CEC-Model: A new competence model for CBR systems based on the belief function theory

Safa Ben Ayed[1,2], Zied Elouedi[1], and Eric Lefèvre[2]

[1] LARODEC, Institut Supérieur de Gestion de Tunis, Université de Tunis, Tunisie,
safa.ben.ayed@hotmail.fr, zied.elouedi@gmx.fr
[2] Univ. Artois, EA 3926, LGI2A, 62400 Béthune, France,
eric.lefevre@univ-artois.fr

Abstract. The high influence of case bases quality on Case-Based Reasoning success gives birth to an important study on cases competence for problems resolution. The competence of a case base (CB), which presents the range of problems that it can successfully solve, depends on various factors such as the CB size and density. Besides, it is not obvious to specify the exactly relationship between the individual and the overall cases competence. Hence, numerous Competence Models have been proposed to evaluate CBs and predict their actual coverage and competence on problem-solving. However, to the best of our knowledge, all of them are totally neglecting the uncertain aspect of information which is widely presented in cases since they involve real world situations. Therefore, this paper presents a new competence model called *CEC-Model (Coverage & Evidential Clustering based Model)* which manages uncertainty during both of cases clustering and similarity measurement using a powerful tool called the belief function theory.

Keywords: case-based reasoning, competence model, cases coverage, belief function theory, uncertainty, clustering

1 Introduction

Among the main concerns within the knowledge engineering field is to offer techniques aiming to assess informational resources. In particular, the community of Case-Based Reasoning (CBR) provides a specific interest to evaluate case bases since their quality presents the key factor's success of CBR systems. In fact, the higher the quality of this knowledge container, the more "competent" it is. Actually, the competence (or coverage) of a CBR system refers to its cabability to solve target problems. That's why, the notion of *case competence* is widely used, also, within the field of Case Base Maintenance (CBM), where most of the CBM policies ([8], [16], [17], [18], [19], etc.) do their best to maintain the most competent cases. However, this key evaluation criterion is difficult to predict since the true character of competence within CBR as well as its sources are not well comprehensible [1]. Moreover, even if we could estimate the competence of an individual case, the estimation of the global case base competence remains

complex because of the lack of clarity towards the relationship between local and global competence contribution. By this way, we find several research, over the years, that are interested on case base competence notion, where some of them offer case competence models for CBs evaluation. Typically, case competence models divide cases into competence groups, then estimate cases coverage using similarity measures. Their theoretical contributions are obviously well defended. However, the embedded imperfection in cases was totally neglected within this area, especially that each case refers to one real world experience. Evidently, events and situations occured within our world are full of uncertainty and imprecision. Therefore, we propose, in this paper, a new case competence model, called *CEC-Model* encoding "Coverage & Evidential Clustering based Model", that aims to accurately evaluate the overall case base coverage using the belief function theory [2] [3]. This theory offers all the necessary tools to manage all the levels of uncertainty in cases. Through Fig. 1, it is straighforward to show the different fields intersection leading to build and construct our new competence model. In a nutshell, CEC-Model divides the case base into groups using the evidential clustering technique called ECM [6]. Then, it uses a distance within the belief functions framework that leads, ultimately, to estimate the global coverage of the case base. Like the competence model on which we are based [1] to estimate the relation between local case competence and global CB competence, our CEC-Model makes some assumption; First, we assume that the set of cases in the CB presents a representative sample of the set of target problems. Second, we assume that the problem space is regular, where we are based on the CBR hypothesis "Similar problems have similar solutions".

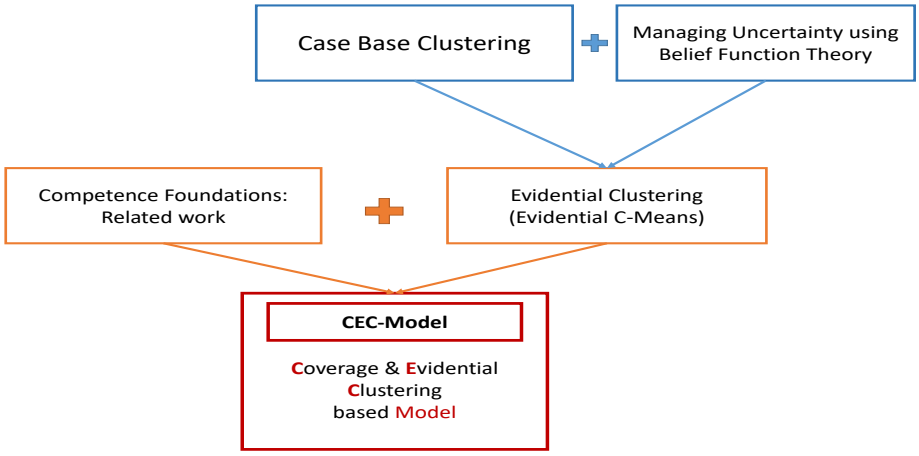

Fig. 1: Towards CEC-model

The reminder of this paper is organized as follows. In section 2, we overview the related work of the Competence concept by offering its foundation, defining

the basic factors affecting the case base competence, and presenting some competence models. Section 3 offers the basic concepts of the belief function theory as well as the used evidential tools for building our model. Throughout Section 4, our new CEC-Model is described in details through its different steps. Finally, our model is supported in Section 5 using an experimental analysis.

2 General Outlook on Case-Base Competence

A case base is said to be "effective" when it is able to offer solutions efficiently and successfully to solve as many target problems as possible. To evaluate the case base effectiveness for CBR systems, two criteria are generally used: Performance (Definition 1) and Competence (Definition 2).

Definition 1. *The Performance is the answer time that is necessary to generate a solution to a target problem.*

Definition 2. *The Competence is the range of target problems that can be successfully solved.*

Contrary to the competence, the performance criterion for a case base can be straightforward measured. Hence, we will focus, in the following of this Section, on the competence criterion by presenting its foundations (Subsection 2.1), its influencing factors (Subsection 2.2) and some existing models to predict the overall case base competence (Subsection 2.3).

2.1 Case Competence Foundations

When we talk about case competence, two main concepts arise: case Coverage (Definition 3) and case Reachability (Definition 4).

Definition 3. *The coverage of one case is the set of target problems that this case is able to solve. It is defined formally as follows [8]:*

$$CB = \{c_1, .., c_n\}, c \in CB, Coverage(c) = \{c' \in CB / Solves(c, c')\} \quad (1)$$

where CB presents the case base and $Solves(c, c')$ is the fact that the case c is able to solve the case c'.

Definition 4. *The reachability of a target problem is the set of cases that can be used to solve it. It is defined such that [8]:*

$$CB = \{c_1, .., c_n\}, c \in CB, Reachability(c) = \{c' \in CB / Solves(c', c)\} \quad (2)$$

These two latter definitions are based on the assumption that the case base presents the representative sample of all target problems. In fact, it is impossible

in the reality to define and fix the entire set of all the target problems. Besides, in that step, we are not intended to explicitly define the predicate "Solves".

For the sake of clarity regarding Definitions 3 and 4, we illustrate in Fig. 2 an example. Let c_1, c_2 and c_3 three cases, and their coverage are labeled with $1, 2$ and 3 respectively. Therefore, $Coverage(c_2) = \{c_2, c_3\}$ and $Reachability(c_2) = \{c_2, c_1\}$. Logically, we assign more interest to cases having a large coverage an

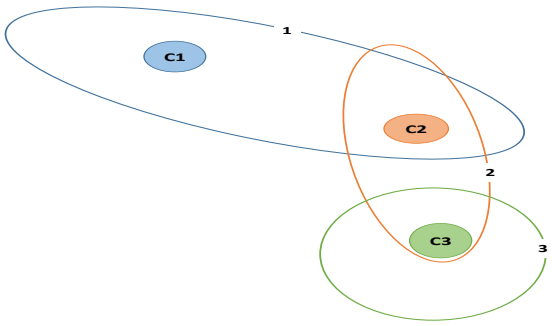

Fig. 2: Concepts of cases Coverage & Reachability

a small rechability set. In this paper, we restrict the competence of the case base to its overall coverage. However, the interaction between local competences is necessary to well estimate the entire case base's coverage. Moreover, several factors can influence the prediction of this criterion.

2.2 Basic factors influencing case base competence

Building an appropriate competence model requires an awareness on the different factors influencing CB's competence as well as understanding how they affect it. Actually, several factors have been studied in the literature. On the one hand, we find statistical properties such that the CB's size, distribution and density of cases [7] [9] [10]. On the other hand, the competence is naturally relied to the problem solving properties such as vocabulary, similarity and adaptation knowledge [11] [12], as well as individual cases coverage [1] [7] [13].
Similarly to some other research [1] [9], we focus in this paper to understand and measure this competence through case base size and density factors as well as the coverage concept.

2.3 Case Competence Modeling

In the literature, various ways are proposed to model cases competence in order to evaluate the ability of case bases on problem solving. Besides, these models can be the basis of numerous case base maintenance approaches. Hence, we present in what follows three among the most known case competence models.

Model 1: Case Competence Categories [7]: Based on the notions of coverage and reachability, Smyth & Keane classify cases according to their competence characterization into four types, where the following Definitions arise.

Definition 5. *Pivotal cases represent single way to solve a specific problem. They are defined such that:*

$$Pivot(c) \quad iff \quad Reachability(c) - \{c\} = \emptyset \tag{3}$$

Definition 6. *Auxiliary cases are totally subsumed by other cases. They do not influence the global competence at all. Hence, they are defined such that:*

$$\begin{gathered} Auxiliary(c) \quad iff \quad \exists c' \in Reachability(c) - \{c\} / \\ Coverage(c) \subset Coverage(c') \end{gathered} \tag{4}$$

Definition 7. *Spanning cases do not directly influence the CB competence. They link together regions covered by the two previous types of cases (Pivotal and Auxiliary).*

Definition 8. *Support cases exist in groups to support an idea. Each support case in a support group provides the same coverage as the other cases belonging to the same group. They are formally defined such that:*

$$Support(c) \quad iff \quad \exists c' \in Reachability(c) - \{c\} / Coverage(c') \subseteq Coverage(c) \tag{5}$$

For further clarification, and by returning to Fig. 2, we mention according to the four previous Definitions that c_1 represents a Pivotal case, c_2 presents a Spanning case, and c_3 is an Auxiliary case. Concerning Support cases, Fig. 3 illustrates three examples of them that cover the same space.

Model 2: Coverage model based on Mahalanobis Distance and Clustering (CMDC) [14]: Based on the idea that the CB's competence is proportional to individual case's contribution, CMDC defines the overall case base competence as follows:

$$Comp\%(CB) = |1 - \frac{\sum_{j=1}^{K} \sum_{i=1}^{N} Cov(c_{ij})}{SizeCB}| \tag{6}$$

where K is the number of groups building the CB, N is the size of the j^{th} group, and $Cov(c_{ij})$ represents the coverage of case i towards cluster (group) j.

After applying the DBSCAN-GM algorithm [15] for clustering cases belonging to the CB, this model proposes a classification of cases into three types in order to calculate $Cov(c_{ij})$ used in Equation 6. The first type concerns *Noisy*

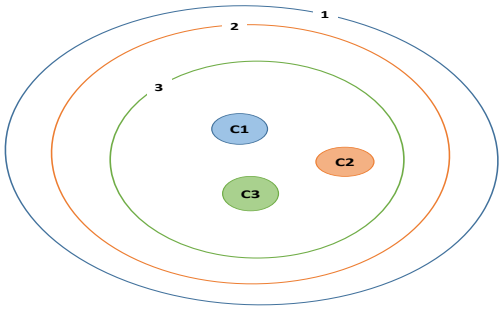

Fig. 3: Example of Support cases towards one Support group

cases where their coverage is null since they are considered as a distortion of values. The second type concerns a closely group of cases existing on the core of one cluster and named *Similar* cases. The coverage of Similar cases is equal to their cardinality within each group. Finally, *Internal* cases represent cases that are situated in the border of each cluster. They only cover themselves and their coverage is equal to one (See Fig. 4).

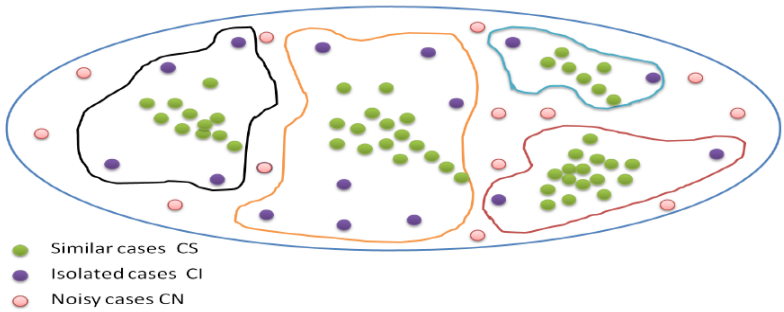

Fig. 4: Cases types defined by CMDC model [14]

Actually, this model was the basis of several policies aiming to maintain case bases for CBR systems, such that [16], [18], [19], etc.

Model 3: Smyth & McKenna (S&M) model [1]: Basically, S&M model tries to estimate the competence by finding and encoding the crucial relationship between individual case (local) and the entire CB (global) competence. To do, S&M identifies the fundamental unit of competence as the *competence group* of cases.In fact, within a very traditional point of view, this unit was "the case" only. To recognize these groups, authors in [1] define a competence group as the

set of cases that have a shared coverage. Formally, this is defined such that:

$$\begin{aligned} For\ G = \{c_1, .., c_p\} \subseteq CB, CompetenceGroup(G)\ \ iff \\ \forall c_i \in G, \exists c_j \in G - \{c_i\} : SharedCoverage(c_i, c_j) \\ \wedge \forall c_k \in CB - G, \neg \exists c_l \in G : SharedCoverage(c_k, c_l) \end{aligned} \tag{7}$$

where

$$\begin{aligned} For\ c_1, c_2 \in CB \\ SharedCoverage(c_1, c_2)\ iff\ Coverage(c_1) \cap Coverage(c_2) \neq \emptyset \end{aligned} \tag{8}$$

To ease the concept of competence group, we indicate using Fig. 2 that the different cases c_1, c_2 and c_3 present only one coverage group since they share their coverage (similarly for Fig. 3).

Furthermore, S&M model allocate a considerable interest to identify Coverage groups since the larger group coverage means a larger ability to solve target problems. By this way, authors in [1] affirm that it is mainly depending on the *size* and *density* of cases. Obviously, the first factor is straightforward calculated. However, the density of cases is defined such that:

$$GroupDensity(G) = \frac{\sum_{c \in G} CaseDensity(c, G)}{|G|} \tag{9}$$

where $CaseDensity(c, G)$ presents the local density of the case c within the group $G \subset CB$, and $|G|$ is the number of cases belonging to G.
Since the coverage of a group must be directly proportional to its size and inversely proportional to its density, the current model defines it as follows:

$$GroupCoverage(G) = 1 + [\ |G|\ (1 - GroupDensity(G))\ ] \tag{10}$$

Undoubtedly, the proposed contributions to model cases competence are interesting. However, they remain limited by their disability to manage uncertainty within knowledge, especially for real experiences (cases). The next Section presents, therefore, a powerful tool used for this matter called the Belief function theory.

3 Belief Function Theory: Basic Concepts

The belief function theory [2] [3], also known by Evidence theory or Dempster-Shafer theory, is a theoretical framework for reasoning under partial and unreliable (uncertain and imprecise) information. It was introduced by Dempster and Shafer [2] [3], and then studied by Smets [4] [5]. As a generalization of other uncertainty management theories [20] [21] [22], belief function theory proved to be effective in various applications. The rest of this Section will recall the main definitions and concepts and the used tools offered within this theory.

Let ω be a variable taking values in a finite set $\Omega = \{w_1, .., w_K\}$ named the frame of discernment. The mass function $m(.)$, which represents the uncertainty and imprecision knowledge about the actual value of ω, is defined as an application from the power set of Ω (2^Ω) in $[0, 1]$ and satisfying

$$\sum_{A \subseteq \Omega} m(A) = 1 \tag{11}$$

Actually, $m(A)$ can be viewed as the degree of belief committed exactly to the subset of events A. A is called focal element if $m(A) > 0$, and the mass function m is equivalent to a probability distribution when all the focal elements are singletons. It is then called Bayesian mass function.

Since two events within the belief function theory are mainly described by their mass functions, it is also interesting to measure the similarity and distance between them. One of the most known and used distances between two pieces of evidence is called the Jousselme Distance of evidence [23].
Given two pieces of evidence m_1 and m_2 on the same frame of discernment, the Jousselme distance between them is defined as follows:

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

where $\underline{\underline{D}}$ is a $2^K \times 2^K$ matrix whose its elements are calculated as follows:

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

To make decision towards the value of ω, the mass function m can be transformed into a pignistic probability distribution $BetP$ [4] such as:

$$BetP(A) = \sum_{B \subseteq \Omega} \frac{|A \cap B|}{|B|} \frac{m(B)}{1 - m(\emptyset)} \quad \forall A \in \Omega \tag{14}$$

Finally, the decision is made by choosing the variable with the highest $BetP$ value.

Concerning the evidential clustering of n objects, the partial knowledge in that time will concern the membership of objects to clusters. Hence, the frame of discernment Ω, in that case, contains the set of all clusters. Basically, an $n \times 2^{|\Omega|}$ credal partition matrix is generated after applying an evidential clustering technique. It offers n mass functions that reflect the membership degrees of belief to each clusters' subset (partition).

The Evidential C-Means (ECM) [6] presents one of the most known evidential clustering techniques. It takes as input the set of n objects and the number K of clusters, and generates as output the credal partition (matrix M) as well as the prototype (center) of each partition (matrix V). Like almost of clustering methods, ECM aims to create dense groups by minimizing distances belonging

to the same cluster and maximize those belonging to different ones. To do, ECM method intend to minimize the following objective function:

$$J_{ECM}(M,V) = \sum_{i=1}^{n} \sum_{j/A_j \neq \emptyset, A_j \subseteq \Omega} |A_j|^{\alpha} m_{ij}^{\beta} d_{ij}^2 + \sum_{i=1}^{n} \delta^2 m_{i\emptyset}^{\beta} \tag{15}$$

subject to

$$\sum_{j/A_j \subseteq \Omega, A_j \neq \emptyset} m_{ij} + m_{i\emptyset} = 1 \qquad \forall i = 1...n \tag{16}$$

where d_{ij} represents the euclidean distance between the object i and the center of the partition j, the parameter α controls the degree of penalization allocated to partitions with high cardinality, and δ and β are two parameters aiming to treat noisy objects.
To minimize the above objective function, an alternation between two steps is performed. The first one consists of supposing that the matrix of centers V is fixed and solving Equation 15 constrained by Equation 16 using the Lagrangian technique. Then, the second phase consists to fix the credal partition M and minimize the unconstrained problem defined only by Equation 15.

During this Section, we only focused on the necessary background within the belief function framework that allow to understand our contribution presented hereafter. More details can be found in [2], [3], [4], [5], [6], and [23].

4 Coverage & Evidential Clustering based Model (CEC-Model)

The purpose of this Section is to present our new case competence model dedicated for this paper. This model is named CEC-Model and able to manage uncertainty within the base knowledge. It also uses the coverage concept as well as the mathematical relation between the competence of a group and the local competence contribution of its individual cases [1] to provide as output a prediction of the global case base competence. Our model can serve, on the one hand, at evaluating the quality of any given case base. On the other hand, it can be the basis for maintaining case bases by finding, for instance, the combination of cases that offer a maximum rate of global competence offered by CEC-Model. For the sake of simplicity, the global process followed by our model to reach its objective in estimating the CB competence rate while managing uncertainty is shown in Fig. 5. First of all, we perform the evidential clustering technique to offer a credal partition of cases that allows to manage uncertainty not only towards the membership of cases to clusters, but also towards their membership to all possible subsets of clusters (partitions). At the second level, the credal partition generated during Step 1, which is a way to model cases membership uncertainty, will be used then, during Step 2, to measure the similarity between cases. Besides, it will be transformed using the pignistic probability (Equation 14), during Step 3, to make the decision about the membership of cases to groups.

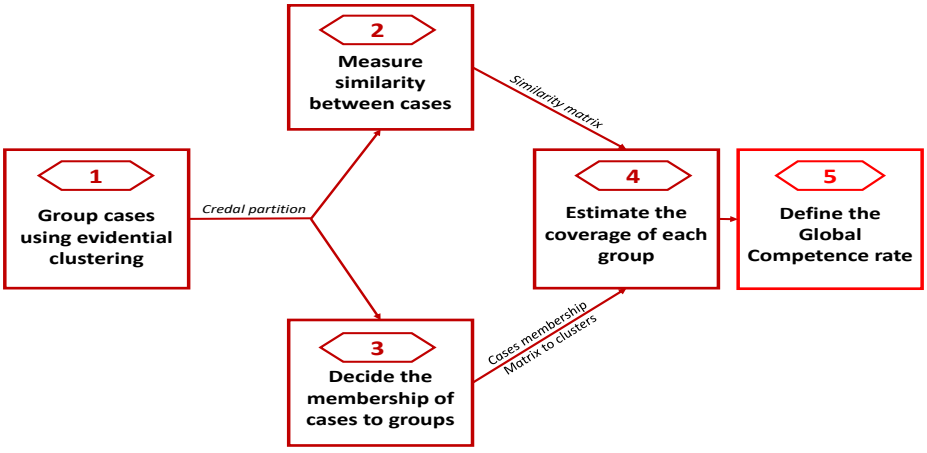

Fig. 5: CEC-Model steps for CB's competence estimation

The outcome of both Steps 2 and 3 will serve then to calculate the individual cases densities regarding their groups, the density rate of each group, and the coverage of the different groups. Finally, the global competence rate of the overall case base defines the purpose of Step 5, where it is estimated by the average of normalized coverages of all groups composing the case base. In what follows, within the reminder of this Section, we will present in more details every step composing our CEC-Model.

4.1 Step 1: Group cases using evidential clustering

In our first Step, we aim to group cases according to their similarities. The more two cases are similar, the more they are able to cover each others. In fact, similarly to several research in competence modeling [1] [14], the idea is that the coverage of one case is defined by the range of cases that are similar to it. Hence, applying a clustering technique based on distances computing offers a simple and reasonable solution to devise the case base into a number of coverage groups. However, the amount of imperfection that is commonly presented in cases knowledge do not allow us to be certain about the membership of cases to the different clusters. For that reason, we make use of the belief function theory and the evidential clustering, more accurately the Evidential C-Means (ECM) technique (see Subsection 3). The idea consists on creating coverage groups with degrees of belief. Finally, the output of this Step is a credal partition containing n pieces of evidence m_i describing the belief's degrees of membership.

4.2 Step 2: Measure similarity between cases within the evidential framework

At this Step, we aim to take advantage of the offered credal partition to measure the similarities between every couple of cases. A case is therefore characterized by

its mass function that defines the membership degrees of belief to every partition of groups. For instance, given three groups G_1, G_2 and G_3, the mass function of case c_i is presented as a vector where it has the following form:

$$\begin{array}{cccc} m_i = [m_i(\emptyset) & m_i(G_1) & m_i(G_2) & m_i(G_1, G_2) \\ m_i(G_3) & m_i(G_1, G_3) & m_i(G_2, G_3) & m_i(\Omega)] \end{array} \tag{17}$$

Let us remind that the sum of all its elements are equal to one.

Now, we have to calculate distances between every two cases through their corresponding pieces of evidence. To do, we choose to use a well known powerful tool within the belief functions community called Jousselme Distance [23], which offers results in $[0, 1]$. Therefore, we build an $n \times n$ distances matrix that we called $CredDist$, where $CredDist(c_i, c_j)$ is the result of Jousselme Distance between m_i and m_j using Equations 12 and 13.
Then, the similarity matrix $CredSim$ is generated as follows:

$$CredSim = Ones - CredDist \tag{18}$$

where $Ones$ is an $n \times n$ matrix filled by 1.

4.3 Step 3: Decide the membership of cases to groups

After computing cases distances with taking into account the uncertainty presented in cases, we move on now from the credal level to the pignistic level where we have to make decision about the membership of cases to the different groups. To do, we transform the mass function of each case to a pignistic probability using Equation 14. Then, we put each case in the group offering the highest pignistic probability value.

4.4 Step 4: Estimate the coverage of each group

The challenge of this step consists on finding the best configuration to model the relationship between a local (individual case) and the global (entire CB) competence contributions. As mentioned in Subsection 2.2, several factors are affecting the interaction between them. For our model, and based on [1], the combination to build the global competence properties of case bases is influenced by two main factors: *Size* and *Density*.
As specified in the Introduction, we assume that the problem space is regular. Hence, cases with high density imply a high degree of mutual similarity. Per contra, sparse cases present low degree of mutual similarity. Consequently, our model calculate the local density of a case c towards the group $G \subseteq CB$ in which it belongs as follows:

$$CaseDensity(c, G) = \frac{\sum_{c' \in G - \{c\}} CredSim(c, c')}{|G| - 1} \tag{19}$$

Afterwards, we calculate the density of each group as the average of all its corresponding cases density using Equation 9.

Ultimately, and based on [1], we define the relationship between the density and the coverage of each group. In fact, dense groups cover smaller target problems space (Density factor). In contrast, groups with higher size cover larger problem space (Size factor). Consequently, we calculate the coverage of each group using Equation 10.

4.5 Step 5: Define the global case base competence rate

Last but not least, we aim at estimating the global competence of case bases based on groups coverage computed during the previous step. In S&M model [1], the global competence is calculated as the sum of all the coverage measurements of groups building the CB. However, we aim in our model to estimate the global competence as a percentage. Then, the proposed global competence rate is calculated as follows:

$$Comp(CB)\% = \frac{\sum_{k=1}^{|\Omega|} GroupCov_n(G_k)}{|\Omega|} \tag{20}$$

where $GroupCov_n(G_k)$ is the normalized coverage of the k^{th} group, defined such that:

$$GroupCov_n(G_k) = \frac{GroupCoverage(G_k) - 1}{|G_k|} \tag{21}$$

The demonstration that gives birth to groups coverage normalization formula (and then the global CB competence rate) is presented as follows:
Let Ω be the frame of discernment containing K groups G_k. $CB = \{c_1, .., c_n\}$ is then divided into $|\Omega|$ groups:

$$\begin{array}{lccl}
\text{We have:} & 0 \leq & CredSim(c, c_i) & \leq 1 \\
0 & \leq & \sum_{c_i \in G - \{c\}} CredSim(c, c_i) & \leq |G| - 1 \\
0 & \leq & \frac{\sum_{c_i \in G - \{c\}} CredSim(c, c_i)}{|G| - 1} & \leq 1 \\
0 & \leq & \boldsymbol{CaseDensity(c, G)} & \leq 1 \\
0 & \leq & \frac{\sum_{c \in G} CaseDensity(c, G)}{|G|} & \leq 1 \\
0 & \leq & \boldsymbol{GroupDensity(G)} & \leq 1 \\
1 & \leq 1 + [|G|(1 - GroupDensity(G))] & & \leq 1 + |G| \\
1 & \leq & \boldsymbol{GroupCoverage(G)} & \leq 1 + |G| \\
0 & \leq & \frac{GroupCoverage(G) - 1}{|G|} & \leq 1 \\
0 & \leq & \boldsymbol{GroupCov_n(G)} & \leq 1 \\
0 & \leq & \sum_{G \in \Omega} GroupCov_n(G) & \leq |\Omega| \\
0 & \leq & \frac{\sum_{G \in \Omega} GroupCov_n(G)}{|\Omega|} & \leq 1 \\
0 & \leq & \boldsymbol{Comp(CB)} & \leq 1
\end{array}$$

5 Experimental Analysis

During the previous Sections, we reviewed the main definitions for competence and coverage modeling, and we proposed a novel model for case bases competence estimation within the frame of belief function theory and evidential

clustering. In this Section, we need to support our model using an empirical evidences. The idea is to demonstrate experimentally that our model competence rate predictions are sufficiently match to the actual competence measurements such as the Percentage of Correct Classification (accuracy). Furthermore, it is more reasonable to define the correlation between their values than focusing on which criterion has the highest values. To start, we present the setup of experimentation. Then, we show how to proceed to support our CEC-Model.

5.1 Experimental Setup

Our CEC-Model algorithm was developed using Matlab R2015a, and tests were performed on real data sets taken from UCI repository [24]. In this paper, we share results offered by three data sets [1] that are described in Table 1. In fact,

Table 1: UCI data sets characteristics

Case base	Attributes	Instances	Classes	Class distribution
Mammographic Mass	6	961	2	516/445
Ionosphere	34	351	2	226/125
Iris	4	150	3	50/50/50

within the context of CBR, attributes are considered as problems description and the class characterizes their solutions. Besides, default values of the ECM algorithm are taken, and the number of clusters is equal to the number of solutions in the CB. Since we will support our model basing on the accuracy criterion, we measure it by applying 10-folds cross validation using the following formula:

$$PCC(\%) = \frac{Number\ of\ correct\ classifications}{Total\ number\ of\ classifications} \times 100 \tag{22}$$

where we used the 1-Nearest Neighbor as a classification method.

5.2 Evaluation criteria

Our experimental study is divided into two parts, where each one carries on one evaluation criterion. Firstly, we are interested to know the correlation between the actual CB's competence (Accuracy) and the estimated global competence rates predicted by our CEC-Model. The different values are the results of a randomly incremental evolution of case bases. Actually, the higher a positive correlation, the more our model is supported. Hence, we measure this correlation using the *Pearson's correlation coefficient* [25] which is bounded between -1 and 1, and defined as follows:

$$r = \frac{\sum_{i=1}^{n}(a_i - \overline{a})(b_i - \overline{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \overline{a})^2}\ \sqrt{\sum_{i=1}^{n}(b_i - \overline{b})^2}} \tag{23}$$

[1] Other CBs are offering similar results but are not presented here due to lack of space.

where a_i (respectively b_i) are the values of the actual CB's competence (respectively the predicted global competence by CEC-Model), and $\overline{a}$ (respectively $\overline{b}$) presents the mean value of a_i (respectively b_i) measurements.

During the second part of our experimentation, we opt to measure the error rate between CEC-Model estimated competence and the PCC values, such that:

$$Error(\%) = \frac{|EstimatedComp - PCC|}{PCC} \times 100 \tag{24}$$

5.3 Results and discussion

For the first part of our experimentation, results are shown in Fig. 6, where the actual and estimated competence are plotted against the size of three different case bases. These results provide a high support in favor of our CEC-Model. In fact, it seems to be an almost perfect closely relationship between every two curves (problem-solving accuracy and CB's competence), and hence a strong correlation between them. For the sake of precision, we further measured this correlation for every CB using Equation 23 and we found high results reflecting a good match between the predicted and the true competence (0.91 for Mammographic Mass, 0.8 for Ionosphere, and 0.83 for Iris). Let us remind that the closer the value to one, the higher the correlation is.

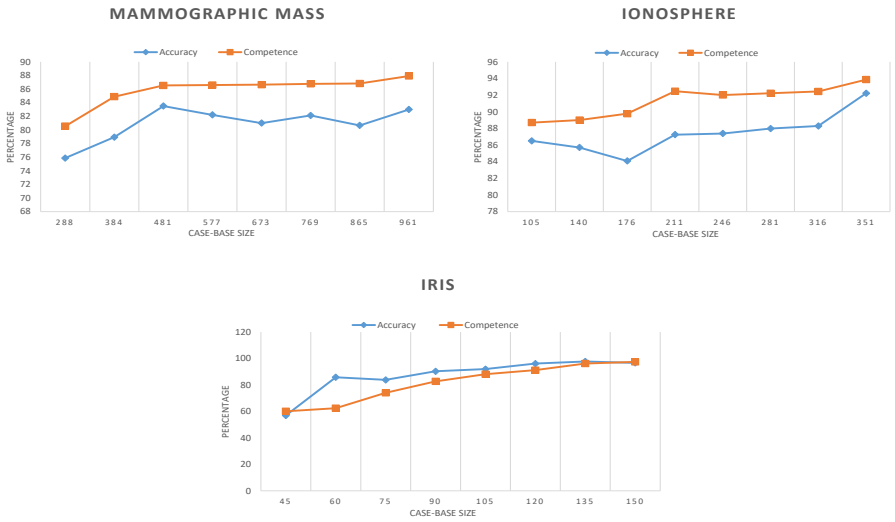

Fig. 6: Comparing estimated competence using our CEC-Model to the CB's accuracy for Mammographic Mass, Ionosphere, and Iris data sets

In the second results part, we note from Table 2 that our CEC-Model offers close competence estimation to the actual accuracy regarding the totality

of the different three tested case bases. Actually, this closeness was measured formally using the error rate criterion (Equation 24), where competitive results were provided comparing to those offered by S&M [1] and CMDC [14] models. For instance, we offered the minimum error rate for Iris data set which is estimated to 0.8%. Besides, the error rate for Mammographic Mass is estimated to environ 5.9%, where it is measured as 21.1% with S&M and environ 13.8% with CMDC (S&M and CMDC models are reviewed in Subsection 2.3).

Table 2: Results in term of Error rate (%)

Case base	S&M	CMDC	CEC-Model
Mammographic Mass	21.10	13.820	5.928
Ionosphere	3.544	0.287	1.779
Iris	4.010	0.927	0.807

6 Conclusion

The evaluation of knowledge resources are regularly a concern of widespread interest in knowledge management systems. In CBR systems, modeling case base competence with managing uncertainty within knowledge is essential to find the real coverage of cases. In this paper, we proposed a new competence model based on a previous work [1] with joining the ability to manage all levels of cases membership uncertainty towards groups building the case base, as well as to satisfy the need of imperfection handling when measure similarities and cases density. To support our model, we tested on data sets from UCI repository [24] with varying their size. Actually, competence estimations offered by our model are quite closely to the actual competence measurement (Accuracy) with a relatively high positive correlation between them.
Since the competence of CBR systems case bases presents the basis of the Case Base Maintenance (CBM) policies, we can, as future work, use our new competence model CEC-Model at the aim of maintaining CBs in order to well detect useless cases for target problems resolution.

References

1. Smyth, B., McKenna, E.: Modelling the competence of case-bases. *European Workshop on Advances in Case-Based Reasoning*. Springer, 208–220 (1998)
2. Dempster, A. P.: Upper and lower probabilities induced by a multivalued mapping. *The annals of mathematical statistics*, 325–339 (1967)
3. Shafer, G.: A mathematical theory of evidence. *Vol. 1. Princeton: Princeton university press* (1976)
4. Smets, P.: The transferable belief model for quantified belief representation. *In Quantified Representation of Uncertainty and Imprecision*, 267–301 (1998)

5. Smets, P.: The combination of evidence in the transferable belief model. *IEEE Transactions on pattern analysis and machine intelligence 12(5)*, 447–458 (1990)
6. Masson, M. H., Denœux, T.: ECM: An evidential version of the fuzzy c-means algorithm. *Pattern Recognition 41*, 1384–1397 (2008)
7. Smyth, B., Keane, M. T.: Remembering to forget: A Competence-Perserving Deletion Policy for CBR systems. *In Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI)*, 377–382 (1995)
8. Smyth, B., McKenna, E.: Competence models and the maintenance problem. *Computational Intelligence, 17(2)*, 235–249 (2001)
9. Lieber, J.: A criterion of comparison between two case bases. *European Workshop on Advances in Case-Based Reasoning*. Springer, Berlin, Heidelberg, 87–100 (1994)
10. Riesbeck, C. K., Schank, R. C.: *Inside case-based reasoning*. Psychology Press (2013)
11. Arshadi, N., Jurisica, I.: Feature Selection for Improving Case-Based Classifiers on High-Dimensional Data Sets. *In FLAIRS Conference*, 99–104 (2005)
12. Ayeldeen, H., Hegazy, O., Hassanien, A. E.: Case selection strategy based on K-means clustering. *In Information Systems Design and Intelligent Applications*. Springer, New Delhi. 385–394 (2015)
13. Smyth, B.: Case-base maintenance. *In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, Springer, Berlin, Heidelberg, 507–516 (1998)
14. Smiti, A., Elouedi, Z.: Modeling competence for case based reasoning systems using clustering. *The 26th International FLAIRS Conference, The Florida Artificial Intelligence Research Society, USA*, 399–404 (2013)
15. Smiti, A., Elouedi, Z.: Dbscan-gm: An improved clustering method based on gaussian means and dbscan techniques. *In 16th International Conference on Intelligent Engineering Systems (INES)*, IEEE, 573–578 (2012)
16. Smiti, A., Elouedi, Z.: SCBM: Soft Case Base Maintenance method based on competence model. *Journal of Computational Science*, (2017)
17. Smyth, B., McKenna, E.: Building compact competent case-bases. *In Proceedings of the international conference on case-based reasoning*, 329–342 (1999)
18. Ben Ayed, S., Elouedi, Z., Lefevre, E.: ECTD: Evidential clustering and case Types Detection for case base maintenance. *In the 14th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA)*, 1462–1469 (2017)
19. Ben Ayed, S., Elouedi, Z., Lefevre, E.: DETD: Dynamic policy for case base maintenance based on EK-NNclus algorithm and case Types Detection. To appear *In the 17th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU)*, Springer (2018)
20. Feller, W.: *An introduction to probability theory and its applications.* Vol. 2. John Wiley & Sons (2008)
21. Zadeh, L. A.: Fuzzy sets. *In Fuzzy Sets, Fuzzy Logic, And Fuzzy Systems: Selected Papers by Lotfi A Zadeh*, 394–432 (1996)
22. Dubois, D., Henri, P.: Possibility theory. *Computational complexity. Springer New York*, 2240–2252 (2012)
23. Jousselme, A. L., Grenier, D., Bossé, É.: A new distance between two bodies of evidence. *Information fusion, 2(2)*, 91–101 (2001)
24. Blake, C.: UCI repository of machine learning databases. *http://www. ics. uci. edu/mlearn/MLRepository. html* (1998)
25. Pearson, K.: Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *In Philosophical Transactions of the Royal Society of London*, 253–318 (1896)