ECTD: Evidential Clustering and case Types Detection for case base maintenance

Safa Ben Ayed
LARODEC, Université de Tunis,
Institut Supérieur de Gestion,
2000 Le Bardo, Tunisia
safa.ben.ayed@hotmail.fr

Zied Elouedi
LARODEC, Université de Tunis,
Institut Supérieur de Gestion,
2000 Le Bardo, Tunisia
zied.elouedi@gmx.fr

Eric Lefèvre
Univ. Artois, EA 3926,
LGI2A,
62400 Béthune, France
eric.lefevre@univ-artois.fr

***Abstract*—The key factor for the success of Case Based Reasoning (CBR) systems is the quality of their case bases as well as the time spent in case retrieval process which is mainly depending on case bases' size. Indeed, the speed of the retrieval process is seriously decreasing when the case base becomes so heavy. To vouch for case bases' quality, a maintenance process must be provided. Hence, a field for Case Base Maintenance (CBM) emerges. However, a lot of works in CBM field suffers from some limitations and they generally reduce case base's competence during maintenance, especially when cases involving imprecise or uncertain information. To deal with these problems, we propose, in this paper, a new CBM approach named ECTD, Evidential Clustering and case Types Detection for case base maintenance, which is able to manage imperfection in cases by using belief function theory. The key idea of ECTD approach is to use machine learning technique, more accurately the evidential c-means (ECM). Then, it divides cases relative to the different partitions of clusters into four types so that we can subsequently perform the case base maintenance.**

***Index Terms*—Case Based Reasoning (CBR), Case Base Maintenance (CBM), uncertainty, belief function theory, clustering, machine learning.**

I. Introduction

The main objective of Artificial Intelligence (AI) is to conceive systems able to understand, emulate and reproduce human reasoning. Case Based Reasoning is a problem solving paradigm which is, similarly to human being, based on reusing past experiences to solve new problems by assuming that similar problems have similar solutions. Therefore, CBR presents a perfect sample of AI. In CBR systems, a problem is solved after following an entire cycle composed of four phases known by the 4RE [1]: REtrieve, REuse, REvise and REtain. First of all, the CBR system searches in its case base the case(s) having the most similar problem description to the target one by using a similarity measure. Then, it tries to adapt its solution in order to build a new one having the ability to solve the target problem. However, this solution can be inadequate, so it should be revised and validated. Finally, the new problem attached to its proposed solution will be stored in the case base as a new case. Like others, this case will serve for new problem resolution. This incremental evolution of case bases causes a great storage requirement, consequently, a very time consuming retrieval process which leads to the seriously decrease of case bases' performance. Accordingly, we observe a strongly emergence of the Case Base Maintenance (CBM) field and a great awareness of it's importance towards the learning phase. Materially, it can be explained by the utility of CBR systems and their achieved success in several domains. Thus, CBM has been defined in [2] as the process aiming to improve the performance of CBR systems: *"Case-base maintenance implements policies for revising the organization or contents (representation, domain contents, accounting information, or implementation) of the case base in order to facilitate future reasoning for a particular set of performance objectives"*. In fact, we find in the literature several CBM policies aiming to maintain case bases for CBR systems. However, they generally suffer from some weaknesses. For instance, we cite the limitation of their complexity when dealing with large case bases or their disability to manage imperfection in knowledge whereas cases which refer to real world situations are full of uncertainty, which is related to the source of information, and imprecision, which is related to the information itself. To deal with the latter problem, there is a number of theories that can be used. One of the most appropriates theories is the belief function theory [16] [17] which presents a powerful tool and accounts for knowledge uncertainty in different levels from the complete ignorance to the total certainty.

For these reasons, we propose, in this paper, a new Case Base Maintenance approach named ECTD, encoding *"Evidential Clustering and case Types Detection for case base maintenance"*, able to deal with imperfection in cases descriptions and retain the most competent in problem solving by using belief function theory as well as machine learning techniques, more precisely Evidential c-means (ECM) [21].

The remainder of this work is organized as follows. In section II, we expose a brief review of CBM policies in the literature. The necessary background related to the theory of belief functions is presented in section III where it provides on the one hand the basic notions of the theory and on the other hand an evidential clustering technique called Evidential c-means [21]. Section IV details the different steps of our new proposed ECTD approach. Finally, experiments and results are given in section V.

II. CBM policies: Related work

In the literature, there are several policies that deal with the case base maintenance and this can only indicate the significance of CBM field towards CBR systems success. Basically, the objective of CBM policies is to reduce the size of case bases in order to cut down the retrieval time. In short, we list some methods that have been made for maintaining case bases by classifying them into four strategies: Data reduction methods based on selection, Selective rule based strategy, Optimization strategy and Partition strategy.

A. Data reduction methods based on selection

The main objective of this class of CBM policies is to select from case bases only the representative cases where their set is able to cover all the rest of cases. For instance, the Condensed Nearest Neighbor approach (CNN) [5] is a data reduction algorithm that selects iteratively case base's prototypes and adds them in a new case base. The idea behind CNN approach is to choose randomly a case from the original case base and test if the edited one can solve it or not. If not, this case will be selected to be added in the edited case base and removed from the original one. In the same road, we mention the Reduced Nearest Neighbor approach (RNN) [6] which is characterized by using the whole case base as an initial reduced set. Then, removing cases until no case from the original case base is misclassified by the remaining cases in the edited case base. Besides, we cite the Selective Nearest Neighbor approach (SNN) [7] which guarantees the retrieval of the minimal cases subset by improving the mix between CNN and RNN rules. The idea behind this approach is that all cases in the training set must be closer to a case in the selective set than any one in the training set.

B. Selective rule based strategy

Authors in [8] proposed an approach for maintaining case bases based on selective rules in order to lead their reduction. This method is divided into two steps: First, applying a method for feature reduction since the CBR is sensitive to inaccurate data such as noises and redundant attributes that can seriously reduce case base's performance. Second, defining selective rules, and only cases satisfying these rules will be selected in order to achieve the dynamic maintenance of CBR systems. These proposed rules concern noisy and redundant cases that we generally aim to remove during case base maintenance.

C. Optimization strategy

Generally, CBM policies that are related to this strategy value cases according to some evaluation criteria which influence the decision making about their deletion or retention. Hence, we highlight some of the most known criteria for evaluating a case base in a CBR system. The first one is the performance of a case bases which is measured by the time spent from the arrival of the new problem until proposing its solution. The second one is the case base's competence which is known as the range of problems that this case base is able to solve. Typically, when we talk about competence criterion, two main concepts arise: the coverage and the reachability. The coverage of one case is the set of all the target problems that this case can successfully solves. However, the reachability of a target problem represents the set of cases that can be used to solve this problem [3] [4].

According to Smiti and Elouedi [9], this strategy is divided into two main categories: Standard deletion methods and methods based on case's competence. For the first category, we cite Random Deletion policy (RD) where the principle is to delete cases randomly when a predefined size limitation is exceeded. Besides, we mention Utility Deletion policy (UD) which is based on Minton's utility [10] serving on case performance estimation. Its idea is to delete all cases having a bad performance, consequently, a negative utility.
Concerning the second category based on cases' competence and the two corresponding concepts coverage and reachability, we find as example RC-CNN approach which is an hybrid algorithm combining the CNN algorithm to Relative Coverage (RC) metric aiming to quantify competence contribution of cases [9]. In the same way, Brighton and Mellish [11] proposed an iterative approach based on the competence criterion for cases deletion called Iterative Case Filtering algorithm (ICF). Its principle is to delete cases having a coverage set size smaller than reachable set size i.e. a case will be deleted if more cases can solve it than it can solve itself.

D. Partition strategy

The greatest strength of CBM policies based on partition strategy is that they divide the original case base into a set of small ones where each one can be treated separately. Generally, these small case bases are created thanks to an unsupervised machine learning technique which is *Clustering*. This technique is suitable for such problem since a given case can be considered as an individual and the notion of distance between cases is well presented. Thus, we mention COID method [12] encoding Clustering, Outliers and Internal case Deletion which aims to select from each cluster only cases influencing case base's quality. First, COID applies DBSCAN [13] as the density based clustering technique and benefits from its ability to detect noisy cases. Then, it detects internal cases as those having the smallest Euclidean distance to cluster's center. Finally, COID is supposed to detect outliers. Thence, it used Interquartile Rang (IQR) to detect univariate outliers and Mahalanobis distance for multivariate ones. In addition, we can mention some extentions from COID approach such as WCOID [14] which adds a feature weihghting step and WCOID-DG [15] which combines WCOID with the Gaussian-means clustering technique in order to well estimate the number of clusters. Moreover, we cite examples of soft CBM policies that, to the best of our knowledge, are not numerous. To begin with, we mention simultaneously both of Fuzzy decision tree CBM policy [23] and Fuzzy-Rough CBM approach [24] which share the first step consisting of feature weights learning, the second consisting of case base partitioning and the fourth step maintaining the case base by representative cases selection. Concerning the third step, authors in [23] opt for fuzzy adaptation rule mining, whereas in [24] authors integrate rough and fuzzy sets theories in order to transfer the case knowledge to adaptation knowledge. For both, this step is the most important since it allows to find non-representative cases for each cluster. Last but not least, we cite the Soft CBM Competence Based Model (SCBM) [25] that uses foremost a soft clustering technique called Soft DBSCAN-GM (SDG) [26] which is based on fuzzy set theory [29]. In fact, the purpose of this first step is to create competence groups and facilitate the second one consisting of case types detection (Noisy, Isolated and Similar cases) after computing Fuzzy Mahalanobis Distances between cases and clusters.

E. Critics and discussion

Obviously, there is in the literature a lot of works aiming to maintain case bases for CBR systems as presented during the previous subsections, knowing that they have not been covered exhaustively. Actually, crisp CBM policies can be interesting but they suffer from a major limitation which is presented by their disability in managing imperfection in CBR case bases. Thus, they easily make confusion while making decision about cases deletion or retention since they are full of imperfection. On the other hand, methods managing uncertainty are quite interesting because they are more able to take into account real data related to imperfection. In fact, we already presented some methods that deal with this kind of knowledge such as [23], [24] and [25]. However, formalisms used by these methods did not manage or cover all aspects of uncertainty. Therefore, The risk of removing representative cases persists. By this way, our proposal for this paper is to extend this axis by using a theory having the capacity to take into account the overall uncertainty. This theory is the Evidence theory as it will presented in the next section.

III. Background related to Evidence theory

In order to be familiar with our contribution in this paper, some background related to the Evidence theory or belief function theory are required. On the one hand, we give an overview on the basic notions and concepts of belief function theory [16] [17]. On the other hand, a clustering technique in an evidential frame called Evidential c-means (ECM) [21] will be presented.

A. Belief function theory

Belief function theory, also called Evidence theory or Dempster-Shafer theory [16] [17], is a theoretical framework to model and quantify imperfect knowledge whether it is partial or unreliable. For reasoning under uncertainty, the belief function theory has interpreted differently using different models, inter alia Smets's Transferable Belief Model (TBM) [18] as a non-probabilistic model.

Let us consider the frame of discernment Θ as a finite set of variables w which refers to K elementary events to a given problem ($\Theta = \{w_1, w_2, ..., w_K\}$). The power set of Θ is the set of all the 2^K possible subsets such that:

$$2^\Theta = \{\emptyset, \{w_1\}, \{w_2\}, ..., \{w_K\}, \{w_1, w_2\}, ..., \Theta\} \qquad (1)$$

The key point of Dempster-Shafer theory is the basic belief assignment (bba) which represents the partial knowledge about the value of w and defined as follows:

$$\begin{aligned} m : 2^\Theta &\to [0,1] \\ A &\mapsto m(A) \end{aligned} \qquad (2)$$

where m satisfies the following constraint:

$$\sum_{A \subseteq \Theta} m(A) = 1 \qquad (3)$$

An element A of Θ is called a *focal element* when $m(A) > 0$, and the set containing all these elements is called a *body of evidence* (BOE). When each element in BOE is a singleton, m is named a *Bayesian bba*. On the other hand, when BOE contains only Θ as a focal element, we are in the complete ignorance situation and m is called *vacuous belief function*. However, when it contains only one singleton of Θ as a focal element, m is presented as a *Certain mass function*.

A bba function is normalized when the mass given to the empty set is constrained to be zero ($m(\emptyset) = 0$). In that case, it corresponds to the *closed-world assumption* [17]. A contrary explanation is that the frame of discernment Θ can be incomplete and the value of w can be taken outer Θ. Accordingly, the mass of belief that is not linked to Θ can allowed to be strictly positive ($m(\emptyset) > 0$). That case corresponds to the *open world assumption* [19].

If we are faced to distinct pieces of evidence (e.g., m_1 and m_2), it is important, also, to show some ways in the literature of *bba* combination. The most popular combination rule which is proposed by [16] uses the conjunctive sum operation $\bigcirc\mkern-15mu\cap$ and defined as follows:

$$(m_1 \bigcirc\mkern-15mu\cap\, m_2)(A) = \sum_{B \cap C = A} m_1(B)\ m_2(C) \qquad \forall A \subseteq \Theta \quad (4)$$

However, the normality constraint ($m(\emptyset) = 0$) can be recovered by using Dempster's rule of combination where $\oplus$ represents the notation of its resulting operation such that:

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

where $\kappa = \sum_{A \cap B = \emptyset} m_1(A)\ m_2(B)$ is commonly known as the combination global conflict.

Ultimately, several solutions have been proposed in the frame of belief function theory in order to guarantee the decision making process. For instance, we cite the *pignistic probability*, denoted $BetP$, which is offered by Transferable Belief Model (TBM) and generally considered as a good way for decision making. If the mass function is normalized, then $BetP$ is defined as follows:

$$BetP(w) = \sum_{w \in A} \frac{m(A)}{|A|} \qquad \forall w \in \Theta \qquad (6)$$

If the mass function is unnormalized ($m(\emptyset) > 0$), then the pignistic transformation must be preceded by a normalization step.

B. Evidential c-means (ECM)

The Evidential c-means (ECM) algorithm is an evidential clustering technique proposed in [21], based essentially on Fuzzy c-means (FCM) algorithm [22] and generalizes both of the hard k-means and FCM. The aim of this technique is to assign each object with degrees of belief to the different subsets of clusters from the frame of discernment.

In ECM algorithm, like FCM, each cluster w_k is presented by its center $\boldsymbol{v_k}$ which is a vector defined in object attribute space. However, unlike FCM, one case can belong not only to a singleton cluster but also to a partition of clusters ($A_j \subseteq \Theta$) that can be called a *meta-cluster* and having a cardinality superior than one ($|A_j| > 1$). Correspondingly, the meta-cluster A_j is also represented by a prototype denoted $\overline{\boldsymbol{v_j}}$ and defined as follows:

$$\overline{\boldsymbol{v_j}} = \frac{1}{|A_j|} \sum_{k=1}^{K} s_{kj} \boldsymbol{v_k} \qquad (7)$$

where K represents the number of clusters, $s_{kj} = 1$ if $w_k \in A_j$ and $s_{kj} = 0$ otherwise.

As almost of the clustering techniques, the purpose is to maximize the distances of objects belonging to different clusters and minimize those belonging to the same one. Hence, in the evidential framework, ECM applies the same principle by minimizing the following objective function for n objects and K clusters:

$$J_{ECM}(M,V) = \sum_{i=1}^{n} \sum_{j/A_j \neq \emptyset, A_j \subseteq \Theta} |A_j|^{\alpha} m_{ij}^{\beta} d_{ij}^{2} + \sum_{i=1}^{n} \delta^2 m_{i\emptyset}^{\beta} \qquad (8)$$

subject to

$$\sum_{j/A_j \subseteq \Theta, A_j \neq \emptyset} m_{ij} + m_{i\emptyset} = 1 \qquad \forall i = 1...n \qquad (9)$$

where M represents the credal partition defined in $\mathbb{R}^{n \times 2^K}$ space, V is the matrix of 2^K clusters centers having p features, m_{ij} denotes $m_i(A_j)$ and d_{ij} indicates the euclidean distance between the i^{th} object and the j^{th} partition's prototype. For α, it consists of controlling the degree of penalization for subsets with high cardinality. Finally, β and δ present two parameters for treating noisy objects.

In order to achieve this minimization, an alternation between two phases is applied. The first one consists of supposing that V is fixed and solving Equation 8 constrained by Equation 9 using the Lagrangian where the calculation details are presented in [21]. Thus, the resulting optimum of M for every partition $A_j \subseteq \Theta$ is defined as follows:

$$m_{ij} = \frac{|A_j|^{-\alpha/(\beta-1)} d_{ij}^{-2/(\beta-1)}}{\sum_{A_k \neq \emptyset} |A_k|^{-\alpha/(\beta-1)} d_{ik}^{-2/(\beta-1)} + \delta^{-2/(\beta-1)}} \qquad (10)$$

and

$$m_{i\emptyset} = 1 - \sum_{A_j \neq \emptyset} m_{ij} \qquad \forall i = 1...n \qquad (11)$$

However, for the second phase, we consider M as fixed and an unconstrained minimization problem has to be solved (Equation 8). After the calculation sequence as shown in [21], the resulting cluster centers matrix V comes from the resolution of the system $HV = B$, where B is a $(K \times p)$ matrix and H is a square $(K \times K)$ matrix as they are defined in [21]. with the following definition:

$$B_{lq} = \sum_{i=1}^{n} x_{iq} \sum_{A_j \ni w_l} |A_j|^{\alpha-1} m_{ij}^{\beta} \qquad l = 1...K \qquad q = 1..p \qquad (12)$$

and H is a square $(K \times K)$ matrix having the following form:

$$H_{lk} = \sum_{i} \sum_{A_j \supseteq \{w_k, w_l\}} |A_j|^{\alpha-2} m_{ij}^{\beta} \qquad k,l = 1...K \qquad (13)$$

IV. ECTD APPROACH

Fundamentally, our purpose is to maintain case bases for CBR systems. Especially, we aim to reduce their case bases where uncertainty is handled in order to retain or rather improve as well as possible their competence and performance in problem resolution. To do that, we developed a case base maintenance policy named ECTD for Evidential Clustering and case Types Detection for case base maintenance.

Our ECTD approach goes through three main steps as shown in Figure 1. First, it uses an evidential clustering technique in order to assign cases with a degree of belief to clusters or also to partitions of clusters since in the evidential frame clusters are overlapping. Consequently, the uncertainty regarding the membership of cases to the different clusters is well handled. Besides, a large case base can be treated in the form of a number of small ones. Second, and before applying the maintenance, ECTD proposes to partition cases into four types according to their states and positions towards the different clusters and according to their competence towards the entire case base. As presented in Figure 1, ECTD imposes to a case to be even Noisy, Similar, Isolated or Internal. Finally, the maintenance is achieved by removing cases associated to undesirable types which are Noisy and Similar.

Hereafter, each step will be detailed independently in order to obtain at the end a general depict of our ECTD approach.

A. First step: Evidential Clustering

During this step, we are supposed to perform an evidential clustering which consists to use the belief function theory in order to handle uncertainty about cases assignment to clusters, where each case can belong to all clusters with a degree of belief. Actually, the evidential clustering of objects which are defined as cases in our frame, bestows a credal partition that allows a case to be assigned to multiple clusters, or rather multiple partitions of clusters which makes it more general than the other theories managing uncertainty.

In this context, the evidential clustering technique used in

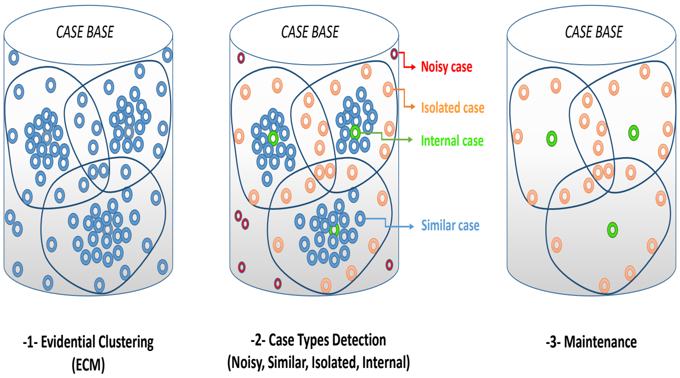
CASE BASE
CASE BASE
Noisy case
Isolated case
Internal case
Similar case
CASE BASE
-1- Evidential Clustering (ECM)
-2- Case Types Detection (Noisy, Similar, Isolated, Internal)
-3- Maintenance

Fig. 1. Steps of our ECTD approach

order to generate the credal partition for our Case Base is the Evidential c-means (ECM) [21] such as it was presented in subsection III.*B*. By this way, ECM presents the first step of our ECTD approach for Case Base Maintenance.
In fact, the output generated by ECM after the convergence of the Equation 8 will be exploited during the next step in order to distinguish between the different case types. This output is represented in the credal partition M as shown in Equations 10 and 11 which allow to handle uncertainty by offering the degrees of belief concerning the assignment of cases to the different subsets of clusters where these subsets or partitions are presented by their centers exposed in the matrix V that is also generated by ECM.

B. Second step: Case Types Detection

Drawing on the three case types presented in [25], we define our four types (Noisy, Similar, Isolated and Internal) to which we assign our own definitions. First of all, we are based on the two basic definitions of cases *coverage* and *reachability* [3] to exhibit the two following definitions that will argue the different case types distinction and explain the adopted way for maintenance:

Definition IV.1. A case $\boldsymbol{x_i}$ is said reachable by the case base if there are a set of cases that are very close to it.

Definition IV.2. Given a subset of cases which are very close to each others, each case of them can cover all the others.

1) ***Noisy*** *cases detection:* Remembering that our first step has been to apply the clustering technique "Evidential c-means". Therefore, it generates for each case a degree of assignment to the different partitions of clusters. The idea here for handling Noisy cases is identical to that used in [28] which consisting to allocate a cluster to which we assign noises. By this way, the empty set partition which corresponds to the "open world assumption" is the cluster that will deal in our context with cases representing a distortion of values and cannot be assigned to any one among the set of clusters. Therefore, ECTD considers that cases having a *"High"* degree of belief to be assigned to the empty set partition are flagged as Noisy. For ECTD approach, a degree of assignment is said *"High"* if and only if it is greater than the sum of all the other degrees of assignments to other partitions. In other words, a case is flagged as Noisy if there is more belief that this case does not belong to any cluster in the frame of discernment Θ than it does. Regardless of that, noisy objects are generally characterized by their remoteness and isolation to each objects gathering or clustering which argues and defends our assumption about Noisy cases definition. Therefore, ECTD defines Noisy cases as follows:

$$\boldsymbol{x_i} \in NC \quad iff \quad m_i(\emptyset) > \sum_{A_j \subseteq \Theta, A_j \neq \emptyset} m_i(A_j) \qquad (14)$$

where $\boldsymbol{x_i}$ is an instance of cases and NC represents the set of all the Noisy cases.
Since the total sum of beliefs towards cases assignment to the different partitions of clusters is constrained to be one (Equation 3), a Noisy case can likewise be defined as follows:

$$\boldsymbol{x_i} \in NC \quad iff \quad m_i(\emptyset) > 0.5 \qquad (15)$$

Fundamentally, this type of cases represents a burden regarding the entire case base. In fact, they cannot cover neither other cases nor themselves. Besides, they cannot be reachable by others i.e. no other case can solve them. Consequently, this type of cases are generally reducing the case base's competence.

2) ***Similar*** *and* ***Isolated*** *cases detection:* Once ECTD detects Noisy cases, it has now to distinguish between two types of cases from the remaining case base which are named Similar and Isolated. This distinction will be done after handling the different distances to clusters centers. Rationally, the clustering algorithm ECM as presented in subsection III.*B* makes that the most of cases are located in the centers of clusters. Therefore, the idea consists of measuring the different distances to different clusters prototypes presenting their centers. Hence, cases having a *"Large"* distance from centers are flagged as Isolated and those having a *"Short"* distances to one cluster's center is flagged as Similar towards this cluster. This description can be better clarified by Figure 1. However, the question herein is: *"How handling distances between cases and clusters centers within an uncertain framework where we should take advantages of the credal partition of cases to all partitions of clusters and not only singleton clusters?"*
The idea of ECTD herein is to exploit as well as possible the cases assignments degrees to the different clusters that are already provided by ECM algorithm in form of bba functions. As mentioned, we are therefore in charge to calculate cases distances in order to distinguish between those which are situated in clusters' cores and those that are more distant while managing uncertainty. To do that, we adapted the Mahalanobis Distance [30] to the belief function theory by using the Belief Covariance Matrix [27]. Hence, we called this distance as *Belief Mahalanobis Distance* (BMD). Conspicuously, this distance has many strengths such that:

- It is appropriate with non-uniform distributions and able to support arbitrary shapes of clusters, and not only spherical as it is the case with the Euclidean one.
- It takes into account the covariance between variables during distances computing.
- It manages so well the uncertainty regarding the membership of cases not only to many clusters but also to many partitions of clusters. Basically, this ability is present thanks to the *Belief Covariance Matrix* Σ which measures the cluster's covariance matrix in p-dimensional space by returning to all the partitions of clusters in which it belongs.

In fact, for a case base containing n instances of multivariate cases $\boldsymbol{x_i}$ defined in p-dimensional space, the *Belief Mahalanobis Distance* (BMD) to a given cluster is defined as follows:

$$BMD(\boldsymbol{x_i}, \boldsymbol{v_k}) = \sqrt{(\boldsymbol{x_i} - \boldsymbol{v_k})^T \Sigma_k^{-1} (\boldsymbol{x_i} - \boldsymbol{v_k})} \tag{16}$$

where $\boldsymbol{v_k}$ is the center of the k^{th} cluster generated during the first step by ECM algorithm whereas Σ_k represents the *Belief Covariance Matrix* [27] of the k^{th} cluster and it has the following form:

$$\Sigma_k = \sum_{i=1}^{n} \sum_{A_j \ni w_k} m_{ij}^2 |A_j|^{\alpha-1} (\boldsymbol{x_i} - \overline{\boldsymbol{v_j}})(\boldsymbol{x_i} - \overline{\boldsymbol{v_j}})^T \tag{17}$$

where A_j is a partition of clusters ($A_j \subseteq \Theta$) with $j = 1, .., 2^K$, k is the cluster's number with $k = 1, .., K$, m_{ij} and $\overline{\boldsymbol{v_j}}$ are respectively the credal partition and the prototypes as defined during the first step by ECM, and by fixing the value of the exponent α, $|A_j|^\alpha$ serves to penalize the belief's allocation to partitions with high cardinality. Remarkably, this belief covariance matrix [27] of the k^{th} cluster exploits the partitions prototypes given during the first step by ECM as well as the credal partition of cases to all subsets containing the cluster k as a focal element in order to well estimate cases dispersion aroud this cluster's center.

After calculating the distance matrix of n cases regarding K clusters, the aim now is to fix a threshold with which we will compare these distances in order to decide if it consists of Similar or Isolated cases. For this reason, we first exclude Noisy cases. Then, we compute the threshold of each cluster as the mean of distances towards the center of this cluster. Therefore, ECTD defines this threshold as follows:

$$Threshold_k = \frac{\sum_{\boldsymbol{x_i} \notin NC} BMD(\boldsymbol{x_i}, \boldsymbol{v_k})}{\#TotalCases - \#NoisyCases} \tag{18}$$

In fact the intuition behind this proposal is that it indicates how much on average a case is close to the center of the distribution and to compare with it. Besides, we exclude Noisy cases because generally the mean as well as the standard deviation are so sensitive to noisy values. So it can affect seriously the results.

Finally, by taking only cases that have not been flagged as Noisy, we are now able to distinguish between Similar and Isolated cases by using the following form:

$$\boldsymbol{x_i} \in \begin{cases} SC_k & if\ \exists k / BMD(\boldsymbol{x_i}, \boldsymbol{v_k}) < Threshold_k \\ IsC & Otherwise \end{cases} \tag{19}$$

where SC_k represents the set of similar cases which are situated near to the core of cluster k (*"Short"* distance) and IsC is the set containing Isolated cases which are more distant (*"Large"* distance).

*3) **Internal** cases detection:* By attaining this phase, we are already flagged each case as one of the following three types: Noisy, Similar or Isolated. However, Similar cases regarding the same cluster are seen as redundant cases. So, we have to vote for only one case for each cluster in order to cover the others after their deletion from the case base. Consequently, our approach chooses to vote for the closest case to each cluster's center and re-flag it as an Internal case. Logically, we obtain finally a number of Internal cases equal to the number of the initial clusters K. Formally, we can define Internal cases such that:

$$\boldsymbol{x_i} \in InC \quad iff \quad \exists k; \neg \exists \boldsymbol{x_j} / BMD(\boldsymbol{x_j}, \boldsymbol{v_k}) < BMD(\boldsymbol{x_i}, \boldsymbol{v_k}) \tag{20}$$

where InC represents the set of Internal cases, $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ are two instances of cases and $\boldsymbol{v_k}$ is the center of the cluster k.

C. Third step: Maintenance

The previous elaborated steps concerning the evidential clustering and the detection of the different case types aim at applying effectively the maintenance task. In general, maintenance of CBR systems and especially of case bases may appear in different forms such as removing or updating a number of cases. Indeed, ECTD is interested in cases elimination. Therefore, it consists foremost of deleting cases representing a distortion of values that can lead to the decrease of CBR systems' ability in problem solving. This first type corresponds to cases flagged as *Noisy*. The second type of cases that have to be removed is *Similar* since this kind of cases can be seen as redundant cases and can be covered by only one case that is already exposed as *Internal*. Hence, the main motivation for removing this type of cases is to alleviate the case base and to get a better response time surely without reducing its competence in problem solving. However, we should keeping *Isolated* cases because they are not reachable by other cases and they can only cover themselves. Besides, their deletion can lead to the risk that some problems can be permanently unsolvable by the case base. Moreover, *Internal* cases must be retained because they cover all the deleted cases that were situated near to the different clusters centers.

V. Experimentation and results

For the experimental investigation regarding our new ECTD approach, we developed it using Matlab R2015a and it is tested on six case bases from U.C.I Repository for machine learning data sets with considering only numeric data. These data sets description including their references, number of instances and number of attributes are shown in Table I. While developing, we fixed the initial number of clusters to the original number of classes given in U.C.I Repository and we considered the default parameters of ECM algorithm. Besides, we did not penalize the belief's allocation to partitions with high cardinality by fixing the exponent of penalization α to one.

TABLE I
CASE BASES DESCRIPTION

	Case base	Ref	# instances	# attributes
1	Glass	GL	214	10
2	Heberman	HB	306	3
3	Iris	IR	150	4
4	Ionosphere	IO	351	34
5	BankNote Authentification	BN	1372	5
6	Phishing Websites	PW	2456	30

Taking in mind that the main purpose of our ECTD approach is to maintain case bases with preserving or even improving their performance and competence during problem resolution, we propose therefore to measure the ECTD efficiency according to the three following criteria and compare the results with the initial non-maintained case bases that we will call the Initial CBR (ICBR).

TABLE II
STORAGE SIZE [S(%)]

Case bases		Storage size [S(%)]	
		ICBR	ECTD
1	GL	100 %	50 %
2	HB	100 %	39.87 %
3	IR	100 %	38.67 %
4	IO	100 %	41.03 %
5	BN	100 %	35.86 %
6	PW	100 %	41.26 %

TABLE III
ACCURACY [PCC(%)]

Case bases		Accuracy [PCC(%)]	
		ICBR	ECTD
1	GL	86.92 %	94.39 %
2	HB	74.18 %	76.23 %
3	IR	98 %	98.28 %
4	IO	86.89 %	87.5 %
5	BN	97.45 %	97.56 %
6	PW	92.18 %	94.17 %

A. Evaluation criteria

- **Storage size [S(%)]:** is the percentage of the remaining case base after maintenance vis-a-vis the initial one. Hence, it consists of the reduction size rate. Actually, it is also used in other CBM approaches such as [14] and [15]. The more the storage size (S%) is reduced, the better maintenance is achieved. Thusly, this criterion is defined such that:

$$S = \frac{Final\ case\ base\ size}{Initial\ training\ case\ base\ size} \times 100 \qquad (21)$$

TABLE IV
RETRIEVAL TIME [T(S)]

Case bases		Retrieval time [T(s)]	
		ICBR	ECTD
1	GL	0.0069	0.0062
2	HB	0.2825	0.0133
3	IR	0.0077	0.0077
4	IO	0.0836	0.0162
5	BN	0.0272	0.0052
6	PW	3.5330	0.6362

- **Accuracy [PCC(%)]:** This criterion refers to the famous classification evaluation measure called Percent of Correct Classification (PCC) and it is generally exposed as a percentage like it is marked within this following expression:

$$PCC = \frac{\#\ well\ classified\ instances}{\#\ total\ classified\ instances} \times 100 \qquad (22)$$

  To calculate the PCC value in our context and comparing between the accuracy of the initial case base and the case base after maintenance, we choose the 1-Nearest Neighbor (1-NN) as a classification algorithm. Then, we performed the method of 10-Fold Cross Validation.
- **Retrieval time [T(s)]:** Since the performance of CBR systems is strongly linked to the time of problem resolution, we choose this criterion as an important one and we applied it around the algorithm 1-NN to measure the classification duration in seconds.

B. Results and discussion

In terms of reduction size, we observe from the results shown in Table II that our ECTD approach has been able to reduce more than half the size of almost all the tested case bases comparing to the initial case bases which contains the totally of cases instances. In fact, for all the tested case bases, ECTD keeps between about 35% and 50 % of cases instances comparing to the Initial CBR (ICBR) with 100%. However, it is still necessary to ascertain their competence towards problem solving after maintenance. Here, we are talking about the accuracy criterion. Similarly, we are faced to an amelioration for all the tested case bases in terms of problem solving or accuracy where the percentage of cases correctly classified by 1-NN is better than that offered by the initial CBR systems as it is shown in Table III. Actually, the accuracy values provided by ECTD are varying between 76.23% for *"Heberman"* data set and 98.28% for *"Iris"* data set. However, for ICBR, their values are varying between 74.18% and 98% for the same data sets. Rationally, this slightly accuracy amelioration has occurred thanks to the deletion of noisy cases which mostly affecting case bases competence. Finally, the reduction of the retrieval time values such that presented in Table IV is an expected results since the number of case bases have been reduced by deleting several cases instances. For example, the retrieval time with the data set *"Phishing website"* passes from about 3.5 seconds to 0.6 seconds since ECTD keeps only about 41% from the initial case base size.

VI. Conclusion

In this paper, we presented a new case base maintenance approach able to manage uncertainty in cases called ECTD. Our new method aims to maintain case bases for CBR systems by removing all the cases representing a burden for the case base. The idea is summed up by applying the ECM [21] as an evidential clustering technique on the case base and exploiting the generated bbas corresponding to the different partitions as well as possible in order to distinguish between four types of cases: Noisy, Similar, Isolated and Internal cases. Finally, the case base maintenance is achieved by deleting Noisy cases in order to ameliorate the case base's competence and Similar cases so as to alleviate the case base. As future work, experimentation can carried out using a case study.

References

[1] A. Aamodt and E. Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. In Artificial Intelligence Communications, pages 39-52, 1994.

[2] D. C. Wilson and D. B. Leake. Maintaining case-based reasoners: Dimensions and directions. In Computational Intelligence, pages 196-213, 2001.

[3] B. Smyth and M. T. Keane. Remembering to forget. In Proceedings of the 14th international joint conference on Artificial intelligence, pages 377-382, 1995.

[4] B. Smyth and E. McKenna. Competence models and the maintenance problem. In Computational Intelligence 17, pages 235-249, 2001.

[5] P. Hart. The condensed nearest neighbor rule (Corresp.). IEEE transactions on information theory, pages 515-516, 1962.

[6] W. Gates. The Reduced Nearest Neighbor Rule. In IEEE Transactions on Information Theory, pages 431-433, 1972.

[7] G. Ritter, H. Woodruff, S. Lowry and T.Isenhour. An algorithm for a selective nearest neighbor decision rule. In IEEE Transactions on Information Theory, pages 665-669, 1975.

[8] H. Zhao, Hui, L. Wang, W. Dong, X. Sun, and Y. Ji. A rule based case maintenance method for the performance of CBR classifier. In Control and Decision Conference (CCDC), pages 4174-4179, 2016.

[9] A. Smiti and Z. Elouedi. Overview of Maintenance for Case based Reasoning Systems, Int. J. Comput. Appl, pages 49-56, 2011.

[10] S. Minton. Quantitative results concerning the utility of explanation-based learning. In Artificial Intelligence 42, pages 363-391, 1990.

[11] H. Brighton and C. Mellish. On the consistency of information filters for lazy learning algorithms. In European conference on principles of data mining and knowledge discovery, pages 283-288, 1999.

[12] A. Smiti and Z. Elouedi. COID: Maintaining case method based on Clustering, Outliers and Internal Detection. In Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, pages 39-52, 2010.

[13] F. Sander, M. Ester, and P. Kriegelh. The algorithm GDBSCAN and its application. In Data Mining and Knowledge Discovery, pages 178-192, 1998.

[14] A. Smiti and Z. Elouedi. WCOID: Maintaining Case-Based Reasoning systems using Weighting, Clustering, Outliers and Internal cases Detection. In the 11th International Conference on Intelligent Systems Design and Applications (ISDA), pages 356-361. IEEE Computer Society, 2011.

[15] A. Smiti and Z. Elouedi. WCOID-DG: An approach for case base maintenance based on Weighting, Clustering, Outliers, Internal Detection and Dbsan-Gmeans. Journal of computer and system sciences, pages 27-38, 2014.

[16] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. The annals of mathematical statistics, pages 325-339, 1967.

[17] G. Shafer. A mathematical theory of evidence. Vol. 1. Princeton: Princeton university press, 1976.

[18] P. Smets. The transferable belief model for quantified belief representation. In Quantified Representation of Uncertainty and Imprecision, pages 267-301. Springer Netherlands, 1998.

[19] P. Smets. The combination of evidence in the transferable belief model. IEEE Transactions on pattern analysis and machine intelligence. pages 447-458, 1990.

[20] T. Denœux, S. Sriboonchitta, and O. Kanjanatarakul. Evidential clustering of large dissimilarity data. Knowledge-Based Systems, pages 179-195, 2016.

[21] M. H. Masson and T. Denoeux. ECM: An evidential version of the fuzzy c-means algorithm. Pattern Recognition 41, pages 1384-1397, 2008.

[22] J. C. Bezdek, R. Ehrlich and W. Full. FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences 10, pages 191-203, 1984.

[23] S. K. Shiu, C. H. Sun and X. Wang and D. S. Yeung. Maintaining Case-Based Reasoning Systems Using Fuzzy Decision Trees. In European workshop on advances in case-based reasoning (EWCBR), pages 285-296, 2000.

[24] G. Cao, Simon C. K. Shiu and Xizhao Wang, A Fuzzy-Rough Approach for Case Base Maintenance. In European workshop on advances in case-based reasoning (EWCBR), pages 118-130, 2001.

[25] A. Smiti and Z. Elouedi. Maintaining Case Based Reasoning Systems Based on Soft Competence Model. In International Conference on Hybrid Artificial Intelligence Systems, pages 666-677. Springer International Publishing, 2014.

[26] A. Smiti and Z. Elouedi. Fuzzy density based clustering method: Soft DBSCAN-GM. In 8th International Conference on Intelligent Systems (IS), pages 443-448. IEEE, 2016.

[27] V. Antoine, B. Quost, H.M. Masson and T. Denœux. CECM: Constrained evidential c-means algorithm. Computational Statistics & Data Analysis, pages 894-914, 2012.

[28] R. N. Dave. Characterization and detection of noise in clustering. Pattern Recognition Letters, pages 657-664, 1992.

[29] L.A. Zadeh. Fuzzy sets. Information and control 8, pages 338-353, 1965.

[30] P. C. Mahalanobis. Mahalanobis distance. In Proceedings National Institute of Science of India, pages 234-256, 1936.