Maintaining case knowledge vocabulary using a new Evidential Attribute Clustering method

S. BEN AYED* and Z. ELOUEDI

LARODEC, Institut Supérieur de Gestion de Tunis, Université de Tunis, 41 Avenue de la liberté, cité Bouchoucha, 2000 Le Bardo, Tunisia
**E-mail: safa.ben.ayed@hotmail.fr*
E-mail: zied.elouedi@gmx.fr

E. LEFEVRE

Univ. Artois, EA 3926, LGI2A, 62400 Béthune, France
E-mail: eric.lefevre@univ-artois.fr

Maintaining the vocabulary of case knowledge within Case Based Reasoning (CBR) presents a crucial task to ensure a high-quality problem-solving and to improve retrieval performance for large-scale CBR systems. To do, we propose, in this paper, a method that manages uncertainty while selecting the best attributes characterizing case knowledge by using belief function theory. Actually, this method is based on a new evidential attribute clustering technique to eliminate redundant and noisy attributes describing cases.

Keywords: Case based reasoning; Maintenance; Case vocabulary; Attribute clustering; Belief function theory; Feature selection.

1. Introduction

Case Based Reasoning is a methodology that aims to solve new problems through reusing the most similar past experiences.[1] Over the years, CBR has known a widespread interest in several domains thanks to its capability to learn incrementally. Actually, the arriving of a new problem triggers a cycle with four steps.[1] First, CBR *retrieves* from the case base the most similar one. Second, it *reuses* the corresponding solution to be adapted to the target problem. Third, the proposed solution is *revised*. Finally, the new case is *retained* in order to extend case base's capability in future problems resolution. To control this case knowledge growth along with preserving its competence, many works are interested in Case Base Maintenance (CBM)

field.[2,3] However, knowledge containers[4] such as *Similarity measures*, *Adaptation rules* and *Vocabulary* are also considerable maintenance targets. Accordingly, researches around Case-Based Reasoner Maintenance (CBRM)[5] are also directed. In this work, we are situated in the maintenance of the vocabulary knowledge container for structual CBR systems. Herein, a case is described using a number of attributes[a] which are mainly serving in matchmaking and case retrieval. Logically, the more a case is described, the best solution is offered. However, some application domains[12] describe cases with a very large number of features which leads to decrease problem-solving performance. Besides, the existence of irrelevant and noisy features can seriously reduce CBR systems' competence. To deal with these problems, we propose, in this paper, an approach that selects only the most 'informative' features using a new evidential attribute clustering method which is based on belief function theory[6,7] to manage all levels of uncertainty towards the membership of features to the different clusters.
The rest of this paper is organized as follows. Two among the used applicable concepts within vocabulary maintenance researches are reviewed in Section 2. The necessary background related to the used evidential clustering technique is briefly introduced in Section 3. Throughout Section 4, we describe the different steps of our proposed method for case knowledge vocabulary maintenance. Finally, Section 5 conducts experimental study on UCI data sets to evaluate our newly method.

2. Applicable concepts for vocabulary maintenance

Many concepts within machine learning studies have been applied in the different methods for maintaining CBR systems. Among them, we review, in Subsections 2.1 and 2.2, *feature selection* and *attribute clustering* concepts. In Subsection 2.3, we explain our motivation behind this work.

2.1. *Applying Feature selection for vocabulary maintenance*

The vocabulary of CBR systems defines the information towards the corresponding field and the way to express them. To maintain vocabulary with considering structured CBR systems, we should select only features that ensure accurate retrieval outcomes. Herein, the problem of Feature Selection (FS) arises. Actually, since FS is an NP-Hard problem aiming specially in irrelevant features elimination, we find numerous FS techniques where some

[a]In this paper, we use *attribute* and *feature* terms exchangeably.

of them were combined with CBR systems.[8,9] Besides, some techniques are leading to select features by assigning weights reflecting their relevance.[10]

2.2. *Applying Attribute clustering for Feature selection*

Attribute clustering is carried out in several researches[11,12] as a feature selection task. Like the standard objects' clustering, attributes belonging to the same cluster are similar and those belonging to different ones are dissimilar. However, the notion of similarity within attributes reflects the relation between them (*e.g in term of correlation, dependency, etc.*). Consequently, that leads us to eliminate dispensable features by selecting only representative one(s) for each cluster.

2.3. *Motivation and discussion*

Actually, using attribute clustering for maintaining case knowledge vocabulary has the advantage of preserving the relation between features which offers a better flexibility at the level of CBR framework. In fact, we can replace any selected representative feature by another one belonging to the same cluster. However, existing researches in this road even cannot manage uncertainty about attributes membership to clusters or they are not able to manage all levels of uncertainty, from the complete ignorance to the total certainty. For that reason, we propose to maintain cases vocabulary using a powerful tool for uncertainty management called belief function theory.[6,7]

3. Belief function theory

The belief function theory (or Evidence theory)[6,7] is a theoretical framework for reasoning with partial and unreliable information. Its basic concepts will first be recalled in Subsection 3.1, and the used evidential clustering algorithm will then summarized during Subsection 3.2.

3.1. *Basic concepts*

Let Ω be a finite set of events called the frame of discernment, and ω is a variable taking values in Ω. The basic belief assignment (*bba*) function m, from 2^{Ω} to $[0, 1]$, represents the partial knowledge towards the real value taken by ω verifying $\sum_{A \subseteq \Omega} m(A) = 1$. Complete ignorance corresponds to $m(\Omega) = 1$, and total certainty is achieved when $m(A) = 1$ and A is a singleton. The subset A is called focal element if $m(A) > 0$. Furthermore, a bba m can be represented by $bel(A)$ as the amount of support given only

to the subset A. In addition, it can be represented by the plausibility $pl(A)$ which is the maximum amount of belief that can be assigned to A, and defined such that $pl(A) = \sum_{A \cap B \neq \emptyset} m(B)$ for all $A \subseteq \Omega$. Concerning the decision making process, choosing the highest pignistic probability $BetP$ presents one of the most powerful techniques, which is defined as follows:

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

3.2. *Evidential dissimilarity data clustering*

The aim of evidential dissimilarity data clustering is to construct a credal partition for dissimilarity data. Actually, the credal partition quantifies the uncertainty of n objects membership to clusters using bba functions where $\Omega = \{\omega_1, .., \omega_c\}$ denotes a set of c clusters. Among such techniques offering a credal partition, we enumerate RECM[13] , EVCLUS[14] and k-EVCLUS[15] . The two latter do not make assumption about the dissimilarity nature, although RECM assumes explicitly that the input dissimilarity is calculated as Squared Euclidean Distances.[13] For that reason, we centralize our work around k-EVCLUS which is an improvement of EVCLUS algorithm.
Let $D = (d_{ij})$ is $n \times n$ dissimilarity matrix where d_{ij} is the degree of dissimilarity between objects x_i and x_j. Besides, let $F_1, ..., F_f$ are f focal sets. Logically, the more two objects are similar, the more plausible that they belong to the same cluster. In fact, it is shown that $pl_{ij} = 1 - \kappa_{ij}$[14] with $\kappa_{ij} = \sum_{A \cap B = \emptyset} m_i(A) m_j(B)$ is the degree of conflict between m_i and m_j. Since similar objects should have mass functions with low degrees of conflict and conversely, the credal partition within k-EVCLUS, presented as a matrix M of size $n \times f$, is the result of the following stress function minimization which is solved using Iterative Row-rise Quadratic Programming (IRQP):

$$J(M) = \eta \sum_{i<j} (\kappa_{ij} - \delta_{ij})^2 \tag{2}$$

where η is a normalizing constant and $\delta_{ij} = \varphi(d_{ij})$ are transformed dissimilarities. Using matrix notations, κ_{ij} is written herein as $m_i^T C m_j$, where C is a square $f \times f$ matrix with general term $C_{kl} = 1$ if $F_k \cap F_l = \emptyset$, and $C_{kl} = 0$ otherwise.
Moreover, k-EVCLUS eliminates redundancy of information within dissimilarity matrix (e.g. given x_1 and x_2 are two very similar objects. For any object x_3 dissimilar from x_1, it is then usually dissimilar from x_2) in order

to reduce the complexity of stress criterion calculation such that:

$$J_k(M) = \eta \sum_{i=1}^{n} \sum_{r=1}^{k} (\kappa_{ij_r(i)} - \delta_{ij_r(i)})^2 \tag{3}$$

with $j_1(i), ..., j_k(i)$ are k integers sampled randomly for $i = 1, .., n$.
Actually, $J_k(M)$ requires $O(nk)$ operations instead of $O(n^2)$ for EVCLUS.

4. Maintaining case knowledge vocabulary in an evidential framework

The main purpose of our proposed method is to maintain case knowledge vocabulary by eliminating on the one hand redundant features which are so correlated, and on the other hand noisy features which lead to distort the problem-solving. Our method is thus summed up by the following steps.

4.1. *Step 1: Creating cases' features relational matrix*

Features' relationship that we take into account in our method reflects the amount of correlation between them. Given a case base CB with n objects and p features, we choose to use the *pearson's correlation coefficient*[16] , denoted by r, to measure the linear association between every two variables. Hence, $R = (r_{_{AB}})$ is our relational matrix and $r_{_{AB}}$ is defined such that:

$$r_{_{AB}} = \frac{\sum_{i=1}^{n}(a_i - \overline{a})(b_i - \overline{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \overline{a})^2}\ \sqrt{\sum_{i=1}^{n}(b_i - \overline{b})^2}} \tag{4}$$

where a_i and b_i are the values of every two attributes A and B respectively for object i, and $\overline{a}$ and $\overline{b}$ are their mean values.

4.2. *Step 2: Generating cases' features dissimilarity matrix*

Definition 4.1. Two features A and B are said to be similar if there is a high correlation between them, and conversely.

According to Definition 4.1, we can thus generate a matrix $D = (d_{AB})$ as a $p \times p$ dissimilarity matrix between features where $d_{AB} = f(r_{_{AB}})$ with f is a function from $[-1, 1]$ to $[0, 1]$. Actually, we have $-1 < r_{_{AB}} < 1$ where three *Situations (S_i)* are therefore arising:[16]

- S_1: If $r_{_{AB}} \simeq -1 \Rightarrow$ High correlation (negative) $\Rightarrow$ High similarity.
- S_2: If $r_{_{AB}} \simeq 1 \Rightarrow$ High correlation (positive) $\Rightarrow$ High similarity.
- S_3: If $r_{_{AB}} \simeq 0 \Rightarrow$ No correlation $\Rightarrow$ High dissimilarity.

Within S_1 and S_2, A and B are offering the same information. Consequently, they are redundant, whereas it is not the case for S_3.

Now, it is straightforward to show that the dissimilarity between two features A and B is computed as follows:

$$d_{AB} = f(r_{AB}) = 1 - |r_{AB}| \tag{5}$$

where r_{AB} represents the similarity between A and B.

4.3. *Step 3: Evidential attribute clustering*

After generating a square dissimilarity matrix for p features, we aim now to group them using a dissimilarity data clustering technique which is able to manage all levels of uncertainty within the input dissimilarity data. For that reason, we use an evidential technique called k-EVCLUS[15] as presented throughout Section 3 where we apply it on the already created dissimilarity matrix for p features during *Step 2*. The output of this attribute clustering procedure is the set of features detected as outliers as well as the credal partition of features' membership to the different clusters.

4.4. *Step 4: Case knowledge vocabulary maintenance*

Ultimately, we aim to define our strategy for case vocabulary maintenance. Actually, we eliminate all noisy features detected during the previous step since they distort the process of problem-solving. On the other hand, we also remove redundant features belonging to the same cluster and keeping only one as their representative. In fact, the membership of features to the different clusters is decided through the pignistic probability transformation from the credal partition as defined in Equation 1. Removing redundant features serves mainly in reducing the execution time of indexing and retrieving cases, which then conduct to improve CBR systems performance.

5. Experimental study: Results and analysis

To measure our method's efficiency, we developed it using R software, testing on UCI repository data sets, evaluating results via accuracy, which is calculated using 10-folds cross validation technique, and retrival time criteria (Table 1). This is done after varying the number of clusters K from 3 to 7 and choosing then the most convenient one[b]. Finally, we compare

[b]The number of clusters (K) offering the highest accuracy for our method: Ionosphere ($K = 3$), Glass ($K = 5$), WDBC ($K = 4$), German ($K = 4$), Heart ($K = 4$), and Yeast ($K = 4$).

results related to our method (AttEvClus-CBR) with those offered by the original non maintained case base (Original-CBR), as well as the updated case bases at the vocabulary level using ReliefF[10] (ReliefF-CBR) as one of the most known FS methods. Like we did with our method, we choose for RefiefF-CBR the most relevant attributes set offering the highest accuracy[c].

Table 1. Evaluation of our proposed vocabulary maintaining method

	Case bases	Original-CBR		ReliefF-CBR		AttEvClus-CBR	
		PCC[a]	Time[b]	PCC	Time	PCC	Time
1	Ionosphere	85.48	1.942	84.88	1.188	**88.33**	0.912
2	Glass	97.64	0.967	98.11	0.882	**98.59**	0.762
3	WDBC	60.16	1.710	96.33	1.112	**96.46**	1.013
4	German	64.6	1.812	**73.4**	1.213	73.25	1.211
5	Heart	57.5	2.103	62.45	1.091	**62.98**	1.028
6	Yeast	55.32	0.954	**99.05**	0.722	**99.05**	0.724

Note: [a] Percentage of correct classifications (%) offered by 5-NN algorithm (to be not sensible to noisy cases). [b] The retrieval time in seconds exerted in 5-NN.

Obviously, results offered by our proposed method, as shown in Table 1, ensure a high-quality case knowledge vocabulary maintenance. In term of accuracy (PCC), we note that our method has been able to increase the problem solving competence for all case bases (CB) comparing to the original ones. For instance, it increases the accuracy for "WDBC" data set from 60.16% to 96.46%. Comparing to ReliefF-CBR, our method has also competitive results by offering the best accuracies for almost all the CBs. These results can be explained by the high quality of the used evidential clustering technique in managing uncertainty and in noisy features detection. In term of retrieval time, our mainly objective is to provide lower values than those offered by the original CBR systems. Indeed, we note that there is a respectable time reduction for all the different CBs. For example, the time decreases from about 2.1s to about 1s for "Heart" data set. Besides, we note a slightly faster process for almost all the CBs comparing to ReliefF-CBR.

6. Conclusion

In this paper, we have developed a method to maintain the vocabulary of CBR systems by eliminating irrelevant and redundant features. To do, we

[c]The number of features (p) offering the highest accuracy for ReliefF: $p = 5$ for Ionosphere, Glass and WDBC. $p = 4$ for German and Heart. And $p = 3$ for Yeast.

applied a new evidential attribute clustering technique that considers the correlation between features and manages uncertainty about their membership to clusters. Finally, it keeps only representative features for clusters.

References

1. A. Aamodt, E. Plaza, Case-based reasoning: Foundational issues, methodological variations, and system approaches. In *AI communications* (IOS press, 1994), pp. 39-59.
2. A. Smiti, Z.Elouedi, SCBM: soft case base maintenance method based on competence model. In *Journal of Computational Science* (Elsevier, 2017), DOI: 10.1016/j.jocs.2017.09.013.
3. S. Ben Ayed, Z. Elouedi, E. Lefevre, ECTD: Evidential Clustering and case Types Detection for case base maintenance. In *International Conference on Computer Systems and Applications* (IEEE, 2017), pp. 1462-1469.
4. M. M. Richter, M. Michael, Knowledge containers. In *Readings in Case-Based Reasoning* (Morgan Kaufmann, 2003).
5. D. C. Wilson, D. B. Leake, Maintaining Case-Based Reasoners: Dimensions and Directions. In *Computational Intelligence* (2001), pp. 196-213.
6. G. Shafer: A mathematical theory of evidence. In *Princeton university press* (Princeton university press, 1976).
7. A. P. Dempster, Upper and lower probabilities induced by a multivalued mapping. In *The annals of mathematical statistics* (1967), pp. 325-339.
8. N. Arshadi, I. Jurisica, Feature Selection for Improving Case-Based Classifiers on High-Dimensional Data Sets. In *FLAIRS Conference* (2005), pp. 99-104.
9. G. Zhu, J. Hu, J. Qi, J. Ma, Y. Peng, An integrated feature selection and cluster analysis techniques for case-based reasoning. In *Engineering Applications of Artificial Intelligence* (Elsevier, 2015), pp. 14-22.
10. I. Kononenko, Estimating attributes: analysis and extensions of RELIEF. In *European conference on machine learning* (Springer, 1994), pp. 171-182.
11. T. Hong, Y. Liou, Attribute clustering in high dimensional feature spaces. In *International Conference on Machine Learning and Cybernetics* (IEEE, 2007), pp. 2286-2289.
12. P. Maji, Fuzzy-rough supervised attribute clustering algorithm and classification of microarray data. In *Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* (IEEE, 2011), pp. 222-233.
13. M. Masson, T. Denœux, RECM: Relational evidential c-means algorithm. In *Pattern Recognition Letters* (Elsevier, 2009), pp. 1015-1026.
14. T. Denœux, M. Masson, EVCLUS: evidential clustering of proximity data. In *Transactions on Systems, Man, and Cybernetics* (IEEE, 2004), pp. 95-109.
15. O. Kanjanatarakul, S. Sriboonchitta, T. Denœux, K-EVCLUS: Clustering large dissimilarity data in the belief function framework. In *International Conference on Belief Functions* (Springer, 2016), pp. 105-112.
16. K. Pearson, Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. In *Philosophical Transactions of the Royal Society of London.* (1896), pp. 253-318.