \sum_{l=1}^{p} d(m_{il}^{\Omega_a}, m_{jl}^{\Omega_a}) \tag{11}$$

and

$$d(m_1^{\Omega_a}, m_2^{\Omega_a}) = \sqrt{\frac{1}{2}(m_1^{\Omega_a} - m_2^{\Omega_a})^t \underline{\underline{D}}(m_1^{\Omega_a} - m_2^{\Omega_a})} \tag{12}$$

with:

$$\underline{\underline{D}}(A,B) = \begin{cases} 1 & \text{if } A = B = \emptyset \\ \frac{|A \cap B|}{|A \cup B|} & \forall A, B \in 2^{\Omega_a} \end{cases} \tag{13}$$

We note that $\frac{1}{p}\sum_{l=1}^{p} d(m_{il}^{\Omega_a}, m_{jl}^{\Omega_a})$ is the dissimilarity between two objects o_i and o_j. The dissimilarity between two objects is the mean of the distances between belief functions values of evidential attributes (evidential values). Each object is affected to the closest cluster (having the minimal dissimilarity value) in an iterative way until reaching the stability of the cluster repartition.

3.2 Independence measure

Definition 1. *Two sources are considered to be independent when the knowledge of one source does not affect the knowledge of the other one.*

The aim is to study mass functions provided by two sources in order to reveal any dependence between these sources. Provided mass functions are stored in evidential databases, thus each evidential database stores objects having evidential values for some evidential attributes. Suppose having two evidential databases EDB_1 and EDB_2 provided by two distinct sources s_1 and s_2. Each evidential database contains about n

records (objects) and p evidential attributes. Each mass function stored in that EDB can be a classification result according to each source. The aim is to find dependence between sources if it exists. In other words, two sources s_1 and s_2 classifying each one n objects. m_{ia} (a^{th} attribute's value for the i^{th} object) provided by s_1 and that provided by s_2 are referred to the same object i. If s_1 and s_2 are dependent, there will be a relation between their belief functions. Thus, we suggest to classify mass functions of each source in order to verify if clusters are independent or not. The proposed method is in two steps, in the first step mass functions of each source are classified then in the second step the weight of the linked clusters is quantified.

1. Step 1: Clustering
   Clustering technique, presented in section 3.1, is used in order to classify mass functions provided by both s_1 and s_2, the number of clusters can be the cardinality of the frame of discernment. After the classification, objects stored in EDB_1 and provided by s_1 are distributed on C clusters and objects of s_2 stored in EDB_2 are also distributed on C clusters. The output of this step are C clusters of s_1, noted Cl_{k_1} and C different clusters of s_2, noted Cl_{k_2}, with $1 \leq k_1, k_2 \leq C$.
2. Step 2: Cluster independence
   Once cluster repartition is obtained, the degree of independence and dependence between sources are quantified in this step. The most similar clusters have to be linked, a cluster matching is performed for both clusters of s_1 and that of s_2. The dissimilarity between two clusters Cl_{k_1} of s_1 and Cl_{k_2} of s_2 is the mean of distances between objects o_i contained in Cl_{k_1} and all the objects o_j contained on Cl_{k_2}:

   $$\delta^1(Cl_{k_1}, Cl_{k_2}) = \frac{1}{n_{k_1}} \sum_{l=1}^{n_{k_1}} D(o_l, Cl_{k_2}) \tag{14}$$

   We note that n_{k_1} is the number of objects on the cluster Cl_{k_1} and δ^1 is the dissimilarity towards the source s_1.

Dissimilarity matrix M_1 and M_2 containing respectively dissimilarities between clusters of s_1 according to clusters of s_2 and dissimilarities between clusters of s_2 according to clusters of s_1, are defined as follows:

$$M_1 = \begin{pmatrix} \delta^1_{11} & \delta^1_{12} & \dots & \delta^1_{1C} \\ \dots & \dots & \dots & \dots \\ \delta^1_{k1} & \delta^1_{k2} & \dots & \delta^1_{kC} \\ \dots & \dots & \dots & \dots \\ \delta^1_{C1} & \delta^1_{C2} & \dots & \delta^1_{CC} \end{pmatrix} \quad \text{and} \quad M_2 = \begin{pmatrix} \delta^2_{11} & \delta^2_{12} & \dots & \delta^2_{1C} \\ \dots & \dots & \dots & \dots \\ \delta^2_{k1} & \delta^2_{k2} & \dots & \delta^2_{kC} \\ \dots & \dots & \dots & \dots \\ \delta^2_{C1} & \delta^2_{C2} & \dots & \delta^2_{CC} \end{pmatrix} \tag{15}$$

We note that $\delta^1_{k_1k_2}$ is the dissimilarity between Cl_{k_1} of s_1 and Cl_{k_2} of s_2 and $\delta^2_{k_1k_2}$ is the dissimilarity between Cl_{k_2} of s_2 and Cl_{k_1} of s_1 and $\delta^1_{k_1k_2} = \delta^2_{k_2k_1}$. M_2 the dissimilarity matrix of s_2 is the transpose of M_1 the dissimilarity matrix of s_1. Clusters of s_1 are matched to the most similar clusters of s_2 and clusters of s_2 are linked to the most similar clusters of s_1. Two clusters of s_1 can be linked to the same cluster of s_2. A different matching of clusters is obtained according to s_1 and s_2. A set of matched clusters is obtained for both sources and a mass function can be used to quantify the independence between the couple of clusters. Suppose that the cluster Cl_{k_1} of s_1 is matched to Cl_{k_2} of s_2, a mass function m defined on the frame of discernment $\Omega_I = \{Dependent\ \bar{I}, Independent\ I\}$ describes how much this couple of clusters is independent or dependent as follows:

$$\begin{cases} m^{\Omega_I}_{k_1k_2}(\bar{I}) = \alpha(1-\delta^1_{k_1k_2}) \\ m^{\Omega_I}_{k_1k_2}(I) = \alpha\delta^1_{k_1k_2} \\ m^{\Omega_I}_{k_1k_2}(\bar{I}\cup I) = 1-\alpha \end{cases} \tag{16}$$

where α is a discounting factor. When $\alpha = 1$, the obtained mass function is a probabilistic mass function which quantifies the dependence of each matched clusters according to each source. A mass function is obtained for each matched clusters Cl_{k_1} and Cl_{k_2}, thus C mass functions are obtained for each source. The combination of that C mass functions $m^{\Omega_I}_{k_1k_2}$ using Dempster's rule of combination is a mass function m^{Ω_I} reflecting the overall dependence of one source towards the other one:

$$m^{\Omega_I} = \oplus m^{\Omega_I}_{k_1k_2} \tag{17}$$

After the combination, two mass functions describing the dependence of s_1 towards s_2 and that of s_2 towards s_1 are obtained. Pignistic probabilities are derived from mass functions using the pignistic transformation in a purpose of making decision about the dependence of sources. A source s_1 is dependent on the source s_2 if $BetP(\bar{I}) \geq 0.5$ otherwise it is independent. $BetP(\bar{I})$ is the pignistic probability of $\bar{I}$ computed from $m^{\Omega_I}_{s_1}(\bar{I})$.

4 Negative and positive dependence

A mass function describing the independence of one source towards another one can inform about the degree of dependence but does not inform if this dependence is positive or negative. In the case of dependent sources, this dependence can be positive meaning that the classification of one source is directly affected by the classification of the other one, thus both sources have the same knowledge. In the case of negative dependence, the knowledge of one source is the opposite of the other one.

Definition 2. *A source is positively dependent on another source when the belief of the first one is affected by the knowledge of the belief of the second one and both beliefs are similar.*
If a source s_1 is negatively dependent on s_2, s_1 is always saying the opposite of what said s_2.

Definition 3. *A source is negatively dependent on another source when their beliefs are different although the belief of the first one is affect by the knowledge of the belief of the second one.*
If matched clusters contain the same objects thus these clusters are positively dependent. It means that both sources are almost classifying objects in the same way. If matched clusters contain different objects thus one source is negatively dependent on the other because it is classifying differently the same objects. A mass function defined on the

frame of discernment $\Omega_P = \{Positive\ Dependent\ P,\ Negative\ Dependent\ \bar{P}\}$ can be built in order to quantify the positivity or negativity of the dependence of a cluster Cl_{k_1} of s_1 and a cluster Cl_{k_2} of s_2 such that Cl_{k_1} and Cl_{k_2} are matched according to s_1 as follows:

$$\begin{cases} m_{k_1k_2}^{\Omega_P}(P|\bar{I}) = 1 - \frac{|Cl_{k_1} \cap Cl_{k_2}|}{|Cl_{k_1}|} \\ m_{k_1k_2}^{\Omega_P}(\bar{P}|\bar{I}) = \frac{|Cl_{k_1} \cap Cl_{k_2}|}{|Cl_{k_1}|} \\ m_{k_1k_2}^{\Omega_P}(P \cup \bar{P}|\bar{I}) = 0 \end{cases} \tag{18}$$

We note that these mass functions are conditional mass functions because they do not exist if sources are independent, thus these mass functions are dependent on the dependency of sources. These mass functions are also probabilistic. In order to have the marginal mass functions, the Disjunctive Rule of Combination proposed by Smets [9] in section 2.2 can be used in order to compute the marginal mass functions defined on the frame of discernment Ω_P. Marginal mass functions are combined using Dempster's rule of combination presented in equation (5), then the pignistic transformation is used to compute pignistic probabilities which are used to decide about the type of dependence and also to enrich the corresponding evidential databases.

5 Example

The method described above is tested on generated mass functions. Mass functions are generated randomly using the following algorithm:

This algorithm is used to generate n random mass functions which decisions (using the pignistic transformation) are not known, whereas in the case of positive or negative dependence decision classes have to be checked.

Algorithm 1 Mass generating

```
Require: $|\Omega|$, $n$ : number of mass functions
for $i = 1$ to $n$ do
    Choose randomly $F$, the number of focal elements on $[1, |2^\Omega|]$.
    Divide the interval $[0, 1]$ into $F$ continuous sub intervals.
    Choose randomly a mass from each sub interval and attribute it to focal elements.
    Attribute these masses to focal elements previously chosen.
    The complement to 1 of the attributed masses sum is affected to the total ignorance $m(\Omega)$.
end for
return $n$ mass functions
```

1. Positive dependence:
   When sources are positively dependent, the decided class (using the pignistic transformation) of one is directly affected by that of the other one. To test this case, we generated 100 mass functions on a frame of discernment of cardinality 5. Both sources are classifying objects in the same way because one of the sources is positively dependent on the other as follows:

Algorithm 2 Positive dependent Mass function generating

```
Require: $n$ mass functions generated using algorithm 1, Decided classes
for $i = 1$ to $n$ do
    Find the $m$ focal elements of the $i^{th}$ mass function
    for $j = 1$ to $m$ do
        The mass affected to the $j^{th}$ focal element is transferred to its union with the decided class.
    end for
end for
return $n$ mass functions
```

Applying the method described above, we obtained this mass function defined on the frame $\Omega_P = \{P, \ \bar{P}\}$ and describing the positive and negative dependence of s_1 towards s_2:
$m(P) = 0.679$, $m(\bar{P}) = 0.297$, $m(\bar{P} \cup P) = 0.024$
Using the pignistic transformation $BetP(P) = 0.691$ and $BetP(\bar{P}) = 0.309$, meaning that s_1 is positively dependent on s_2. The marginal mass function of the positive and negative dependence of s_2 according to s_1:
$m(P) = 0.6459$, $m(\bar{P}) = 0.3272$, $m(\bar{P} \cup P) = 0.0269$
Using the pignistic transformation $BetP(P) = 0.6593$ and $BetP(\bar{P}) = 0.3407$, meaning that s_2 is positively dependent on s_1.

2. Negative dependence:
When sources are negatively dependent, one of the sources is saying the opposite of the other one. In other words, when the classification result of the first source is a class A, the second source may classify this object in any other class but not A. Negative dependent mass functions are generated in the same way as positive dependent mass functions but the mass of each focal element is transferred to focal elements having a null intersection with the decided class. In that case, we obtain this mass function of the dependence of s_1 according to s_2:
$m(P) = 0.0015$, $m(\bar{P}) = 0.9909$, $m(\bar{P} \cup P) = 0.0076$
Using the pignistic transformation $BetP(P) = 0.0053$ and $BetP(\bar{P}) = 0.9947$, meaning that s_1 is negatively dependent on s_2. The marginal mass function of the dependence of s_2 according to s_1:
$m(P) = 0.0011$, $m(\bar{P}) = 0.9822$, $m(\bar{P} \cup P) = 0.0167$
Using the pignistic transformation $BetP(P) = 0.00945$ and $BetP(\bar{P}) = 0.99055$, meaning that s_2 is negatively dependent on s_1. These mass functions are added to the corresponding evidential databases to enrich them. $m^{\Omega_I}_{k_1k_2}$ are not certain mass functions, thus some degree of total ignorance appears in $m(\bar{P} \cup P)$ when using the DRC.

6 Conclusion

Enriching evidential databases with dependence information can inform users about the degree of interaction between their sources. In some cases where one source is completely dependent on an another one, the evidential database of that source can be discarded when making a decision. In this paper, we suggested a method estimating the dependence degree of one source towards another one. As a future work, we may try to estimate the dependence of one source according to many other sources and not only one source.

References

1. Bach Tobji, M.-A., Ben Yaghlane, B., Mellouli, K.: A New Algorithm for Mining Frequent Itemsets from Evidential Databases. In Information Processing and Management of Uncertainty (IPMU'2008), pp. 1535–1542. Malaga, Spain (2008).
2. Ben Hariz, S., Elouedi, Z. and Mellouli, K.: Clustering Approach Using Belief Function Theory. In: Euzenat, J., Domingue, J. (eds.) AIMSA'2006, LNCS (LNAI), vol. 4183, pp. 162–171. Springer, Heidelberg (2006).
3. Chebbah, M., Martin, A. and Ben Yaghlane, B.: About sources dependence in the theory of belief functions. In the 2^{nd} International Conference on Belief Functions (BELIEF'2012). Compiègne, France (2012).
4. Dempster, A. P.: Upper and Lower probabilities induced by a multivalued mapping. Annals of Mathematical Statistics, vol. 38, pp. 325–339 (1967).
5. Denoeux, T.: A *k*-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Transactions on Systems, Man and Cybernetics, vol. 25(5), pp. 804–813 (1995).
6. Hewawasam, K.K.R.G.K., Premaratne, K. and Subasingha, S.P., Shyu, M.-L.: Rule Mining and Classification in Imperfect Databases. In Int. Conf. on Information Fusion, pp. 661–668. Philadelphia, USA (2005).
7. A.-L. Jousselme, D. Grenier and E. Bossé, "A new distance between two bodies of evidence," *Information Fusion*, vol. 2, pp. 91–101, (2001).
8. Masson, M. -H., Denoeux, T.: ECM: an evidential version of the fuzzy c-means algorithm. Pattern Recognition, vol. 41, pp. 1384–1397 (2008).
9. Smets, P.: Belief Functions: the Disjunctive Rule of Combination and the Generalized Bayesian Theorem. International Journal of Approximate Reasoning, vol. 9, pp. 1–35 (1993).
10. Smets, P., Kruse, R.: The transferable belief model for belief representation. Uncertainty in Information Systems: From Needs to Solutions, pp. 343-368 (1997).
11. Shafer, G.: A mathematical theory of evidence. Princeton University Press (1976).