International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:4, No:1, 2010

Methods for Case Maintenance in Case-Based Reasoning

A. Lawanna and J. Daengdej

Abstract**—**Case-Based Reasoning (CBR) is one of machine learning algorithms for problem solving and learning that caught a lot of attention over the last few years. In general, CBR is composed of four main phases: *retrieve* the most similar case or cases, *reuse* the case to solve the problem, *revise* or adapt the proposed solution, and *retain* the learned cases before returning them to the case base for learning purpose. Unfortunately, in many cases, this retain process causes the uncontrolled case base growth. The problem affects competence and performance of CBR systems. This paper proposes competence-based maintenance method based on deletion policy strategy for CBR. There are three main steps in this method. Step 1, formulate problems. Step 2, determine coverage and reachability set based on coverage value. Step 3, reduce case base size. The results obtained show that this proposed method performs better than the existing methods currently discussed in literature.

Keywords**—**Case-Based Reasoning, Case Base Maintenance, Coverage, Reachability.

I. Introduction

CASE-Based Reasoning (CBR) is an algorithm of solving new problems based on the solutions of similar past problems. The well-known 4R processes of traditional CBR [1] are *retrieve*, *reuse*, *revise*, and *retain*. That is solving a problem by CBR involves:

Retrieve: Obtaining a problem description, measuring the similarity of the current problem to previous problems stored in a case base (or memory) with their known solutions,

Reuse: Reuse the solution of one of the retrieved cases, possibly after adapting it to account for differences in problem descriptions.

Revise: The solution proposed by the system is then evaluated (e.g., assessed by a domain expert).

Retain: The problem description and its solution can then be retained as a new case, and the system has learned to solve a new problem.

While there is a number of research issues related to all these 4 steps of CBR, one of the issues that catch large amount of attention of CBR researchers is degrading of CBR system's performance after a few runs. In this case, the "retain" process is the one that causes the uncontrolled case base growth which can result in dropping of competence and performance of the system for the next cycle.

Therefore, many CBR researchers develop the Case Base Maintenance (CBM) methods in order to response to this problem [2], [3]. The CBM methods relate to deleting cases, adding selected cases, or partitioning cases which are the theoretical and conceptual difference methods for accounting in the CBM. Particularly, those methods deal with the three main issues discussed in section II. Up to now, on one can guarantee which one is the best method. Some of them succeed to reduce cases but cannot preserve the competence of the system.

In response to this problem, we propose the Determining Coverage and Reacgability and Reducing Cases (DRCBM) method explained in section III. Our experiments and results are shown in section IV. Thereafter, we evaluate the three comparative studies based on the competence and performance criteria detailed in section V. finally, section VI is the summary conclusion.

II. Case base maintenance

A. Case deletion

Generally, deleting cases methods are developed for solving the uncontrolled case base growth. The oldest and simplest deletion is Random Deletion (RD) which can simply reduce cases but difficulty in preserving the competence while the high utility value of cases is deleted. Thus, Minton [4] proposed the Utility Deletion (UD) instead of RD. Conceptually it deletes the lowest utility value of cases based on Minton's equation. However, the competence of the system is still dropping. Therefore, Smyth and Keane (1995) proposed the Footprint Deletion and Footprint Utility Deletion (FD&FUD) which are claimed to be a competence preserving deletion policy [5]. The policy determines the coverage and reachability (C&R) set based on a simple nearest neighbor denotes:

Coverage of a case is the set of target problems that it can be used to solve.

Reachability of a target problem is the set of cases that can be used to provide a solution for the target.

The C&R set then can be categorized into a type hierarchy based on their coverage potential and adaptation power follows:

A. Lawanna, Department of Information Technology, Assumption University, Bangkok, Thailand Phone: 02-7191079; fax: 02-7191639; (e-mail: adtha@scitech.au.edu).

J. Daengdej, Department of Information Technology, Assumption University, Bangkok, Thailand Phone: 02-7191079; fax: 02-7191639; (e-mail: adtha@scitech.au.edu).

Pivotal cases which are generally outliers, being too isolated to be solved by any other case, affect competence when they are deleted.

Spanning cases, their coverage spaces span regions of pivotal cases. They do not affect the competency.

Support cases are a special class of spanning cases and do not affect the competence.

Auxiliary cases are the cases that do not affect the competence at all.

The deletion policy then selectively deletes cases from a case base guided by the classification of the cases until a limit on the case base size is reached. The algorithm was empirically shown to preserve the competency of a CBR system and to outperform a number of previous deletion based strategies. However, deleting a pivotal case may reduce the competence because by definition there is at least one problem that can no longer be solved, namely the problem that corresponds to the pivotal case itself. Of course, in practice there will be a range of problems in the region of the pivot which can no longer be solved [9].

B. Case addition

Zhu and Yang argued that FD&FUD policy is not guaranteed the competence to be preserved when auxiliary, spanning, or supporting cases are deleted because their similarity value may close to the case representative (or centroid). Another argument, they carried out one theory that proves the coverage value of FD&FUD decreases when numbers of cases are deleted. Therefore, they proposed a addition policy, cases in an original case base are repeatedly selected and *added* to an empty case base until a certain size limit is reached, producing an updated case base which high coverage guarantee (at least 63% coverage) by placing a lower bound on the competence of the resulting case base [6].

From our study, we found that case addition policy by Zhu and Yang can longer provide a high coverage value among cases for preserving the competence of the system. However, case addition can no longer preserve the performance of the CBR systems because of time complexity which relate to the operation, $O(n^2)$. Indeed, for each added case it is necessary to re-examine the whole original case base which can be fastidious. For case addition, the problem description and the system deduced solution form the case that is added [7].

C. Case partitioning

The partitioning policy consists of dividing the case base into several clusters. It enables for case selection, in an increasing manner, the attributes which are rich in information and which can cover the structure of the case base [8].

Overall, the cases in the initial case base are representative, accurate, and diverse. Each case is regarded as a cluster and itself is called key case [9] and we can partition the case library into several clusters by using the weighted distance metric such as decision tree [7] and K-Means clustering [2], [10]. The method partitions cases into clusters that can be converted to new smaller case-bases [11]. By consideration of the clustering group samples (cases) into partitions, such that samples within a cluster are similar to one another and dissimilar to samples in other clusters [12]. They can find the most representative cases for each cluster [13], [14], [15], [16]. However, we found that one of the drawbacks of partitioning is present during the classification and class selection procedure. When a border element is poorly classified, it is possible to have no answer while it could have been found in the neighbouring class.

D. Research Issues Related to CBM

According to the survey, we observed that the CBM methods deal with the following lists the main issues involve:

1). Problem Formulation

The CBM methods currently assume that the case-base contains a representative sample description of problems. This is reasonable since a CBR system could not be a good problem solver if the case-base were not representative.

From the investigation, we have found that most CBR researches use a single unknown value for their work [17] detailed in table I.

For example, according to case number 1, the company has sold a computer which has 16 MB of RAM, 2.5 GB hard disk and 15 inch monitor for 1,950. Similar interpretation is also applied to all other cases. Applied adaptation rule: the case is then adapted by reducing the size of monitor from 17 inch (case number 4) down to 15 inch, and the final price is reduced to 2,200.

TABLE I
A SET OF TARGET PROBLEMS

Case Number	RAM (MB)	Hard Disk (GB)	Monitor (inch)	Price (Dollars)
1	16	2.5	15	1,950
2	64	4.0	21	5,430
3	32	3.2	15	2,450
4	32	1.7	17	2,500
5	64	4.0	15	3,100
Problem1	32	1.7	15	?
Problem2	16	2.5	?	1,900
Problem3	16	?	15	1,500
Problem4	?	4.0	17	2,400
Problem5	32	4.0	?	?
Problem6	16	?	?	17,00
Problem7	?	?	21	2,100
Problem8	32	?	?	?
Problem9	?	?	?	2,200

In the traditional method, case solution answer to the problem is only on price of computer. In our work, we observed that a set of target problems is not based on only a single unknown value. In the real world phenomenon, it can involve the multi unknown value. For example, a problem that has 16 MB of RAM, 2.5 GB hard disk and ? inch monitor for 1,900 dollars or a problem that has 32 MB of RAM, ? GB hard disk and ? inch monitor for 2,000 dollars.

From the survey, we claimed that the traditional CBM methods [19] are no longer considering this situation. Relevant to the ability of CBM, we believe that it can solve this issue.

Therefore, in our paper we offer the maintenance method that can solve this problem before reducing cases.

2). Determining C&R set

This intense interest is highlighted by Smyth and Keane (1995). They believe that the key concept in categorizing cases is determining C&R set. Consequently, they define the coverage set as a case in the set of target problems that can be used to solve and the reachability set of a target problem is the set of cases that can be used to provide a solution for the target. Up to now, most of CBM methods determine C&R set before maintaining (deleting, adding, or partitioning) cases because they believe that it can provide the best case representatives.

Obviously, computing these sets for every case and target is impossible; the space of target problems is, in general, simply too vast. A more tractable solution is to assume that the case-base itself is a sample of the underlying distribution of target problems. Now, we can estimate the coverage of a case by the set of cases that can be solved by its retrieval and adaptation, and the reachability of a case by the set of cases that can bring about its solution. Smyth and Keane commented that coverage and reachability cannot be calculated because the possible set of problems is in general too vast [6]. Thus the assumption is made that the problem distribution in the case base is representative and a heuristic approach is used.

In Fig 1 is drawn in order to show example of C&R set where T is denoted as a set of target problems and S is a set of case solutions. There are five target problems $\{T_1\text{-}T_5\}$ and five case solutions $\{S_1\text{-}S_5\}$ which give the C&R set (e.g., (T_1, S_1), (T_2, S_1), (T_3, S_1), (T_4, S_1), (T_5, S_2), (T_5, S_3), (T_5, S_4), and (T_5,S_5)).

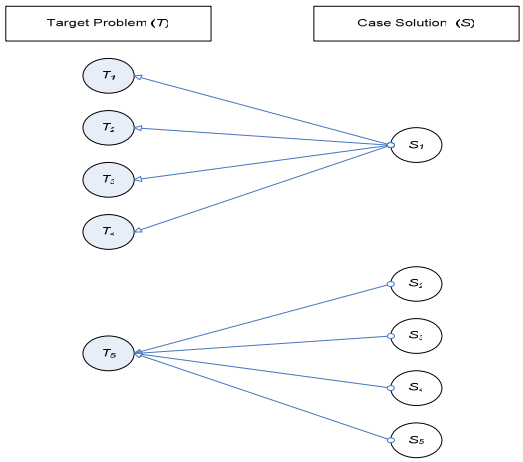

Fig. 1 C&R set

The traditional methods can determine C&R set (e.g., by nearest neighbor approach) based on a single unknown value. Our paper will show the determining C&R based on either a single unknown value or the multi unknown value.

3). Reducing a case base size

An interesting question is "between case deletion and case addition which one is the appropriate technique to maintain cases?" Bogaerts and Leake proposed the following four CBM techniques already implemented in IUCBRF [18]. Each is described by its policies for addition to and deletion from the case base.

From the survey, we have found that the deletion, addition, and many CBM methods can reduce case base size but could not preserve the competence and performance of the system. Therefore, the adaptation cost is required to fix these serious problems [5], [6].

The contribution of this paper is in the application of the case deletion strategy. C&R set to our knowledge has been studied at this level contrary to methods from case addition strategy by Zhu and Yang (1998) and deletion strategy by Smyth and Keane(1995). Consequently, we propose a case maintenance method by deleting the least quality cases in the case base. This quality of cases is based on the competence and performance measure.

Table II shows definition of terms used in our algorithm and equations.

III. The Proposed Determining C&R and Reducing Case Base Size (DRCBM) Method

TABLE II
DEFFINITION

Symbol	Quantity
C	Case
T	A set of formulated target problems
Q	A set of formulated target problems in different unknown value
$Q_{(\alpha,\alpha-n)}$	A number of possibility
N_t	The total number of formulated target problems
S	Case Solution
D	The candidate of the case solution
K	Deleted case
θ	The Obtained case
η	The initial case base
α	Attribute
$Q_{(\alpha,n)}$	A number of possibility

A. Step I: Formulate a set of target problems

More precisely, the acquisition is performed during a CBR session: the target problem is automatically solved by adaptation of the retrieved case and, after that, the solution is presented to the user who, depending on his/her expertise level, may be able to detect that the solution is not satisfactory and why that is not the case.

Traditionally, this optimization problem has been formulated**.** Many target problems require creative solutions. However, such problems are typically weakly-structured and underspecified (open-ended). We investigate the potential of analogical reasoning for this type of problems. We carry out

experiments in a CBR environment. We demonstrate that formulating a set of target problem with analogies leads to more competence and performance based on the Problem Formulation Algorithm.

Problem Formulation Algorithm

Let C = {C_1,C_2,…,C_n }be the set of cases in a case base;

C_n ={ $\alpha_1, \alpha_2, ..., \alpha_m$ } be the set of attributes in cases;

α_m = {n_1 , n_2 , …, n_i} be the set of unknown value in the attributes;

n_j ={ $Q_{(\alpha,\alpha-1)}, Q_{(\alpha,\alpha-2)}, ...Q_{(\alpha,\alpha-n)}$ } be a set of formulated target problems in different unknown value.

1. The case C_1 represents the attributes, α . Case represents attributes $\alpha_1, \alpha_2, ..., \alpha_m$.

2. The unknown value n_1 exists in the attribute of a case. Unknown value orders the set of question.

3. If there is a case with $Q_{(\alpha,\alpha-n)}$ where n ≠ o and n ≠ α , then

Select a case with $Q_{(\alpha,\alpha-n)}$

End

The Problem Formulation relies on "number of unknown value in the attribute of each case". Generally, number of possibility, $Q_{(\alpha,\alpha-n)}$ can be calculated by using (1).

$$Q_{(\alpha,\alpha-n)} = \frac{\alpha!}{(\alpha-n)!\,n!} \quad (1)$$

A set of formulated target problems can be calculated by using (2)

$$T = \sum \left(Q_{(\alpha,\alpha-1)}, Q_{(\alpha,\alpha-2)}, ...Q_{(\alpha,\alpha-n)} \right) \quad (2)$$

Finally, the total number of formulated target problems (N_t) can be calculated by using (3)

$$N_t = T * C \quad (3)$$

For example, if C=50 and α =4 then $Q_{(4,1)}$ =4, $Q_{(4,2)}$ =6, $Q_{(4,3)}$ =4, T=14, and N_t=700. The calculating process follows:

$$Q_{(4,1)} = \frac{4!}{(4-1)!1!} = 4;$$

$$Q_{(4,2)} = \frac{4!}{(4-2)!2!} = 6;$$

$$Q_{(4,3)} = \frac{4!}{(4-3)!1!} = 4;$$

$$T = \sum \left(Q_{(4,1)} + Q_{(4,2)} + Q_{(4,3)} \right) = 4+6+4=14 ;$$

$N_t = T*C = 14(50) = 700$

The different initial case base and attributes can be processed with the same method. The algorithms for properties with multiple value answers will be expanded and generalized in order to produce the results expected by the case builders.

Step II: Determining C&R

Viewing at how an individual case takes part in the problem solving process we observe that C&R set has an effect on its competence and performance. Intuitively, cases with large coverage sets seem likely to be giving off large competence contributions. Better case (large coverage value) can be reserved for the next CBM cycle. The best case (maximum coverage value) represents the cases with the least gap from target problem.

In the previous example, we present fig 2. for describing how to achieve the largest coverage set.

Fig 2 shows the super set of target problems; for instance, {T5, T6, T7}⊂{T1}, {T5, T8, T9}⊂{T2}, {T6, T8, T10}⊂{T3}, {T7, T9, T10}⊂ {T4}, {T11, T12, T13, T14}⊂ {T5}, {T11, T12, T13, T14}⊂{T6}, {T11, T12, T13, T14}⊂{T7},{T11, T12, T13, T14}⊂ {T8},{T11, T12, T13, T14}⊂{T9}, and {T11, T12, T13, T14}⊂{T10}. Thus, we found that {T1, T2,…, T4}can give the largest coverage set.

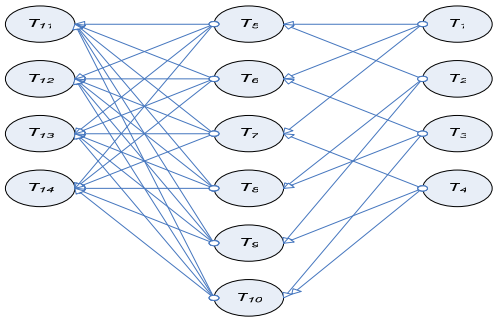

Fig.2 A super set

Consequently, we apply Case-Based Problem Solving to propose the case called a pair-wise association (T,S) where T is a problem and S is a solution of T . Solving a problem means associating a licit solution with it. Reasoning from cases means solving a problem called the target problem, using the case base.

Problem Solving Algorithm

T = new target problem

S = find solution (T)

D = find candidate (S)

The similarity measure of (T,S) is performed using the k-nearest neighbor's algorithm (k-NN). It is a simple method for classifying objects based on closest training examples in the feature space. An object is classified by a majority vote of its neighbors. k is a positive integer, typically small. If k =1, then the object is simply assigned to the class of its nearest neighbor. The k-nearest neighbors are determined according to a distance function like the Euclidian distance. To determine neighbors of an object, we calculate the distances from this object to the whole points in a reference data. Then we sort these distances in an ascending order and select top k points, these are called k-nearest neighbors[1].

What value of k is optimal? It is not necessarily an obvious solution. How to choose *k*? Do we use 1 nearest neighbor, 10 nearest neighbors, 50 nearest neighbors? The best choice of k depends upon C&R set; generally, larger values of *k* reduce the effect of noise on the classification, but make boundaries

between classes less distinct. However, in our research, we assume that noise data is not realized. Therefore, choosing a small value for k may lead the algorithm to over fit the data.

The ability to measure C&R is the key to understanding competence in CBR. This research investigates C&R, competence and problem-solving capacity of case base with one of its aims being to develop a method to create these aspects of a case-base and the group of cases within it. Coverage assumes a finite problem space and attempts to measure the number of points within this problem space covered by the case-base. This empirical coverage applies where the cases are represented by attribute with value providing a finite problem space. The coverage of each C&R is then measured by the coverage ratio.

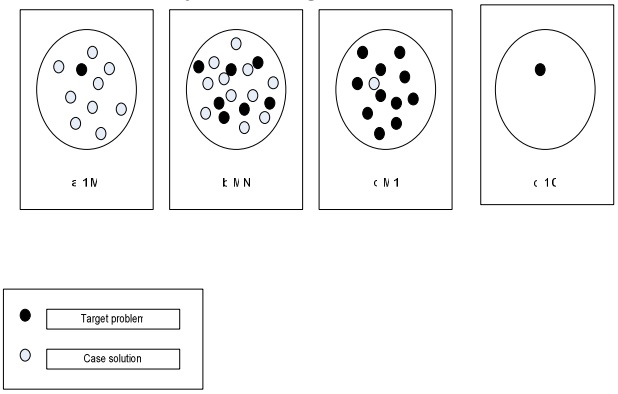

Fig. 3 C&R set

Fig 3 shows four types of C&R set which are described as follows:

Fig.3(a) shows that a single target problem can be solved by many case solutions. It presents low coverage ratio which equals 1:*M*, this situation is not appropriate because reachability set is high.

Fig. 3(b) shows that many target problems can be solved by many case solutions. It presents coverage ratio which equals *M:N*, also this situation is not appropriate because reachability set is still high.

Fig. 3(c) shows that many target problems can be soled by one case solution. It presents coverage ratio which equals *M*:1, this situation is the best because many target problems can be solved by one case.

Fig. 3(d) shows that a single target problem can not be solved by any case solution. This situation is not considered in this research because the possibility is low.

Thereafter, we demonstrate the coverage value for representing cases that affect to the competence and performance of the system. Fig. 4 shows the coverage ratio of the previous example which can be calculated using the C&R set.

For instance;

Coverage set (1) = 4 ; Reachability set (1) = 1;
Coverage set (2) = 6; Reachability set (2) = 2;
Coverage set (3) = 4; Reachability set (3) = 3;

Coverage ratio (1) = $\frac{4}{1} = 4$;

Coverage ratio (2) = $\frac{6}{2} = 3$;

Coverage ratio (3) = $\frac{4}{3} = 1.33$;

In order to provide a case base with good competence, its coverage ratio must be high and its reachability value must be low. Therefore, the largest coverage value of this example can be calculated by using (4) which equals 3.5

$$\text{Coverage value} = \frac{T}{\theta} \quad (4)$$

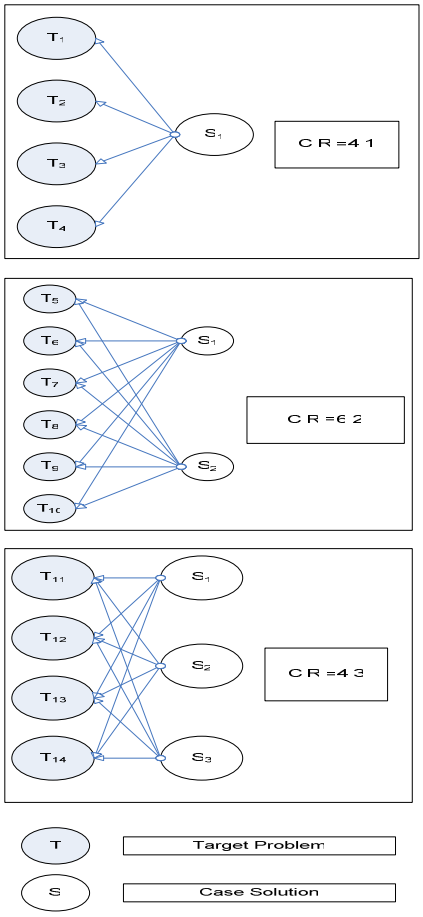

Fig. 4 The coverage ratio

Consequently, the coverage value is applied to lead the deletion of cases. Relevant to this fact, DRCBM method consists of reducing the case base size while maintaining a maximal competence detailed in step III..

B. Step III: Reducing case base size

In this step, we propose deletion policy in order to reduce the number of cases needed to learn. The main purpose of this method is to maximize the competence and performance of the CBR system and at a moment reduce, as much as possible, the size of case-base. The experiments using different domains,

most of them from the UCI repository, show that the deletion techniques can maintain the competence obtained by the initial case base. The DRCBM decides whether to delete a case or not by the deletion algorithm. The algorithm is motivated by the need to delete cases in order to maintain the competency of a case base at a reasonable size.

Deletion Algorithm

If there is a case with $Q_{(\alpha,\alpha-n)}$ where $n \neq o$ and $n \neq \alpha$ then

Select case with $Q_{(\alpha,\alpha-n)}$

EndIf there is a case with $Q_{(\alpha,\alpha-1)}$ then

Select a case with $Q_{(\alpha,\alpha-1)}$

EndIf

The algorithm aims the content of the case base for selecting which cases to keep and which cases to delete. Beside these, this algorithm ensures that $Q_{(\alpha,\alpha-1)}$ is formulated by choosing those features which maximally coverage and minimally reachability between the candidate cases. The ability of the DRCBM method to select optimal cases can be used to successfully reduce the case base without losing valuable information.

IV. Experiments and Results

In this section, we show the experimental result of E.coli data-set (336 cases, 5 attributes) which is available from the UCI Machine Learning Repository www.ics.uci.edu/~mlearn/ MLRepository .html)

TABLE III
A TYPE OF TARGET PROBLEM

$Q_{(5,1)}$	$Q_{(5,2)}$	$Q_{(5,3)}$	$Q_{(5,4)}$
5	10	10	5

Table III shows a set of formulated target problems which are { $T_1, T_2, \ldots, T_{30}$}. Also, we found that

$Q_{(5,4)} = \{T_1,T_2,..,T_5\}$; $Q_{(5,2)} = \{T_6,T_7,..,T_{15}\}$;

$Q_{(5,3)} = \{T_{16},T_{17},..,T_{25}\}$; $Q_{(5,1)} = \{T_{26},T_{27},..,T_{30}\}$;

Coverage set (1) = 5; Reachability set (1) = 1;
Coverage set (2) = 10; Reachability set (2) = 2;
Coverage set (3) = 10; Reachability set (3) = 3;
Coverage set (4) = 5; Reachability set (3) = 4;
Coverage ratio (1)= 5; Coverage ratio (2) = 5;
Coverage ratio (3) = 3.33; Coverage ratio(4) = 1.2;

Coverage value $= \frac{T}{\theta} = \frac{30}{5} = 6$;

Thereafter, we explain the problem solving method by applying k-NN algorithm based on a chosen distance function to measure similarity value between the query-instance and all the training samples. However, a new target problem is generated for example, T1 = (mcg = 0.5, gvh=0.5, aac=0.5, alm1=0.2, alm2=?) The Euclidean Distance between point problem and case1 is d(Problem and Sequence name of case 1)= sqt (0.5-0.49)2+(0.5-0.29)2+(0.5-0.56)2+(0.2-0.24)2=0.22 This form of computation is carried out for all the training samples. The minimum distance value interprets the maximum similarity among cases. According to this part, it results three situation as follow:

Situation i) M:1, many of case solutions can be provided as a set of case solution when Euclidean Distance is greater than 0.069.

Situation ii) 1:1 we found k=1, results SYGA ECOLI mcg = 0.51, gvh=0.49, aac=0.53, alm1=0.14, alm2=0.26 when Euclidean Distance(min)=0.069

Situation iii) 0:1 has not found in this experiment because all values of the attributes are not existed.

Concerning $Q_{(5,4)}$, it was listed the largest coverage value. On the other hand, $Q_{(5,1)}$ offers a lower coverage value.

After deleting cases, the following statistics are produced:

$$\text{Competence}(\%)=\left(1-\frac{\theta}{N_t}\right)100\% = \left(1-\frac{5}{10{,}080}\right)100\% = 99.95\%$$

$$\text{Reduction}(\%)=\left(1-\frac{\theta}{N_t+\eta}\right)100\% = \left(1-\frac{5}{10{,}080+336}\right)100\% = 99.95\%$$

Beside this, we do four experiments by initiating 50 cases and selecting 4,5,6, and 7 attributes respectively resulted in table IV.

Table IV, Concerning the $Q_{(\alpha,\alpha-1)}$, it was listed the largest coverage value. On the other hand, $Q_{(\alpha,\alpha-n)}$ offers a lower coverage value. For example, if α =4 then Coverage value $= \frac{T}{\theta} = \frac{14}{4} = 3.5$. After deleting cases, the following statistics are produced:

TABLE IV
ASSESSMENT OF THE PROPOSED METHOD

T	N_t	θ	K	C	Coverage value	Reduction (%)	Competence (%)
14	700	4	696	54	3.5	99.47	93.71
30	1,500	5	1,495	55	6	99.68	99.67
62	3,100	6	3,094	56	10.33	99.81	99.81
126	6,300	7	6,393	57	18	99.89	99.89

$$\text{Competence}(\%)=\left(1-\frac{\theta}{N_t}\right)100\% = \left(1-\frac{4}{700}\right)100\% = 93.71;$$

$$\text{Reduction}(\%)=\left(1-\frac{\theta}{N_t+\eta}\right)100\% = \left(1-\frac{4}{700+50}\right)100\% = 99.47$$

This form of calculating process is carried out for all the rest of experiments.

V. Evaluation

A “good” case base is able to solve target problems for as many queries as possible correctly and effectively. The criteria by which one can judge the effectiveness of a case base are given [6], [10]. The important criteria that contribute to the

evaluation of a case base are: *competence* and *performance*.

• Competence is the range of target problems that can be successfully solved by the reachability set. Coverage ratio must be high and its reachability rate must be low.

The results show the very good case base performance which is in relation to the performance and competence. This good performance is expressed through the decreasing retrieval time with high reduction rate. Therefore an optimal case base is obtained. The aim of our deletion techniques is to reduce the case base size while maintaining the competence and performance of the system.

$$\text{Coverage value} = \frac{T}{\theta} \quad (5)$$

$$\text{Competence (\%)} = \left(1-\frac{\theta}{T}\right)100\% \quad (6)$$

The competence concern with the range of target problems that a given system can solve, it also depends on the problem-solving ability of the system and must involve the retrieval and adaptation process of a system. The number of cases can be readily measured, but the problem of how to measure the problem-solving ability of a case in terms of its retrieval and adaptation characteristics is not so simple.

• Performance is the problem-solving time that is necessary to compute a solution for case targets. This measure is bound directly to adaptation power. In this paper, we focus reduction rate and accuracy.

$$\text{Reduction (\%)} = \left(1-\frac{\theta}{N_t+\eta}\right)100\% \quad (7)$$

Performance relies critically on the accuracy, precision, and the cases stored in the case base. Mostly CBR systems apply retrieval methods whose efficiency is based on the case base size, and under these conditions the addition of redundant cases degrade effectiveness of the system by increasing retrieval time.

Three different CBM methods are compared for this experimental study 1) FD&FUD 2) case addition with cases ordered according to their coverage values; 3) DCCBM with cases ordered according to their C&R set and the associated algorithm. In order to strengthen the comparison, six different datasets are used.

Iris, Ecoli, Acute Inflammations, Liver Disorders, Abolone, and Computer Hardware data-sets are available from the UCI Machine Learning Repository (www.ics.uci.edu/~mlearn/ MLRepository.html). The table V illustrates the comparison of the three methods using the four data-sets.

Table V shows the results for the FD&FUD, case addition and DRCBM method. Our results are positive. It can be clearly seen that the DCCBM method is more efficient than the other ones by achieving a better cases reduction rate with a finer competence for the four data-sets. The reduction rate given by the developed method is sensibly higher than the one given by the two traditional methods.

TABLE V
COMPARATIVE STUDY

Data set	Property	FD& FUD	Case Addition	DRCBM
Iris (150 cases, 4 attributes, 8,400 problems)	Obtained case	14	9	4
	Reduction(%)	99.84	99.89	99.95
	Coverage ratio	1	1.56	3.5
	Competence(%)	99.83	99.89	99.95
Ecoli (336 cases, 5 attributes, 10,080 problems)	Obtained case	30	19	5
	Reduction(%)	99.71	99.82	99.95
	Coverage ratio	1	1.58	6
	Competence(%)	99.70	99.81	99.95
Acute Inflammations (120 cases, 6 attributes, 7,440 problems)	Obtained case	62	39	6
	Reduction(%)	99.18	99.48	99.92
	Coverage ratio	1	1.59	10.3
	Competence(%)	99.17	99.48	99.92
Liver Disorders (345 cases, 7 attributes, 43,470 problems)	Obtained case	126	80	7
	Reduction(%)	99.71	99.82	99.98
	Coverage ratio	1	1.58	18
	Competence(%)	99.71	99.48	99.98
Abalone (4,177 cases, 8 attributes, 1,060,958 problems)	Obtained case	254	161	8
	Reduction(%)	99.98	99.98	100
	Coverage ratio	1	1.58	31.75
	Competence(%)	99.98	99.99	100
Computer Hardware (209 cases, 9 attributes, 106.590	Obtained case	510	323	9
	Reduction(%)	99.52	99.70	99.99
	Coverage ratio	1	1.58	56.67
	Competence(%)	99.52	99.70	99.99

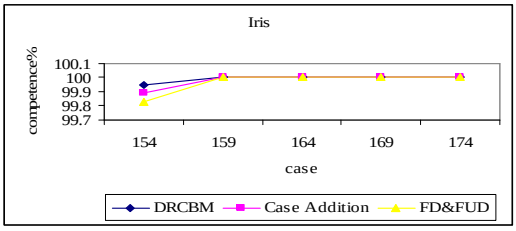

Fig. 5 The competence of comparative studies on Iris dataset

Fig. 5 shows the competence occurred for various cases of Iris dataset. As we can see from the figure, the competence is more in the case of DRCBM when compared to FD&FUD and case addition.

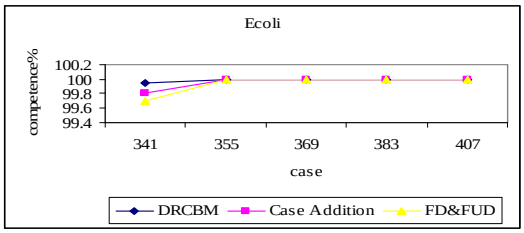

Fig. 6 The competence of comparative studies on Ecoli dataset

Fig. 6 shows the competence occurred for various cases of Ecoli dataset. As we can see from the figure, the competence is less in the case of FD&FUD and case addition when compared

to DRCBM.

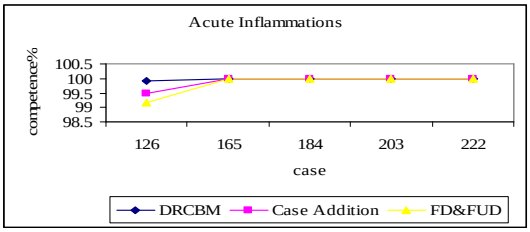

Fig. 7 The competence of comparative studies on Acute Inflammations dataset

Fig. 7 shows the competence occurred for various cases of Acute Inflammations dataset. As we can see from the figure, the competence is less in the case of FD&FUD and case addition when compared to DRCBM.

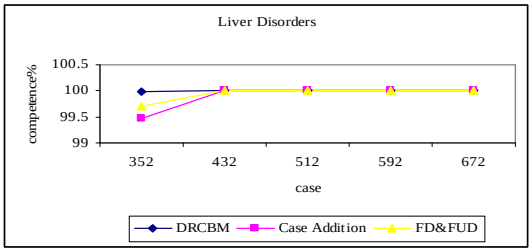

Fig. 8 The competence of comparative studies on Liver Disorders dataset

Fig. 8 shows the competence occurred for various cases of Acute Inflammations dataset. As we can see from the figure, the competence is less in the case of FD&FUD and case addition when compared to DRCBM.

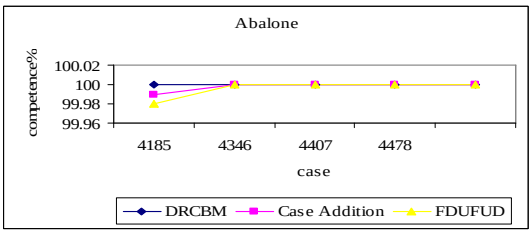

Fig. 9 The competence of comparative studies on Abalone dataset

Fig. 9 shows the competence occurred for various cases of Abalone dataset. In this case, the competence is less in the case of FD&FUD and case addition when compared to DRCBM which equals 100%.

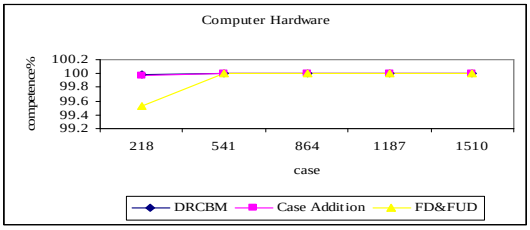

Fig. 10 The competence of comparative studies on Computer Hardware dataset

Fig. 10 shows the competence occurred for various cases of Computer Hardware dataset. As we can see from the figure, the competence is less in the case of FD&FUD and case addition when compared to DRCBM which equals 99.99%.

VI. Conclusion

This paper demonstrated that while traditional CBM methods are effective in controlling the uncontrolled case base size growth from a competence and performance perspective, they may lead to time complexity in many CBR systems. The solution proposed can formulate a set of target problems based on a single unknown value and multi unknown value. The set of target problems interprets size of C&R set which results four types of the coverage ratio (coverage set : reachability set) which are (1:M), (M:N), (M:1), and (1:0). We found that the best competence requires many target problems solved by one case solution (M:1). Thereafter, the coverage value will be calculated. The maximum coverage value will be used for guiding the DRCBM method in order to delete cases while preserving the performance of the systems.

The proposed method was evaluated by using two traditional CBM methods (FD&FUD and case addition) and six datasets. The obtained results were positive in terms of case base reduction size and best competence.

References

[1] D.W. Aha, "Feature Weighting for Lazy Learning Algorithms," Data Mining Perspective. Norwell. Mass.: Kluwer., pp.13-32., 1998.
[2] Q. Yang and J. Wu, "Keep It Simple: A Case-Base Maintenance Policy Based on Clustering and Information Theory," Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence, 102-114, 2000.
[3] R. Thomas and B. Roth, "Knowledge Maintenance of Case-Based Reasoning Systems—The SIAM Methodology," volume 262 of Dissertationen zur Künstlichen Intelligenz. Akademische Verlagsgesellschaft Aka GmbH / IOS Press., Berlin, Germany, 2003.
[4] S. Minton, "Qualitative Results Concerning the Utility of Explanation-Based Learning," Artificial Intelligence, vol 42, pp. 363-391, 1990.
[5] B. Smyth and M. Keane, "Remembering to forget: A Competence-Preserving Case Deletion Policy for Case-Based Reasoning Systems," In Proceedings of the 13th International Joint Conference on Artificial Intelligence., pp. 377-382, 1995.
[6] J. Zhu and Q. Yang, "Remembering to add: Competence-preserving case-addition policies for case based reasoning," 1998.
[7] S. Massie, S. Craw, and N. Wiratunga, "Complexity-guided case discovery for case based reasoning," In: The Twentieth National Conference on Artificial Intelligence, pp. 216–221, 2005.
[8] K. Haouchine, B. Chebel-Morello, and N. Zerhouni, "Competence-Preserving Case-Deletion Strategy for Case-base Maintenance.," *Uncertainty, Similarity, andKnowledge Discovery in Case-Based Reasoning workshop. 9th European Conference on CBR.*
[9] G. Cao, S. Shiu, and X. Wang X, "A fuzzy-rough approach for case base maintenance Cataltepe Z, Abu-mostafa YS, Magdon-ismail M (1999) No free lunch for early stopping. Neural Computation, pp. 11:995–1009, 2001.
[10] R. W. Lucky, "Automatic equalization for digital communication," *Bell Syst. Tech. J.*, vol. 44, no. 4, pp. 547–588, Apr. 1965.
[11] M. Salam´o and E. Golobardes, "Global, local and mixed rough sets case base maintenance techniques," In: Proceedings of the 6th Catalan Conference on Artificial Intelligence, IOS Press, pp 127–134, 2004.
[12] N. Arshadi and I. Jurisica, "Maintaining Case-Based Reasoning Systems: A Machine Learning Approach," Advances in Case-Based Reasoning: Proc. Seventh European Conf., pp. 17-31, 2004.
[13] R. Pan, Q. Yang, J.J. Pan, and L. Li , "Competence Driven Case-Base Mining," *AAAI .2005*, pp. 228-233, 2005.
[14] S.S. Wang and Yeung., "Transferring Case Knowledge to Adaptation Knowledge: An Approach for Case-base Maintenance," Computational Intelligence., 2001.

[15] C. Yang, R. Orchard, B. Farley, and M. Zaluski M, "Authoring Cases from Free-Text Maintenance Data," Machine Learning and Data Mining in Pattern Recognition , Springer Berlin, Heidelberg., 2004.
[16] N. Zhi-wei, L. Yu, and L. Feng-gang, "Case base maintenance based on outlier data mining," Proc.4th Intl.Conf.on Machine Learning and Cybernetics., Guangzhou., pp. 2,861-2,864, 2005.
[17] J. Daengdej, "Adaptable Case Base Reasoning Techniques For Dealing With Highly Noise Cases," The University of New England, 1998.
[18] S. Bogaerts and D. Leake, "IUCBRF Lesson: Case Base Maintenance Policies," 2005.
[19] B. Smyth, "Case–base maintenance," In Proceedings of the 11th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, Springer-Verlag, 1998.

Adtha Lawanna has completed B.Sc in Chemistry from Chiang Mai University, Chiang Mai, Thailand in 1996, the first M.Sc in Chemical Technology from Chulalongkorn University, Bangkok, Thailand in 2000, and the second M.Sc. in Information Technology from Assumption University, Bangkok, Thailand in 2004. He is now working as a Lecturer in Department of Information Technology, Faculty of Science and Technology of Assumption University, Bangkok, Thailand and also doing his Ph.D(IT).

Dr. Jirapun Daengdej is working as Assistant Professor Associate Dean in Faculty of Science and Technology at Assumption University, Bangkok, Thailand. His research interest is in Case-Based Reasoning, Software Engineering&Testing, and Artificial Intelligent.