La maintenance des bases de cas dans un cadre évidentiel

Maintenance of case bases in an evidential framework

S. Ben Ayed[1] Z. Elouedi[1] E. 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

Résumé :

Les systèmes de raisonnement à base de cas se caractérisent par un apprentissage incrémental à partir des expériences passées. Toutefois, cette évolution peut devenir incontrôlable et menacer leur succès qui dépend essentiellement du temps de recherche et de la qualité des bases. Pour garantir cette qualité, la maintenance des bases de cas (CBM) est nécessaire. Cependant, beaucoup de travaux réalisés dans ce domaine réduisent leur compétence en terme de résolution de problèmes surtout quand les cas reposent sur des informations imparfaites. Afin de régler ce problème, nous proposons, dans cet article, une nouvelle approche de Maintenance des Bases de Cas dans un cadre Évidentiel (ECBM). Cette méthode est capable de gérer l'imperfection par l'utilisation de la théorie des fonctions de croyance. L'idée clé est d'utiliser une technique d'apprentissage automatique gérant l'incertitude, puis de distinguer entre les différents types de cas pour ensuite effectuer la maintenance.

Mots-clés :

Raisonnement à base de cas, maintenance des bases de cas, théorie des fonctions de croyance, clustering.

Abstract:

Case Based Reasoning (CBR) systems are characterized by an incremental learning from past experiences. However, this evolution can be uncontrolled and threatens their success which depends essentially on the retrieval process time and case bases quality. In order to guarantee this quality, a case base maintenance (CBM) is necessary. In fact, several works that have been established within this field reduce the case bases competence toward problem resolution, particularly, when cases involve imperfect information. Our aim, in this paper, is to deal with this problem by proposing a new Evidential CBM approach (ECBM) using the belief function theory. Its key idea is to use a machine learning technique managing uncertainty. Then, it differentiates between the different case types in order to perform the maintenance.

Keywords:

Case based reasoning, Case base maintenance, Belief function theory, Clustering

1 Introduction

Le Raisonnement à Base de Cas (CBR) est un paradigme de résolution de problèmes par analogie qui repose sur la réutilisation des expériences passées pour la résolution de nouveaux problèmes. Dans les systèmes CBR, un problème cible suit tout un cycle pour qu'il soit résolu [1]. Ensuite, chaque problème résolu sera enregistré dans la base ce qui exige une grande capacité de stockage et ralentit le processus de recherche conduisant à la dégradation de la performance de ce type de systèmes. Par conséquent, nous observons une forte émergence du domaine de la Maintenance des Bases de Cas (CBM). Ainsi, CBM a été défini dans [2] comme étant le processus visant à faciliter le raisonnement et à améliorer la performance des systèmes CBR par l'implémentation des politiques révisant l'organisation et le contenu des bases. Les politiques CBM existantes souffrent généralement de quelques limitations parmi lesquelles leur incapacité à gérer l'imperfection des informations alors que les cas représentant des situations réelles sont souvent entachés d'incertitude et d'imprécision. Pour faire face à ce problème, plusieurs théories peuvent être utilisées. La théorie des fonctions de croyance [6][7] est parmi les plus appropriées car elle permet de gérer différents niveaux d'incertitude. Pour ces raisons, nous proposons, dans cet article, une approche dans un cadre évidentiel permettant la maintenance des bases tout en gérant l'imperfection des cas par l'utilisation de la théorie des fonctions de croyance et plus particulièrement en utilisant la technique des C-moyennes évidentielles (ECM) [10].

Le reste de cet article est organisé comme suit.

Dans la section 2, nous exposons quelques travaux associés au domaine du CBM. Les outils de base de la théorie des fonctions de croyance ainsi que la méthode de clustering [10] sont présentés dans la section 3. La section 4 détaille les différentes étapes de notre nouvelle approche Évidentielle du CBM (ECBM). Les expérimentations et leurs résultats sont présentés dans la section 5. Enfin, la section 6 sera consacrée à la conclusion.

2 L'état de l'art des politiques CBM

La plupart des méthodes CBM visent à réduire la taille des bases de cas pour les systèmes CBR tout en conservant leur compétence en terme de résolution de problèmes.

Condensed Nearest Neighbor (CNN) [3] est parmi les méthodes les plus connues qui consiste à sélectionner itérativement les prototypes des bases. L'idée de CNN est de choisir aléatoirement un cas à partir de la base originale et de tester si la nouvelle base pourra le résoudre ou non. Si cette base n'arrive pas à résoudre le problème, alors ce cas sera sélectionné pour être ajouté dans la nouvelle base et retiré de l'originale.

Nous pouvons également citer la méthode de *Reduced Nearest Neighbor* (RNN) [4] qui commence par l'utilisation de la base entière comme étant sa nouvelle base réduite. Puis, elle enlève les cas jusqu'à que tous les cas soient bien classés par la nouvelle base.

Par ailleurs, les auteurs dans [5] proposent une série de méthodes d'apprentissage à base d'instances appelées IBL1, IBL2 et IBL3. L'idée de IBL2, par exemple, est de commencer par un ensemble d'apprentissage vide. Pour chaque cas, s'il est mal classé par cet ensemble d'apprentissage, il y sera ajouté.

Les différentes méthodes que nous venons de citer sont incapables de gérer l'incertitude des informations, et cela risque de supprimer des cas importants vis-à-vis de la compétence des bases. Pour cela des approches de CBM reposant sur des théories de l'incertain ont vu le jour. On retrouve, par exemple, l'approche soft appelée SCBM [12] qui divise la base en utilisant la technique de clustering Soft DBSCAN-GM [13] puis se focalise sur la détection des cas qui devront être supprimés pour la maintenance.

Toutefois, les formalismes utilisés par les méthodes qui gèrent l'imperfection ne couvrent pas tous les aspects d'incertitude. Donc, nous proposons, dans cet article, d'étendre cet axe par l'utilisation de la théorie des fonctions de croyance qui est capable de prendre en compte différents niveaux d'incertitude, de l'ignorance totale jusqu'à la certitude complète.

3 La théorie des fonctions de croyance

3.1 Les concepts de base

La théorie des fonctions de croyance, aussi connue sous le nom de la théorie de l'évidence [6][7], permet de modéliser et de gérer les donnés incertaines.
Soit le cadre de discernement Ω contenant un ensemble fini de variables ω_k avec $k = \{1, \cdots, K\}$ qui réfèrent à K événements élémentaires d'un problème donné. A partir de ce cadre, nous définissons les 2^K sous ensembles possibles. Le point clé de cette théorie est la fonction de masse de croyance (bba) m qui représente la part de croyance attribuée sur les différents sous-ensembles de Ω et qui est définie comme suit :

$$\begin{aligned} m &: 2^{\Omega} \rightarrow [0, 1] \\ &\sum_{A \subseteq \Omega} m(A) = 1 \end{aligned} \tag{1}$$

Un élément A de Ω est nommé un élément focal si $m(A) > 0$. Quand les éléments focaux sont des singletons, alors m est appelée une bba bayésienne. Si la fonction de masse ne contient qu'un seul élément focal qui est Ω, nous sommes dans l'ignorance totale et m est appelée une fonction de masse vide. Dans

le cas, où la fonction de masse m ne possède qu'un seul élément focal qui est un singleton, alors m sera présentée comme une fonction de masse certaine.
Une bba est normalisée si la masse associée à l'ensemble vide est contrainte d'être nulle ($m(\emptyset) = 0$). Dans ce cas, elle correspond à l'hypothèse du monde clos [9]. Au contraire, si l'on considère que Ω peut être incomplet, alors la masse de croyance liée à $\emptyset$ peut être positive ou nulle ($m(\emptyset) \geq 0$). Ce cas correspond à l'hypothèse du monde ouvert [9].
L'une des solutions proposées pour la prise de décision repose sur la probabilité pignistique, notée $BetP$ [8]. Si la fonction de masse est normalisée, alors $BetP$ sera définie comme suit :

$$BetP(w) = \sum_{w \in A} \frac{m(A)}{|A|} \qquad \forall w \in \Omega \quad (2)$$

Si la fonction de masse est non normalisée, alors la transformation pignistique doit être précédée par une étape de normalisation.

3.2 Algorithme évidentiel des C-moyennes (ECM)

L'algorithme des C-Moyennes Évidentielles est une technique évidentielle de clustering proposée dans [10]. Elle est fondée essentiellement sur l'algorithme des C-moyennes floues [11]. L'objectif d'ECM consiste à affecter chaque objet avec des degrés de croyance aux différents sous-ensembles de clusters. Dans l'algorithme ECM, chaque cluster w_k est présenté par son centre $\boldsymbol{v_k}$ qui est un vecteur défini dans l'espace d'attributs des objets. Cependant, un objet peut appartenir aussi à une partition de clusters ($A_j \subseteq \Omega$) ayant une cardinalité supérieure à un ($|A_j| > 1$). A_j est alors appelé un méta-cluster et est représenté par un prototype noté $\overline{\boldsymbol{v_j}}$ tel que :

$$\overline{\boldsymbol{v_j}} = \frac{1}{|A_j|} \sum_{k=1}^{K} s_{kj} \boldsymbol{v_k} \quad (3)$$

où K représente le nombre total des clusters, $s_{kj} = 1$ si $w_k \in A_j$ et $s_{kj} = 0$ sinon.

Comme la plupart des techniques de clustering, le but est de minimiser la distance entre les objets appartenant au même cluster et de maximiser celles des objets des clusters différents. Dans le cadre évidentiel, ECM applique ce principe par la minimisation de la fonction objective suivante pour n objets et K clusters :

$$J_{ECM}(M,V) = \sum_{i=1}^{n} \sum_{j/A_j \neq \emptyset} |A_j|^{\alpha} m_{ij}^{\beta} d_{ij}^2 + \sum_{i=1}^{n} \delta^2 m_{i\emptyset}^{\beta} \quad (4)$$

sous la contrainte

$$\sum_{j/A_j \subseteq \Omega, A_j \neq \emptyset} m_{ij} + m_{i\emptyset} = 1 \quad \forall i = 1...n \quad (5)$$

où M représente la partition crédale définie dans l'espace $\mathbb{R}^{n \times 2^K}$, V est la matrice de 2^K centres de clusters ayant p attributs, m_{ij} désigne $m_i(A_j)$ et d_{ij} indique la distance euclidienne entre l'objet i et la partition j. Le paramètre α sert à contrôler le degré de pénalisation des partitions ayant une haute cardinalité. Enfin, β et δ présentent deux paramètres pour le traitement des objets bruités.
Afin de réaliser cette minimisation, une alternance entre deux phases est appliquée. La première consiste à supposer que V est fixe et résoudre l'Équation 4 contrainte par l'Équation 5 en utilisant le Lagrangien où les détails du calcul sont présentés dans [10]. Ainsi, nous obtenons comme résultat la partition crédale M relatif à l'appartenance des objets à tous les sous-ensembles de Ω.
Pour la deuxième phase, on considère que M est fixe et un problème de minimisation non contraint doit être résolu (Équation 4). Après la séquence de calculs comme présentée dans [10], la matrice résultante des centres des clusters V est obtenue à partir de la résolution du système $HV = B$, où B est une matrice de taille $(K \times p)$ et H est une matrice carrée de taille $(K \times K)$ comme elles sont définies dans [10].

4 Une nouvelle approche CBM dans un cadre Évidentiel (ECBM)

Notre objectif, pour cet article, est de maintenir les bases de cas pour les systèmes CBR, plus précisément, de réduire leurs tailles tout en gérant l'incertitude et en améliorant leur compétence et performance. Pour ce faire, nous proposons une approche de Maintenance des Bases de Cas dans un cadre Évidentiel qui est composée de trois étapes principales.

Lors de la première étape, l'objectif est de partitionner la base en utilisant une technique de clustering évidentielle afin de pouvoir affecter les cas avec un degré de croyance aux partitions de clusters puisque dans le cadre évidentiel les clusters se chevauchent. Par conséquent, l'incertitude envers l'appartenance des cas aux différents clusters est bien gérée. Dans la seconde étape, nous classons les cas selon quatre types (cas bruité, similaire, isolé et interne) que nous définirons. Enfin, dans la dernière étape, la maintenance est réalisée par la suppression des cas associés aux types peu souhaitables.

4.1 Étape 1 : Le clustering évidentiel

Pour cette étape, l'objectif est d'effectuer une technique de clustering évidentielle qui consiste à utiliser la théorie des fonctions de croyance pour gérer l'incertitude quant à l'appartenance des cas aux clusters. En effet, le clustering évidentiel des cas offre une partition crédale qui permet à un cas d'appartenir à plusieurs partitions de clusters. Dans notre contexte, la technique de clustering évidentielle utilisée afin de générer cette partition crédale est les C-Moyennes Évidentielles [10]. De cette façon, ECM constitue la première étape de notre approche puisqu'elle répond suffisemment à nos exigences. En effet, les sorties générées par ECM, après la convergence de l'Équation 4, seront exploitées durant les étapes suivantes afin de différencier les types de cas selon leurs caractéristiques. Ces résultats sont la partition crédale M comme présentée dans les Équations 10 et 11 et la matrice des centres des différentes partitions des clusters V.

4.2 Étape 2 : La différenciation des différents types de cas

Dans une base, nous proposons de diviser les cas en quatre types selon leurs caractéristiques :

- Les cas bruités : Ils ont des valeurs qui, logiquement, ne peuvent pas faire partie de la base.
- Les cas isolés : Ils ont des valeurs un peu différentes par rapport à la plupart des cas dans la base.
- Les cas similaires : Ils représentent la plupart des expériences dans la base.
- Les cas internes : Ils représentent les prototypes des différents clusters ou concepts de la base.

1. La détection des cas bruités :

Pour rappel, la première étape d'ECM génère pour chaque cas un degré de croyance concernant l'appartenance aux différentes partitions de clusters. L'idée alors est de manipuler les cas bruités de la même manière que celle utilisée dans [15] où il consiste à allouer un cluster afin d'y attribuer les bruits. Dans notre cas, l'ensemble vide qui correspond à l'hypothèse du monde ouvert est le cluster qui s'occupe des cas représentant une distorsion des valeurs et qui ne peuvent pas être affectés à un cluster du cadre de discernement. Par ailleurs, les bruits sont généralement caractérisés par leur éloignement et leur isolement des regroupements des objets. Donc, les cas bruités sont les cas alloués à l'ensemble vide avec un niveau de croyance "élevé". Pour notre approche, un degré d'affectation est dit "élevé" si et seulement s'il est supérieur à la somme de tous les autres degrés. Ainsi, nous définissons les cas bruités comme suit :

$$\boldsymbol{x_i} \in Br \;\; ssi \;\; m_i(\emptyset) > \sum_{A_j \subseteq \Omega, A_j \neq \emptyset} m_i(A_j) \quad (6)$$

où $\boldsymbol{x_i}$ est une instance de cas et Br représente l'ensemble contenant tous les cas bruités.

Généralement, les cas étiquetés comme bruit ne sont pas intéressants dans les bases où ils ne peuvent aider à résoudre ni les autres cas ni eux-mêmes. Par conséquent, ces cas mènent à réduire la compétence des bases.

2. La distinction entre les cas similaires et les cas isolés :

Une fois les cas bruités détectés, il faut maintenant distinguer entre les cas similaires et les cas isolés. Cette distinction sera faite par la manipulation des distances.

En effet, l'algorithme de clustering ECM fait que la plupart des cas seront autour des centres des clusters. Alors, l'idée consiste à mesurer les distances entre les cas et les centres des différents clusters. Par conséquent, les cas éloignés des centres sont étiquetés comme cas isolés et ceux proches d'un centre d'un cluster sont étiquetés comme similaires par rapport à ce cluster-là. Cependant, la question qui se pose à présent est : "Comment manipuler ces distances, dans un cadre de travail évidentiel, en profitant de la partition crédale des cas afin de bien gérer l'incertitude ?"

L'idée de notre approche est d'exploiter les bbas qui sont déjà fournis par ECM et de calculer des distances afin de distinguer entre les deux types de cas (similaire et isolé) tout en gérant l'incertitude. Pour ce faire, nous avons adapté la distance de Mahalanobis [16] à la théorie des fonctions de croyance en utilisant la matrice de covariance évidentielle [14]. Ainsi, nous nommons cette nouvelle distance : Distance Évidentielle de Mahalanobis (DEM). Clairement, DEM a plusieurs avantages tels que :

- Elle est appropriée pour les distributions non uniformes et capable de supporter les formes arbitraires des clusters et pas uniquement sphériques comme c'est le cas avec la distance euclidienne.
- Elle prend en compte la covariance entre les variables pendant le calcul des distances.
- Elle gère bien l'incertitude vis-à-vis de l'appartenance des cas non seulement à un cluster mais aussi à toutes les partitions de clusters. La manipulation de celles non singletons sert à prendre en compte la totalité de l'incertitute. Cette fonctionnalité est obtenue grâce à la matrice de covariance évidentielle Σ qui calcule la matrice de covariance d'un cluster dans un espace p-dimentionnel tout en exploitant la partition crédal M afin de gérer l'incertitude de l'appartenance des cas aux différents clusters.

En effet, pour une base de cas contenant n instances multivariées $\boldsymbol{x_i}$ définies dans un espace p-dimentionnel, la DEM entre un cas et un cluster est définie telle que :

$$DEM(\boldsymbol{x_i}, \boldsymbol{v_k}) = \sqrt{(\boldsymbol{x_i} - \boldsymbol{v_k})^T \Sigma_k^{-1} (\boldsymbol{x_i} - \boldsymbol{v_k})} \tag{7}$$

où $\boldsymbol{v_k}$ est le centre du cluster k et Σ_k représente la matrice de covariance évidentielle [14] du cluster k ayant la forme suivante :

$$\Sigma_k = \sum_{i=1}^{n} \sum_{A_j \ni w_k} m_{ij}^2 |A_j|^{\alpha-1} (\boldsymbol{x_i} - \overline{\boldsymbol{v_j}})(\boldsymbol{x_i} - \overline{\boldsymbol{v_j}})^T \tag{8}$$

où A_j est une partition avec $j = 1, .., 2^K$, k est le numéro du cluster avec $k = 1, .., K$, m_{ij} et $\overline{\boldsymbol{v_j}}$ sont respectivement la partition crédale et leurs centres générés par ECM. L'exposant α sert à pénaliser l'allocation des croyances aux partitions ayant des cardinalités élevées.
Σ_k exploite ainsi les prototypes des partitions obtenues à l'issue de l'étape 1, ainsi que la partition crédale des cas envers toutes les partitions contenant le cluster k comme un élément focal afin de bien estimer la dispersion des cas autour des centres de différents clusters.

Après le calcul de la matrice de toutes les distances DEM de n cas par rapport aux k clusters, l'objectif est de fixer un seuil pour comparer ces distances et de décider, par la suite, s'il s'agit d'un cas similaire ou isolé. Pour cela, nous excluons tout d'abord les cas déjà étiquetés comme bruits. Ensuite, nous calculons un seuil pour chaque cluster qui est égal à la moyenne des distances par rapport à son centre. Nous

définissons ce seuil de la manière suivante :

$$Seuil_k = \frac{\sum_{\boldsymbol{x_i} \notin Br} DEM(\boldsymbol{x_i}, \boldsymbol{v_k})}{\#TotalCas - \#CasBruits} \quad (9)$$

L'intuition de cette proposition est d'indiquer combien en moyenne un cas est proche du centre de la distribution et de se comparer avec celui-ci. C'est pourquoi, nous excluons les cas bruités car la moyenne est très sensible aux valeurs bruitées qui peuvent affecter de façon non négligeable le résultat. Ainsi, nous pouvons maintenant distinguer entre les cas similaires et isolés en utilisant la forme suivante :

$$\boldsymbol{x_i} \in \begin{cases} Sm_k & if\ \exists k / DEM(\boldsymbol{x_i}, \boldsymbol{v_k}) < Seuil_k \\ Is & sinon \end{cases} \quad (10)$$

où Sm_k représente l'ensemble des cas similaires qui sont situés près du cœur du cluster k et Is représente l'ensemble contenant les cas isolés qui sont plus éloignés.

3. L'étiquetage des cas internes :

En arrivant à cette phase, nous avons déjà étiqueté chaque cas soit comme bruit, soit similaire, soit isolé. Toutefois, les cas similaires par rapport à un même cluster sont considérés comme redondants. Donc, nous devons retenir un seul cas pour chaque cluster afin de couvrir les autres après leur suppression de la base. Par conséquent, notre approche choisit de sélectionner le cas le plus proche au centre de chaque cluster et de l'étiqueter comme un cas interne. Logiquement, nous obtenons à la fin un nombre de cas internes égal au nombre initial des clusters K. Formellement, on peut définir l'ensemble de ces cas comme suit :

$$\boldsymbol{x_i} \in In\ ssi\ \exists k; \neg \exists \boldsymbol{x_j} / DEM(\boldsymbol{x_j}, \boldsymbol{v_k}) < DEM(\boldsymbol{x_i}, \boldsymbol{v_k}) \quad (11)$$

où In représente l'ensemble des cas internes, $\boldsymbol{x_i}$ et $\boldsymbol{x_j}$ sont deux instances de cas et $\boldsymbol{v_k}$ est le centre du cluster k.

4.3 Étape 3 : La maintenance

Les étapes précédentes visent à conduire à une maintenance des bases efficace. En général, la maintenance des systèmes CBR, plus précisément des bases de cas, peut apparaître sous différentes formes telles que la suppression ou la mise à jour d'un nombre de cas. Notre contribution s'intéresse à l'élimination des cas. Tout d'abord, il s'agit de supprimer les cas qui diminuent l'aptitude des systèmes CBR à résoudre les problèmes. Ces cas correspondent aux cas bruités. Par ailleurs, les cas étiquetés comme similaires sont à éliminer puisqu'ils sont considérés comme étant des cas redondants et peuvent être couverts par les cas déjà définis comme internes. La motivation derrière leur suppression est d'alléger la base et d'avoir un meilleur temps de réponse sans réduire sa compétence à résoudre les problèmes. D'un autre côté, nous retenons impérativement les cas isolés car aucun autre cas ne peut les couvrir et leur suppression risque d'aboutir à des problèmes définitivement insolubles. Nous conservons également les cas internes car ils vont couvrir tous les cas similaires.

5 L'expérimentation

Pour la phase d'expérimentation, nous avons développé notre contribution avec Matlab R2015a. Notre approche a été testée sur six bases à partir d'*U.C.I Repository* [17]. Ces bases sont : Breast-Cancer (BC), Climate Model (CM), Ionosphere (IO), Ecoli (EC), Mice-Proteine (MP) et Iris (IR). Nous avons fixé le nombre de clusters égale au nombre de classes originales et nous avons considéré les paramètres par défaut d'ECM. Nous n'avons pas pénalisé l'allocation des croyances aux partitions ayant une cardinalité élevée en fixant l'exposant de pénalisation α à 1.

En gardant à l'esprit que notre objectif principal est de maintenir les bases de cas tout en conservant ou en améliorant la performance et la compétence durant la résolution des problèmes, nous mesurons l'efficacité de notre approche CBM dans le cadre Évidentiel (ECBM) selon les trois critères d'évaluation qui sont présentés dans la section 5.1. Les résultats obtenus avec

Tableau 1 – Taille du Stockage [S(%)]

CB	Taille de stockage (%)				
	ICBR	ECBM	CNN	RNN	IBL2
BC	100	64.71	10.54	**8.49**	90.04
CM	100	**11.85**	32.22	28.7	32.4
IO	100	41.03	21.37	**16.52**	90.03
EC	100	58.33	36.9	30.95	**24.46**
MP	100	41.04	24.79	**19.87**	20.28
IR	100	38.67	16	10.67	**8.67**

Tableau 2 – Précision [PCC(%)]

CB	Précision (%)				
	ICBR	ECBM	CNN	RNN	IBL2
BC	96.19	**99.09**	70.84	62.07	97.72
CM	84.07	**89.06**	81.03	81.29	88
IO	86.89	**87.5**	62.67	48.28	85.13
EC	82.73	**90.81**	62.1	60.58	69.8
MP	88.02	**88.24**	86.51	87.37	78.12
IR	98	98.28	70.83	56.25	**100**

notre méthode sont comparés avec les résultats des bases initiales non maintenues (ICBR) ainsi qu'avec quelques autres approches CBM les plus connues : CNN [3], RNN [4] et IBL2 [5] dans la section 5.2.

5.1 Les critères d'évaluation

— **Taille du Stockage [S(%)] :** est le pourcentage de la taille de la base maintenue par rapport à la taille initiale. C'est le degré de réduction de la taille de la base. Plus la taille de Stockage (S%) est réduite, plus la maintenance est bien atteinte. Ce critère est défini comme suit :

$$S = \frac{Taille\ finale\ de\ la\ base}{Taille\ initiale\ de\ la\ base} \times 100 \quad (12)$$

— **Précision [PCC(%)] :** Ce critère fait référence à la mesure du pourcentage des cas bien classés (PCC) et il est présenté comme suit :

$$PCC = \frac{\#\ Cas\ Bien\ Classés}{\#\ Total\ Cas\ Classés} \times 100 \quad (13)$$

Pour calculer la valeur du PCC et effectuer les différentes comparaisons, nous avons choisi "le voisin le plus proche" (1-NN) comme algorithme de classification. Ce pourcentage de bonne classification a été obtenu par validation croisée (*10-Folds Cross Validation*).

— **Temps de la Recherche [T(s)] :** Puisque la performance des systèmes CBR est fortement liée au temps de résolution des problèmes, nous avons considéré ce critère comme important et nous l'avons appliqué autour de l'algorithme 1-NN pour mesurer le temps de classification en secondes.

5.2 Les résultats et la discussion

Les résultats obtenus selon les différentes méthodes retenues et pour les différentes bases de cas sélectionnées sont donnés dans les Tableaux 1, 2 et 3.

En terme de réduction de la taille (Tableau 1), notre approche est capable de réduire de plus de la moitié la taille initiale des bases (100%), contrairement à IBL2 qui ne modifie quasiment pas certaines bases telles que BC et IO. Toutefois, certaines approches vont au delà de la réduction que nous obtenons avec notre approche. Par exemple, notre approche a laissé environ 64% de la base BC alors que RNN la réduit à 8%. Mais il est nécessaire de prendre en compte d'autres critères pour évaluer la pertinence de la maintenance. Il est nécessaire de s'assurer de la valeur du critère de précision (Tableau 2). En fait, nous atteignons environ 99% de précision pour la même base (BC), alors qu'avec RNN nous remarquons une mauvaise précision évaluée à 62%. De plus, nous observons qu'en terme de précision nous obtenons de meilleurs résultats pour quasiment toutes ces bases même dans leur version d'origine. Celà est bien évidemment expliqué d'une part par l'élimination des cas bruités qui dégradent la compétence des bases en résolution des problèmes, et d'autre part par l'efficacité de la stratégie suivie pour la détection des cas redondants. Nous concluons

Tableau 3 – Temps de la Recherche [T(s)]

CB	Temps du recherche (s)				
	ICBR	ECBM	CNN	RNN	IBL2
BC	0.01	0.007	0.006	**0.005**	**0.005**
CM	0.034	**0.004**	0.013	0.009	0.005
IO	0.084	0.016	0.010	**0.005**	1.386
EC	0.010	0.006	0.005	0.007	**0.004**
MP	0.038	0.008	0.006	0.006	**0.005**
IR	0.008	**0.005**	0.009	0.010	0.010

à partir des bases qui n'étaient pas bien réduites qu'elles ne contiennent pas beaucoup de bruits ou de cas redondants. Enfin, nous observons à partir du Tableau 3 une réduction en terme de temps de recherche par rapport à la base non maintenue, ce qui est cohérent car ce critère dépend essentiellement du nombre d'instances dans la base. En fait, nous avons amélioré la performance en terme de temps pour toutes les bases. Concernant la comparaison avec les autres méthodes, nous remarquons que les valeurs sont très proches pour ce critère. Malgré cela, nous avons enregistré les meilleurs temps de recherche pour quelques bases tels que 0.004s pour CM et 0.005 pour IR. En se comparant avec IBL2, nous observons une différence remarquable au niveau de la base IO où ECBM nécessite 0.016s comme un temps de recherche alors qu'avec IBL2, il a fallu environ 1.386s.

6 Conclusion

Dans cet article, nous avons présenté une nouvelle approche de maintenance des bases de cas des systèmes CBR dans un cadre évidentiel qui est capable de gérer l'incertitude dans les descriptions des cas. Il s'agit de maintenir ces bases par la suppression des cas bruités et redondants. L'idée est de regrouper les cas dans des clusters où chaque cas appartient à toutes les partitions des clusters avec un degré de croyance en utilisant ECM [10], puis exploiter les bba générées autant que possible afin de distinguer entre les différents types de cas. Enfin, la maintenance des bases est réalisée par la suppression des bruits afin d'améliorer la compétence des bases, ainsi que les cas étiquetés comme similaire afin de les alléger. Comme travail futur, nous pouvons étendre notre approche en ajoutant une heuristique visant à determiner le nombre initial de clusters adéquat pour le clustering.

Références

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

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

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

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

[5] D. Aha, W. Kibler, M. Albert. Instance-based learning algorithms. *Machine learning*, 1991, pp. 37-66.

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

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

[8] P. Smets. The transferable belief model for quantified belief representation. *In Quantified Representation of Uncertainty and Imprecision*, 1998, pp. 267-301.

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

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

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

[12] A. Smiti, Z. Elouedi. Maintaining Case Based Reasoning Systems Based on Soft Competence Model. *In International Conference on Hybrid Artificial Intelligence Systems*, 2014, pp. 666-677.

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

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

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

[16] P. C. Mahalanobis. Mahalanobis distance. *In Proceedings National Institute of Science of India*, 1936, pp. 234-256.

[17] C. Blake, C. J. Merz. {UCI} Repository of machine learning databases, 1998.