Modeling Reuse on Case-Based Reasoning with Application to Breast Cancer Diagnosis

Carles Pous[1], Pablo Gay[1], Albert Pla[1], Joan Brunet[2], Judit Sanz[3,4], Teresa Ramon y Cajal[3], and Beatriz López[1]

[1] Universitat de Girona, Campus Montilivi, edifice P4, 17071 Girona
carles.pous@udg.edu,beatriz.lopez@udg.edu
http://eXiT.udg.edu/
[2] Institut Català d'Oncologia, Girona
Jbrunet@iconcologia.net
http://www.ico.net/
[3] Hospital de la Santa Creu i Sant Pau
juditsanzbuxo@gmail.com,tramon@santpau.cat
http://www.santpau.es
[4] Hospital Universitari Sant Joan de Reus

Abstract. In the recent years, there has been an increasing interest on the use of case-based reasoning (CBR) in Medicine. CBR is characterized by four phases: retrieve, reuse, revise and retain. The first and last phases have received a lot of attention by the researchers, while the reuse phase is still in its infancy. The reuse phase involves a multi-facet problem which includes dealing with the closeness to the decision threshold used to determine similar cases, among other issues. In this paper, we propose a new reuse method whose decision variable is based on the similarity ratio. We have applied the method and tested in a breast cancer diagnosis database.

1 Introduction

Case-Based Reasoning (CBR) is experiencing a rapid grow and development in Medicine [3], mainly due to their relationship with evidence-based practice: case-based reasoning focuses on the reuse of previous experiences or cases to solve new problems [4,10]. A case-based system accompanies a conclusion of a problem with the cases found similar in a case base, thus presenting compelling justification for the decision.

Case-based reasoning consists of four phases, that are repeated for each new situation [1]. The first step is to seek for past situation or situations similar to the new one (*Retrieval*). In this stage it is necessary to define a metric function and decide how many cases to retrieve. *Reuse* is the second phase and it consists in using the extracted cases to propose a possible solution. Once the solution is proposed it has to be revised (*Revise*) by a human expert or automatically. After the revision process, it has to be decided if it is useful to keep the new situation in the case base in order to help on the future diagnosis of new situation. This last stage is known as *Retain*. Case-based reasoning has evolved in the last years, and there are currently a big number of available techniques for each phase of the system.

D. Dochev, M. Pistore, and P. Traverso (Eds.): AIMSA 2008, LNAI 5253, pp. 322–332, 2008.

Nevertheless, most of the advances have been achieved at the retrieval and retain phase [14,18,6,15]. In the reuse phase, advances have been obtained depending on the system purpose: diagnosis, classification, tutoring and planning (such as therapy support). Regarding diagnosis and classification, most of the systems rely on adaptation methods that consist on copying the solution of the most similar case or a combination of them (the class of the majority of them, for example). More elaborated methods have been defined for therapy, guidelines and protocols (planning tasks). For example, in [12] a reuse phase for breast cancer treatment protocol decision is based on reformulating the problem in subparts (paths) and then elaborate an adaption for each part. In [9] a particular adaptation method is proposed for deciding the reuse of an antibiotic treatment, that depends on the features used in the retrieval phase. Adaptation methods for planning are influenced by the work done in other application domains, as software design [8]. However, most of the methods are problem specific [16,17]. For example, in [11] the method proposed is based on a library of feature adaptation plans.

According to [12], adaptation in Medicine is complex, needs to deal with the lack of relevant information about a patient, the applicability and consequences of the decision, the closeness to the decision thresholds and the necessity to consider patients according to different viewpoints. In fact, [17] points out about the difficulty on giving autonomy to this CBR step in Medicine. So, the authors propose a user-interactive approach for this CBR stage. In this sense, we believe that all of the approaches found in the literature are complementary and necessary: the use of generic approaches (as the majority rule), specific knowledge (as a library of feature adaptation plans), and user interaction.

In this paper, we introduce a new a simple generically applicable approach to perform case adaptation. Our research concerns reuse for classification, since our goal is to diagnose whether a patient has cancer or not. Our method, takes into account the closeness of the retrieved cases, refining the approach provided by the majority rule. The simplicity of our approach means that the development cost required to apply it is low. We have experimentally tested our method with two cancer data bases, and we have got uncourageous results.

The paper is structured as follows. First, a brief description of an elemental CBR system is given. Section 3 describes the main contribution of the paper, the proposed reuse methodology. Afterwards, the description of the experiments done are reported, following with the results achieved. Finally, we end with the conclusions and future work.

2 Case-Based Reasoning Approach

We are faced with the problem of building a CBR system to support breast cancer diagnosis. As a first approach, we have developed a system similar to the one described in [5],which includes the basic functions. So we have the complete CBR cycle covered with simple functions that enable the study of the reuse stage performance.

Concerning to the retrieve stage, we have used the distance based on the Mathematical distance for numeric attributes and the Hamming distance for categorical ones. Regarding missing values, statisticians have identified three situations [20]. Missing completely at random (MCAR) is the case when the probability of missing a value is the same for all variables. On the other hand, not missing at random (NMAR) values occurs when the probability of missing a value is also dependent on the value of the another missing variable. At last, the missing values classified as missing at random (MCR) happens when the probability of missing a value is only dependent on other variables. We do not distinguish between MCAR,NMAR or MCR values. So, when the attribute of the test case and the memory case are both missed, the distance is considered 0. Second, when the value of the attribute either the test case or the memory case are missed, the distance is also taken as 0. Finally, when both are missed, the corresponding distance function is applied (Mathematical or Hamming).

Thus, the *local distance* between two cases concerning attribute a with a value of x_a and y_a for each case is given by the following equation:

$$d(x_a, y_a) = \begin{cases} 0 \quad if\ either\ x\ or\ y\ are\ unknown \\ Hamming(x_a, y_a) \quad if\ x_a = y_a\ and\ a\ categorical \\ |x_a - y_a| \quad if\ x_a = y_a\ and\ a\ numerical \end{cases} \tag{1}$$

where

$$Hamming(x_a, y_a) = \begin{cases} 1 \quad if\ x_a \neq y_a \\ 0 \quad if\ x_a = y_a \end{cases} \tag{2}$$

The *global distance* is the average of local distances obtained for all the attributes. Finally the similarity between a test case T and a memory case C, $sim(C, T)$, is computed as the inverse of the global distance as follows:

$$sim(C, T) = 1 - \frac{\sum_{i=1}^{n} d(x_i, y_i)}{n} \tag{3}$$

where n is the amount of attributes in the application.

Note, then, that we are using a very simple data retrieval phase. In this paper we do not focus on optimize the number of features to use, or on investigating different similitude functions and attributes weighting. We are focussing on the reuse methods.

3 Reuse Method

The retrieval phase returns a set of k similar cases, $C_1, \ldots, C_k$ to the current test case T. Often they are ranked according to the similarity degree (i.e. $sim(C_1, T) > sim(C_2, T) > \ldots > sim(C_k, T)$). Then, in the reuse phase the solution to the problem posed by T should be computed. Particularly, when dealing with a classification problem, the class corresponding to T should be determined.

In the particular case of breast cancer diagnosis, two classes conforms the solution space of a problem: cancer (+) and no cancer (-). According to the

majority rule, the new problem is classified in the same class than the majority of the retrieved cases. However, in domains like the one we are dealing with, it is the case that in most situations the amount of positive and negative cases retrieved are the same. So, the majority rule is not a valid classification criteria.

An alternative approach is the one proposed by Bilska-Wolak and Floyd who define a decision variable, DV_{bw}, as the ratio of "+" cases in $C_1, \ldots, C_k$ to all similar cases [5]. This definition represents an intuitive approach for describing the likelihood of cancer in a case. Thus,

$$DV_{bw} = \frac{\text{Number of + cases in } \{C_1, \ldots, C_k\}}{k} \tag{4}$$

Then, the decision variable is compared against a given threshold τ. Whenever the current value of the decision variable equals or goes beyond this threshold, the test case T is classified as positive or not.

This approach presents several disadvantages regarding the closeness degree of the similar cases to the test case, as shown in the following example. Let us suppose that we have four similar cases, two positives and two negatives: $C_1^+, C_2^+, C_3^-, C_4^-$. In this case, the decision variable according to [5] is $DV_{bw} = \frac{2}{4} = 0.4$. If τ is set to 0.5, the test case T will have the "+" class assigned. This decision is independent of the similarity degree of the cases. For example, in Table 1 several possibilities are given. In the first one, when all similarities are the same, the results cannot be improved. In the second situation, it is clear that positive cases are closer to the test case that the negative ones, and the decision variable should catch that. Conversely, in the third analyzed scenario, negative cases are more likely. Finally, in the last situation, as in the first scenario, there is no room for precision on the decision variable.

Table 1. Different situations with the same DV_{bw} value

$sim(C_1^+, T)$	$sim(C_2^+, T)$	$sim(C_3^-, T)$	$sim(C_4^-, T)$	DV_{bw}	T class ($\tau = 0.5$)
0.5	0.5	0.5	0.5	0.5	+
0.8	0.8	0.5	0.5	0.5	+
0.4	0.4	0.5	0.5	0.5	+
0.4	0.5	0.4	0.5	0.5	+

To take into account the closeness degree of similarity between the test and a memory case in the decision variable we propose an alternative definition as the ratio of similarities of + cases to the addition of the similarities of all similar cases. Formally,

$$DV_{pous} = \frac{\sum_{C_i^+} sim(C_i^+, T)}{\sum_{i=1}^{k} Sim(C_i, T)} \tag{5}$$

where C_i^+ are the cases in $C_1, \ldots, C_k$ which belong to the + class (suffering cancer).

As it is possible to observe in Table 2, our definition of the decision variable captures the closeness of the similarity to the test case involved in positive and negative cases. So, with the same threshold $\tau = 0.4$ the results are slightly different than the ones obtained with the DV_{bw} (see Table 1).

Table 2. Different situations with the same DV_{pous} value

$sim(C_1^+, T)$	$sim(C_2^+, T)$	$sim(C_3^-, T)$	$sim(C_4^-, T)$	DV_{pous}	T class ($\tau = 0.5$)
0.5	0.5	0.5	0.5	0.5	+
0.8	0.8	0.5	0.5	**0.6**	+
0.4	0.4	0.5	0.5	**0.4**	-
0.4	0.5	0.4	0.5	0.5	+

4 Experimental Set-Up

We have implemented our simple CBR system with the reuse method in Java. In order to test our methodology, we have used two data sets: The Breast Cancer Wisconsin (Diagnostic) Data Set [19] and our own cancer data set (HSCSP). The former is composed by 699 instances, 100 attributes each (integer values). Features are computed from a digitized image of a fine needle aspirate of a breast mass. They describe characteristics of the cell nuclei present in the image. There are some missing values.

The other data set used was provided by the Hospital de la Santa Creu i Sant Pau (HSCSP) from Barcelona. It consists of 871 cases, with 628 corresponding to healthy people and 243 to women with breast cancer. There are 1199 attributes for each case. They correspond to people habits (smoker or not, diet style, sport habits,...), desease characteristics (type of tumor,...), and gynaecological history among others. Since there are redundant, wrong and useless information a preprocess was carried out. Also, the preprocess was used to obtain data corresponding to independent individuals, since there are patients in the database that are relatives. After this operations the final database was constituted of 612 independent cases, with 373 patients and 239 healthy people. From 1199 attributes, 680 are considered useless, redundant or too much incomplete. About the rest of attributes, 192 are discrete, 279 numeric and 34 text (such as postal address, etc.). We have used discrete and numeric attributes for our experimentation (471). Another preprocessing step has also been applied to normalize numerical values.

As our method depends on two parameters, the number of cases retrieved (k) and the threshold used to classify the case according to the corresponding decision variable (τ), we have defined several experiments to be carried out varying them. Particularly we have varied each parameter as follows:

- k: from 1 to 10, step 1
- τ: from 0.1 to 1.0, step 0.1.

A cross-validation methodology has been followed. Figure 1 illustrates the process when using a stratified cross validation methodology with a fixed length of

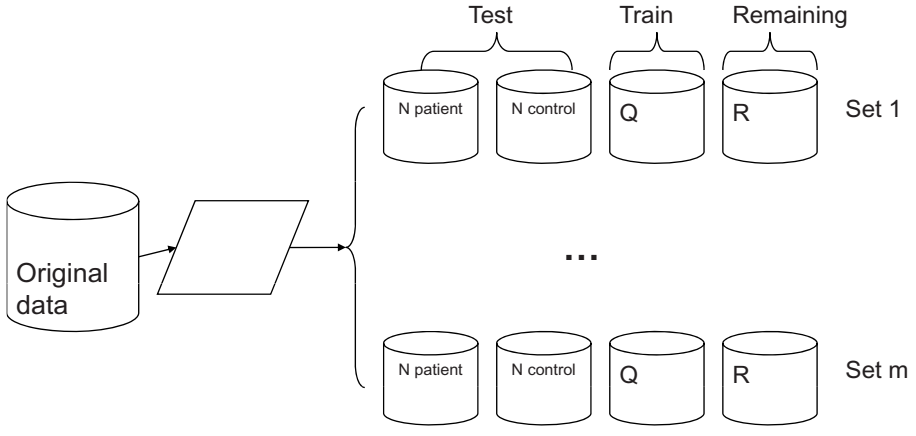

Fig. 1. Data sets generated for stratified cross validation on HSCSP data

test files (N cases per file) and a fixed number of data sets (m). So, from each original data set, up to 10 different training and testing sets have been generated. So, a run have been computed with each data set and each pair of parameter values.

5 Results

We have analyzed the four different possible outcomes: true positives (TP), true negatives (TN), false positive (FP) and false negative (FN). True positive and negative are the correct classifications. A false positive occurs when the outcome is incorrectly predicted as cancer (+); conversely, a false negative occurs when the outcome incorrectly predicts a no-cancer value (-). Next, the true positive rate (tp), and the false positive rate (fp) have been computed as follows:

$$tp = \frac{TP}{TP + FN} \tag{6}$$

$$fp = \frac{FP}{FP + TN} \tag{7}$$

The true and false positive rates have been used to visualize the results in ROC graphs. These graphs have been used in cost-sensitive learning because of the ease with which class skew and error cost information can be applied to them to yield cost-sensitive decisions [7]. In the x-axis, the fp values are plotted, while in the y-axis, the tp ones. Any point in the two-dimensional graph represents an experiment result. The ideal situation would be a system with tp=1 and fp=0, that is a point located just at the right upper part of the graphic.

In order to average the results of the cross validation process, we have followed the threshold averaging methodology explained in [7]. Thus, since the Pous method have two parameters, first we have averaged the results obtained for each k value given a threshold τ. Next, we have averaged the results of 10

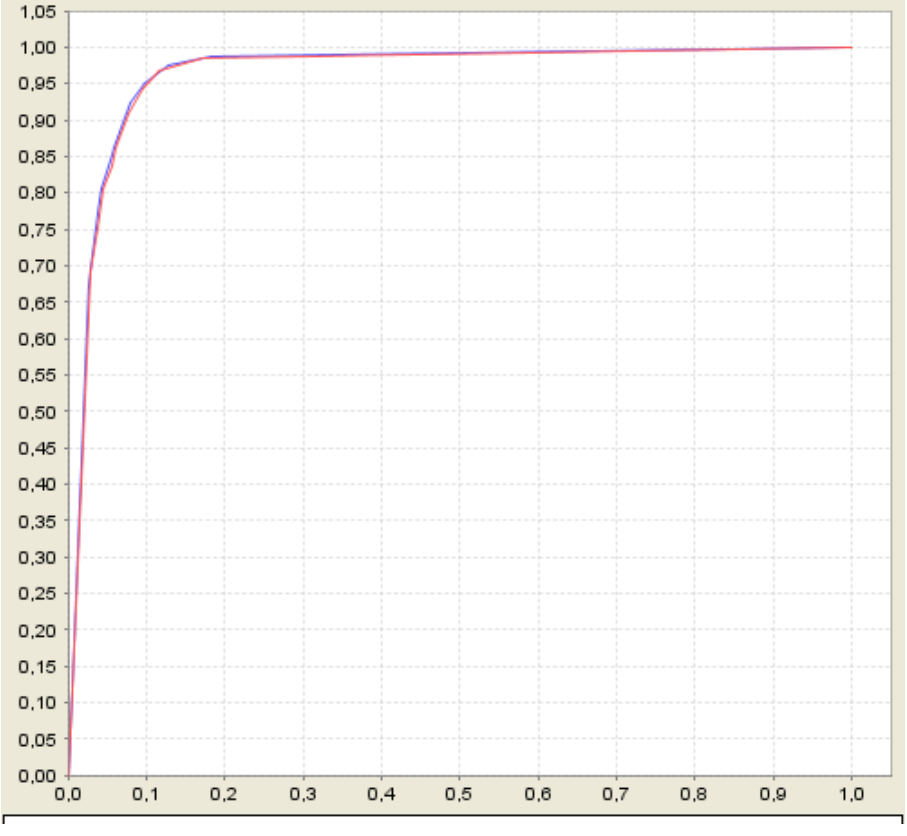

Fig. 2. Results on the Winsconsin data set. fp on the x-axis and tp on the y-axis.

runs for every possible τ. So, any point at the ROC graph corresponds to the results of the method according to the a given τ value.

The results on the Wisconsin data set are shown in figure 2. As it is possible to observe, the tp rate is above the 95% in almost all the ocassions. Regarding the HSCSP data set, results are shown in figure 3. With this data set, we have obtained a lower tp rate than with the Wisconsin data set. However, the complexity of the data set is different (10 attributes versus 471). As a first approach to the case-based breast cancer diagnosis, we believe that these results are good enough: With a $\tau = 0.7$, we got a tp (sensitivity) of 70 % while maintaining the fp (1-specificity) rate to 20 %. Of course, there is room for improvement with the incorporation in a future work of more appropriate retrieval, attribute selection and other methods.

6 Related Work

This section describes the main papers related to the adaption stage of a CBR cycle, when used in several medicine domains. Our starting point has been the

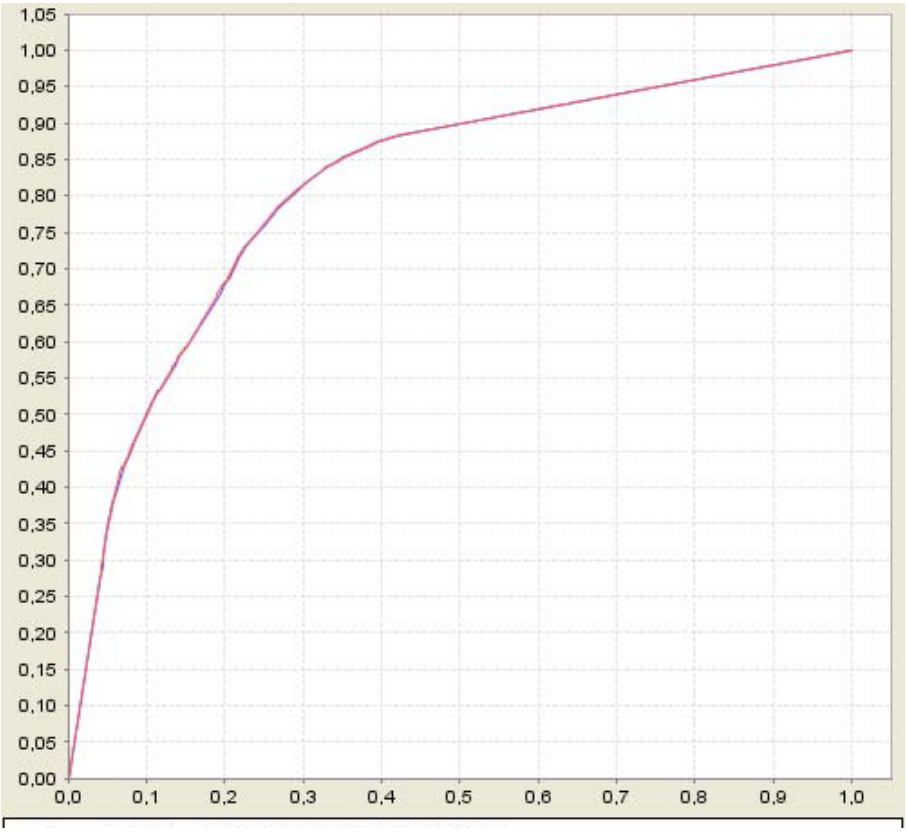

Fig. 3. Results on the HSCSP data set. fp on the x-axis and tp on the y-axis.

work of Bilska-Wolak and Floyd on breast cancer [5]. The authors propose the probabilistic measure presented in equation 4. As a previous step, they use a similarity threshold SIM, empirically estimated, in order to decide if a case is similar or not to the current case. The estimation is based on the goodness of the distance function used in the retrieval phase according to an exhaustive analysis of all the features. In their case, this approach is feasible, since they are dealing with a reasonable number of variables (10). However, in other scenarios, as it is our, this exhaustive approach is unfeasible. Regarding the reuse phase, the differences between our method is expressed in equations 4 and 5.

Our methodology is tightly related to weighted knn-methods. Such methods follow a voting scheme to decide the labeling of a test case, as the majority rule, but using a weight function for every vote [13]. When taking the weight function as the similarity one, we have the same results. Note, however, that we are not just assigning the class with the high aggregated weight, but computing the ratio of the weights of cancer that matched the test case to the total number of k matched cases.

Concerning to reuse methods for CBR systems in classification tasks, in [11] a knowledge-based planning is proposed as a general method. The execution of a case adaptation plan leads to the diagnosis of multiple diseases for a single problem. This mechanism is based on a network of feature adaption plans, that is, specific knowledge costly to acquire.

On the other hand, in [2] a sophisticated adaptation method is presented for individual diabetes prognosis for long-term risks. In some sense, Armengol and her colleagues mix in the DIRAS CBR system the retrieval and the reuse phases. They are continuing refining the case selection from the retrieval phase according to the most discriminant feature of the problem at hand. When all the selected cases belong to the same class, the new case is labeled with it.

[12] develops an adaptation method based on paths. That is, any problem is reformulated in subparts. So, according to the retrieval results of each of the subparts,the corresponding adaptation method is applied to obtain a final solution. The adaptation method described in [12] is defined for breast cancer treatment protocols.

Regarding breast cancer data, most of the CBR systems focus on image analysis, that is, the study of mammographies. Although mammographies are important (they reduce mortality up to the 30-40% [5]), most of the biopsies analyzed are benign (75-80%). So, developing decision support systems based on other clinical data is important in order to reduce annoying and costly biopsies. Most of the CBR systems dealing with non-image data, however, have elaborated retrieval functions, with attribute selection and weighted mechanism, and few attention is paid to the reuse phase. For example, in [6] a nonparametric regression method is introduced for estimating a minimal risk of the distance functions. Our future work includes the incorporation of retrieval methods like this.

7 Conclusions and Discussion

There is an increasing interest on the use of CBR techniques for medical diagnosis. However, most of the advances concentrate on the retrieval phase, in which similar cases are selected from memory. Regarding the reuse phase, researchers have focalized on planning tasks, as therapy or protocols, while for diagnosis the majority rule seems to be the most simple and general method utilized.

In this paper, we have presented a new general method for reuse, that is also simple to compute. It takes into account the similarity degree of the selected cases when determining the class of the new case. The method has been tested in two cancer databases: Wisconsin (from the UCI repository) and HSCSP (our own), and the results obtained are encouraging.

As a future work, we wish to complement our general reuse method, with other knowledge specific methods and user interaction abilities, according to the requirements of the medical applications. From the complementarity of all of these techniques should arise a reuse stage for CBR useful for physicians. We are also planning to integrate more accurate retrieval functions, as well as other attribute selection and weighing methods in our initial CBR prototype.

Acknowledgments. This research project has been partially funded by the Spanish MEC project DPI-2005-08922-CO2-02, Girona Biomedical Research Institute (IdiBGi) project GRCT41 and DURSI AGAUR SGR 00296 (AEDS).

References

1. Agnar, A., Plaza, E.: Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI Communications 7(1), 39–59 (1994)
2. Armengol, E., Palaudáries, A., Plaza, E.: Gindividual prognosis of diabetes long-term risks: a cbr approach. Methods on Information in Medicine 40(1), 46–51 (2001)
3. Bichindaritz, I., Marling, C.: Case-based reasoning in the health sciences:what's next? Artificial Intelligence in Medicine (36), 127–135 (2006)
4. Bichindaritz, I., Montani, S., Portinale, L.: Special issue on case-based reasoning in the health sciences. Appl. Intelligence (28), 207–209 (2008)
5. Bilska-Wolak, A.O., Floyd Jr., C.E.: Development and evaluation of a case-based reasoning classifier for prediction of breast biopsy outcome with bi-radstm lexicon. Medical Physics 29(9), 2090–2100 (2002)
6. Dippon, J., Fritz, P., Kohler, M.: A statistical approach to case based reasoning, with application to breast cancer data. Computational Statistics and Data Analysis 40, 579–602 (2002)
7. Fawcett, T.: Roc graphs: Notes and practical considerations for data mining researchers. Technical Report HPL-2003-4, HP Labs (2003)
8. Gomes, P., Pereira, F.C., Carreiro, P., Paiva, P., Seco, N., Ferreira, J.L., Bento, C.: Case-based adaptation for uml diagram reuse. In: Negoita, M.G., Howlett, R.J., Jain, L.C. (eds.) KES 2004. LNCS (LNAI), vol. 3215, pp. 678–686. Springer, Heidelberg (2004)
9. Heindl, B., Schmidt, R., Schmid, R.G., Haller, M., Pfaller, P., Gierl, L., Pollwein, B.: A case-based consiliarius for therapy recommendation (icons): computer-based advice for calculated antibiotic therapy in intensive care medicine. Computer Methods and Programs in Biomedicine 52(2), 117–127 (1997)
10. Holt, A., Bichindaritz, I., Schmidt, R., Perner, P.: Medical applications in case-basd reasoning. The Knowledge Engineering Review 20(3), 289–292 (2005)
11. Hsu, C.-C., Ho, C.-S.: A new hybrid case-based architecture for medical diagnosis. Information Sciences 166, 231–247 (2004)
12. Lieber, J., d' Aquin, M., Badra, F., Napoli, A.: Modeling adaptation of breast cancer treatment decision protocols in the KASIMIR project. Applied Artificial Intelligence (28), 261–274 (2008)
13. Macleod, J.E.S., Luk, A., Titterington, D.M.: A re-examination of the distance-weighted k-nearest neighbor classification rule. IEEE Transactions on Systems, Man and Cybernetics 17(4), 689–696 (1987)
14. Sollenborn, M., Nilsson, M.: Advancements and trends in medical case-based reasoning: An overview of systems and system development. In: American Association for Artificial Intelligence, Proceedings of the 17th International FlAIRS Conference, Special Track on Case-Based Reasoning, pp. 178–183 (2004)
15. Perner, P.: Concepts for novelty detection and handling based on a case-based reasoning process scheme. In: Perner, P. (ed.) ICDM 2007. LNCS (LNAI), vol. 4597, pp. 21–33. Springer, Heidelberg (2007)

16. Salem, A.-B.M.: Case-based reasoning technology for medical diagnosis. In: Proc. World Academy of Science, Engineering and Technology, vol. 25, pp. 9–13 (2007)
17. Montani, S., Bellazzi, R., Portinale, L., Gierl, L., Schmidt, R.: Casebased reasoning for medical knowledge-based systems. International Journal of Medical Informatics 64, 355–367 (2001)
18. Shie, J.-D., Chen, S.-M.: Feature subset selection based on fuzzy entropy measures for handling classification problems. Applied Artificial Intelligence (28), 69–82 (2008)
19. Wolberg, W., Street, N., Mangasariam, O.: UCI machine learning repository. breast cancer wisconsin (diagnostic) data set (2007)
20. Zhang, S., Qin, Z., Ling, C.X., Sheng, S.: "missing is useful": Missing values in cost-sensitive decision trees. IEEE Transactions on Knowledge and Data Engineering, 17(12), 1689–1693 (2005)