Searching Good Similarity Functions Using Genetic Programming for a Case-Based Reasoning Classifier System

Joan Camps and Josep Maria Garrell and Elisabet Golobardes
and David Vernet
{joanc,josepmg,elisabet,dave}@salleURL.edu

Enginyeria i Arquitectura La Salle, Universitat Ramon Llull,
Passeig Bonanova 8, 08022 - Barcelona, Spain

Abstract. This paper presents some results obtained with the hybridization of Case-Based Reasoning (CBR) with Genetic Programming (GP). CBR is used in a classification environment so it tries to classify new cases using a set of historical retained cases. One key point is the similarity function used to retrieve similar cases from the case base. There are some functions to calculate this distance or similarity between cases. The aim of this paper is to automatically find a good similarity function for the specific domain in which the classifications are made. To reach this goal a GP approach is used. The test bed used comes from a real life problem, the diagnosis of breast cancer using data collected from digital mammography.

1 Introduction

Case-Based Reasoning (CBR), one of the methodologies described in analogical machine learning, is a good method to predict or classify items. To be able to do this prediction, it is indispensable to use an objective similarity measure to compare with other situations retained in the memory of the system. This measurement is called *similarity function* and it is used in many areas but in the analogical reasoning, especially [13, 2, 14]. Depending on the problem to classify, some standard similarity functions could be better or worse. Usually, to improve the classification rate (accuracy in most systems), the similarity function needs to be adapted for each problem.

In this paper, an automatic method to obtain a similarity function is presented. The approach presented is a hybridization between Case-Based Reasoning (CBR) and Genetic Programming (GP) [8, 9]. In some way, the GP searches a good similarity function, the best if possible, to be used in the CBR system in a specific domain. The specific similarity functions obtained by employing GP are evaluated using the classification rate with the use of the CBR.

To test this method, a specific medical problem is used, the prediction of breast cancer. Some features are extracted from several mammographs which is the input of our classification system. Every one of this set of features is a case to be used in the CBR.

After a basic explanation of the CBR fundamentals, the design of the hybrid system is described with the detailed variants in the fitness calculation and the complete configuration used. This section is followed by the experimentation design and the results obtained in the runs set. In the last section some conclusions are collected and a possible extension is proposed.

2 Case-Based Reasoning Fundamentals

Case-Based Reasoning (CBR) is one of the methodologies of analogical machine learning. The system is inspired by the analogy used by the humans when they need to solve some new situation. They find the situation in their memory, trying to find the most similar stored situation. In the same way, this methodology uses a case base that is a set of cases that hopefully represents all the possible ones. When a new case comes up, the CBR finds the most similar case in its memory classifying the new one.

The basic CBR cycle [1] is formed by 4 phases (see Figure 1) :

1. *Retrieval Phase*: In this process the system looks for the most similar cases from the set of known cases.
2. *Reuse Phase*: The selected cases in the last phase are adapted trying to make them close to the case to be classified.
3. *Revise Phase*: The correctness and viability of the proposed solution is revised.
4. *Retain Phase*: At the end, the new relevant information obtained is stored.

The CBR can be used for many different kinds of problems. In this work it is used as a classifier. In the case base, each one is related to a determined class. The process goal is to predict the class where the new case belongs. Not all the phases of the original CBR are used in this hybrid when a new similarity function is found. Only the Retrieval Phase is used because there is not need any kind of adaptation of the used cases (Reuse Phase) neither any revision of the proposed solution (Revise Phase) due the case structure is a constant set of attributes with a known domain for each one. In the same way, the Retain Phase is not used in this work because the goal is to work always with the same knowledge in all the process. The CBR configuration used in the hybrid is centered only in the Retrieval Phase. In a further work the aim is to use all the possibilities of the CBR but now is working just like a nearest neighbor algorithm.

In the Retrieval Phase, to recover the closest cases from the case base is necessary the use of a similarity function that permits to obtain an objective value as a measure of the distance between two cases. There are many possible similarity functions. Some of these functions are: the metric of Minkowsky, Mahalanobis, Camberra, Chebychev and Quadratic Correlation. The most typical used formula is the metric of Minkowsky [4],

$$Mink(Case\_x, Case\_y, r) = \sqrt[r]{\sum_{i=1}^{F} w_i \times |x_i - y_i|^r} \quad (1)$$

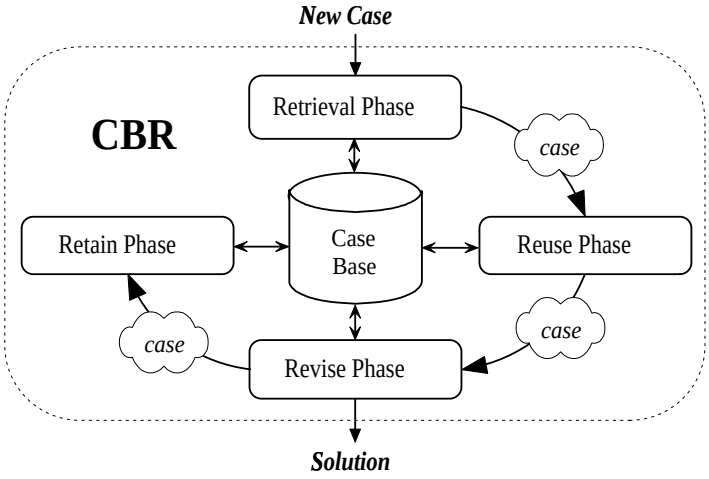

Fig. 1. Case-Based Reasoning Cycle.

Case_x and *Cas_y* are two cases to be compared; F is the total number of attributes or features that every case has; x_i and y_i represent the value of the *i-th* attribute for the *Case_x* and *Case_y*, respectively; w_i is the weight of the *i-th* attribute; and r is the applied degree to the formula. The most known release of the Minkowsky's formula is the Euclidian distance:

$$Eucl(Case\_x, Case\_y) = \sqrt[2]{\sum_{i=1}^{F} |x_i - y_i|^2} \tag{2}$$

As is possible to be observed, the degree is $r = 2$ and all the weights are 1. But another simplified and fast function is also possible, the Hamming distance, simpler in the calculation:

$$Hamm(Case\_x, Case\_y) = \sum_{i=1}^{F} |x_i - y_i| \tag{3}$$

Despite these possibilities, in this work any of the previous functions are used. The learning of the hybrid system just will be to find the best similarity function or, at least, someone better than the classical ones.

3 Hybridazing CBR with GP

The final aim of this work is to improve a rate of classification or prediction in a specific domain. To achieve this results, a hybrid system has been developed using Genetic Programming and Case-Based Reasoning. Usually, CBR, in

order to classify, uses someone of the classic similarity functions described before. The goal of the hybrid system is to find a new similarity function, the best if is possible, for the specific domain where the problem is defined. There are some previous works which related the Case-Based Reasoning with the Evolutionary Computation [3, 5]. In this way, some works use Genetic Algorithms as a extractor of the most relevant features from the available attribute set [7, 12].

As a new proposal, in this paper a hybrid system between CBR and GP is described. The hybridization can be considered from two points of view: from the CBR and from de GP. In the former, the CBR system uses a external similarity function to classify a set of test cases. In this process, a percentage of right classifications is obtained. From this point of view, the GP is a system that proportionate the external similarity function. In the later point of view, the GP is evolving a set of functions needing a system to obtain their level of goodness, otherwise, the fitness of each individual. In this point of view, the CBR is the system that gives this value to the GP system.

In the training process the GP starts evaluating a random generate function set. Each individual of the GP is a possible similarity function and it is represented by a tree made up by a set of operators, parameters and some constants. The process is the classical cycle of Genetic Programming: individual evaluation, selection of some ones, replication and crossing between them, and other secondary operators that can be applied like the mutation. The difference in this process is into the evaluation phase; the fitness of every issue is obtained from de CBR cycle. In other words, for every individual of the population, a CBR cycle is executed to obtain the value of its goodness (see Figure 2).

When the evolved similarity function is obtained, this one can be used for the classical CBR cycle to predict the new cases in this domain.

The similarity function must to be able to return a real value meaning the distance between two cases, one from the case base and the other to be classified. This function can use both attributes set, from both cases. Consequently, the terminal set of the GP has, at least, the double of the parameters that the number of attributes of each case.

The calculation of the fitness for every individual of the GP is a important decision to take. The fitness is used to select some of these individuals with more or less probability to contribute in the next generation of the GP. In some preliminary tests a row fitness was used as a measurement of every individual obtaining good results but not better than the obtained ones with the classical CBR system using the Euclidian similarity function [6]. This first fitness calculation is a global percentage of right classifications (FC1).

$$FC1 = \frac{CorrectClassifiedCases}{NumberOfCases} \tag{4}$$

To improve this situation, other kind of fitness calculations have been designed. If the situation is an unbalanced test set, the obtained function could be unbalanced too. The aim is to balance the correct result rate using the accuracy, a proportionate calculation between all the possible classes. This goal can be

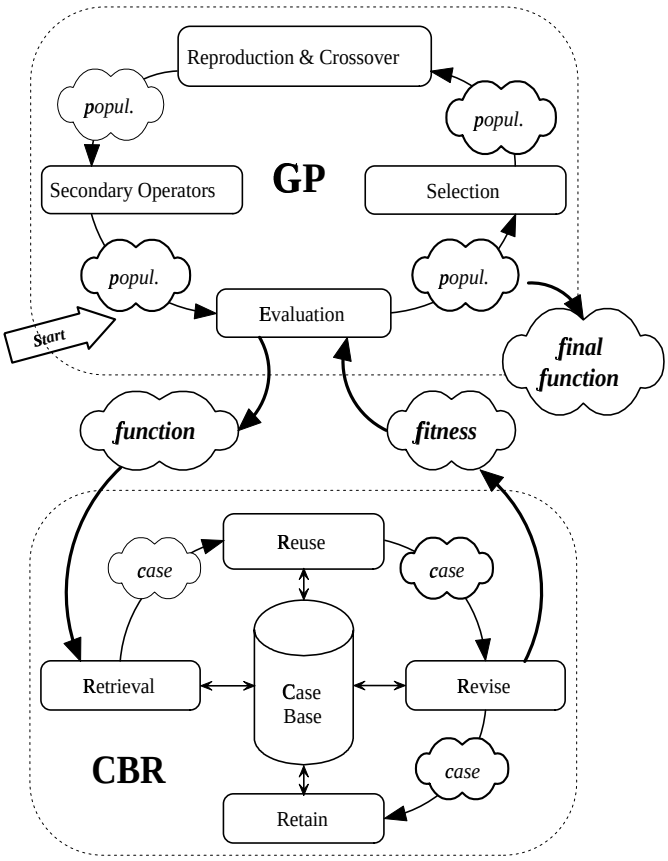

Fig. 2. Basic Hybrid Cycle.

considered like a multi-objective optimization technique. Two approximations have been designed: a proportionate addition of each possible classes, and a multiplication of the correct result rate for every class. While the former, named FC2 here, is indirectly balancing the test set, the later, named FC3, improve the maximization of all correct class result rate at the same time, allowing to prior the most balanced obtained functions.

$$FC2 = \frac{\sum_{i=1}^{Classes} \frac{CorrectClassifiedCases_i}{NumberOfCases_i}}{Classes} \tag{5}$$

$$FC3 = \prod_{i=1}^{Classes} \frac{CorrectClassifiedCases_i}{NumberOfCases_i} \tag{6}$$

Also, a detailed GP configuration is needed: adjusting the evolutionary parameters, the terminal set and the operator set. A elitist generational process

with a roulette-wheel selection is used for all tests. The crossover probability is 0.5 and the mutation probability 0.5. For each run 100 iterations of the GP cycle has been made. And there are 100 individuals per run with a maximum depth of 3 and a ramped half-and-half initial population.

The terminal set contains the ephemeral random constant and twice the number of parameters that the number of the attributes of each case. In the prediction of breast cancer, each case is made up of 21 attributes. The found function will calculate a distance between two cases. Consequently, in this problem, a set of 42 parameters is needed. The ephemeral random constant have a real domain between -10 and 10, and a constant perturbation probability of 0.75.

The operator set is made up of 10 atomic functions: *addition*, *subtraction*, *multiplication*, *division*, *squareroot*, *cuberoot*, *square*, *cube*, *absolutevalue* and $sum21$, a special operator that corresponds with the next expression:

$$sum21 = \sum_{i=1}^{21} subexp_i \times |x_i - y_i| \tag{7}$$

Where $subexp_i$ is a sub-tree which works like a weight to the i-th attribute. X_i and Y_i are the i-th attributes of the cases that have been comparing.

4 Experimentation

In this section the classification problem used to test the system is explained briefly. However, the training running set using a expert selected training set of cases is explained too.

The problem used to test the hybrid system is the prediction of breast cancer. The aim is to be able to predict if a person has or not a breast cancer. The obtained data to get this goal is a set of mammographs [11]. There are many techniques to predict breast cancer but, to make a fast prediction, the most used technique for the experts is the analysis of the microcalcifications contained in the mammography. The digitalized mammography is segmented to show better the microcalcifications and to obtain a set of objective data to be used in the train and the test system. Concretely, in this problem a set of 21 features is obtained for every mammography. A set of 286 processed mammogramphs is the total number of cases used to test and train the hybrid system. Every case has its correct diagnosis in one of the Positive or Negative class.

In a preliminary work [5, 6], using the hybrid with the FC1 process to obtain the fitness in every evaluated formula, some results were obtained but not better than the classical prediction obtained using the CBR system with its classical similarity functions [10]. In these different runs only the number of individuals per population and the number of iterations were changed.

The training and testing sets have been designed by some experts. As a result of this process, a training set with 216 cases and a test set with 70 cases were obtained. In the training set there is a 34.26% of cases with breast cancer. In the test set this percentage is about 30%.

The maximization of the correct classifications is the aim of the hybrid system but this is not only the goal. In medical environments a specific language and some conditions are required to trust in a prediction. The sensitivity is defined like the proportion of true positive predictions and the specificity is the percentage of true negative predictions. In other words, the sensitivity describes the capacity of a system to predict when a mammography reflects a breast cancer, and the specificity describes the capacity of a system to ensure that there is not a breast cancer. In terms of classification, the sensitivity is the rate of the positive class and the specificity is the rate of the negative class. In medical environments is required not only the maximization of the global accuracy but also the maximization of this both terms just described.

Different kinds of fitness calculation have been designed to try to improve all these rates. While the FC1 fitness calculation is only the rate of correct classifications, the FC2 and FC3 try to improve the sensitivity and the specificity at the same time.

5 Results

In this section all the results are analyzed and summarized. For each kind of fitness calculation system a set of 10 runs has been throw. The global classification rate average is compared with the obtained accuracy using classical similarity functions in CBR. The partial average for every proposed fitness calculation is also showed making a comparison between these differences.

Table 1 shows the results of the classification rate using CBR. The similarity functions used in these first classifications are Hamming, Euclidean, Cubic and Clark. There are three values for each similarity function: sensitivity, specificity and accuracy. The global average of all the results using the hybrid system are shown in the bottom of the table. As can be appreciated, the average of all the

Runs	Sensitivity	Specificity	Accuracy
`Hamming`	80.95	69.39	72.86
`Euclidian`	66.67	75.51	72.86
`Cubic`	66.67	77.55	74.29
`Clark`	66.66	81.63	77.14
`Hybrid`	78.64	73.17	77.00

Table 1. Comparing the results of the classical similarity functions used in CBR with the average of all the results obtained with the GP-CBR hybrid system.

runs in the hybrid between CBR and GP is in the level of the best execution using the Clark similarity function in CBR. Despite these results, the average of sensitivity and specificity are more balanced using the similarity functions obtained in the hybrid system.

The best found solutions are from the FC3 and FC1 run set, while the worst solutions are from the FC2 and FC1 run set.

The best global solution in accuracy is the same that appears with the best specificity rate. Has a 84.29% of accuracy and a specificity of 90.48%, appearing twice. Otherwise, the best sensitivity rate is a function from the FC1 runs set that obtain a 100%, but is the worst in specificity, only a 19.05%. The worst solution in accuracy and sensitivity is a FC2 function, with a 70.0% of accuracy and a 65.31% in sensitivity.

Table 2 shows a comparison between the averages obtained using the proposed different kinds of fitness calculations. Like the table before, the sensitivity (Sensit.), the specificity (Specif.) and the accuracy (Accur.) are shown in three columns but, in this case, with the standard deviation of each one of these measurements. As can be seen in the Accuracy column, the worst average is the

Runs	**Sensit. (Std.Dev.)**		**Specif. (Std.Dev.)**		**Accur. (Std.Dev.)**	
FC1	86.53	7.15	58.10	21.27	78.00	2.35
FC2	72.86	4.82	79.05	6.81	74.71	2.34
FC3	76.53	3.23	82.38	5.96	78.29	3.86

Table 2. Comparing the results of the different systems of fitness calculation.

obtained with the FC2 fitness calculation. The other calculation systems have a very similar rate in this column. It is interesting to pay attention in the comparison between the sensitivity and specificity in FC1 and FC3 fitness calculations and, especially, in their standard deviations.

In the FC1 fitness calculation system, the best obtained similarity function has 82.86% of accuracy and a 80.95% of specificity. Otherwise, the best in sensitivity is the worst in the other rates.

The FC2 fitness calculation seems the worst. The best accuracy is 77.14% and appears three times, the best in sensitivity is 81.63% and the best in specificity is 85.71%, appearing four times. The worst has a accuracy of only 70.0% and a sensitivity of 65.30%. Another function is the worst in specificity, with a 71.42% rate.

The FC3 fitness calculation system is the best of all. The best found solution (Equation 8) is the best in the three measurements: 84.29% of accuracy, 81.63% of sensitivity and 90.48% of specificity, appearing twice.

$$
\begin{aligned}
Sim\ (Case\_X, Case\_Y) = {} & Y_{12}(X_1 - Y_1) + X_{21}(X_2 - Y_2) + {X_9}^2(X_3 - Y_3) + \\
& 7.26566(X_4 - Y_4) + X_7(X_5 - Y_5) + X_{17}(X_6 - Y_6) + 10.5642(X_7 - Y_7) + \\
& X_{17}(X_8 - Y_8) + {X_9}^2(X_9 - Y_9) + 8.91974(X_{10} - Y_{10}) - 3.58618(X_{11} - Y_{11}) \\
& -5.77795(X_{12} - Y_{12}) + X_9(X_{13} - Y_{13}) - 4.91022(X_{14} - Y_{14}) + \\
& Y_{19}(X_{15} - Y_{15}) + Y_7(X_{16} - Y_{16}) - 7.81784(X_{17} - Y_{17}) + X_{15}(X_{18} - Y_{18}) \\
& -3.58618(X_{19} - Y_{19}) - 5.46136X_{18}(X_{20} - Y_{20}) + X_9(X_{21} - Y_{21})
\end{aligned}
\tag{8}
$$

The worst found solution is the worst in the three measurements too: 74.29% of accuracy, 73.47% of sensitivity and 76.19% of specificity, appearing twice.

6 Conclusions and Further Work

The similarity function selection for a CBR classifier is a key point in order to reach good prediction results. In this work, this selection was performed using Genetic Programming. The different similarity functions proposed by the GP have been analyzed with the CBR to obtain a correctness rate, the fitness. The results obtained with the different runs show the proposed hybrid system as a good improvement to find a specific similarity function for a CBR classification system in a certain domain. This resultant improvement supposes an increment in the computational cost. As usually, better similarity functions are obtained but this needs greater search time.

In these kinds of predictions, like breast cancer, it is very important to maximize the sensitivity and the specificity. In a failed prediction it is better to predict a False Positive than a False Negative. The aim is to improve not only the accuracy, but also the sensitivity and specificity. With this goal, different modes of fitness calculation have been implemented and tested. The FC3 fitness calculation formula appears as the best. In this modality, the accuracy obtained is the best and the sensitivity and specificity also show improvement.

It is interesting to note that in the breast cancer prediction problem, the results obtained by the hybrid system are better that the others obtained using CBR with standard similarity functions. Actually, a specific similarity function is found for this problem.

The results obtained with this system could be better applying a better method to adjust the random constants than the perturbation constant. In a further work we will try with different methods to perform this improvement. Other fitness calculation formulae can be tested in the same line than the FC3 tested here. The design of this formulae can be developed with the perspective of the multi-objective optimization, with the aim to improve both, sensitivity and specificity, but with the possibility to give more priority to one or another.

Acknowledgements

We would like to thank the *Ministerio de Sanidad y Consumo, Instituto de Salud Carlos III, Fondo de Investigación Sanitaria* for its support under grant number FIS 00/0033. The results of this work were obtained using the equipment co-funded by the *Direcció General de Recerca de la Generalitat de Catalunya (D.O.G.C 30/12/1997).* And finally, we would like to thank Enginyeria i Arquitectura La Salle for their support to our AI Research Group.

References

1. Aamodt, A., Plaza, E.: Case-Based Reasoning: Foundations Issues, Methodological Variations, and System Approaches. AI Communications **7** (1994) 39–59
2. Aha, D., Harrison, P.: Case-Based Sonogram Classification. Navy Center for Applied Research in AI, Washington, D.C. AIC-93-041, **NRL/FR/5510-94-9707** (1994)
3. Ahluwalia, M., Bull, L.: Coevolving Functions in Genetic Programming: Classification using K-nearest-neighbour. Proceedings of the Genetic and Evolutionary Computation Conference (1999)
4. Bachelor, B.: Pattern Recognition: Ideas in practice. Plenum Press (1978)
5. Golobardes, E., Llorà, X., Salamó, M., Martí, J.: Computer Aided Diagnosis with Case-Based Reasoning and Genetic Algorithms. Journal of Knowlegde-Based Systems **15**, Elsevier Science (2002) 45–52
6. Golobardes, E., Nieto, M., Salamó, M., Camps, J., Calzada, G., Martí, J., Vernet, D., Generació de funcions de similitud mitjançant la Programació Genètica pel Raonament Basat en Casos. Butlletí de l'ACIA: 4t Congrés Català d'Intel·ligència Artificial **25** (2001) 100–107
7. Kelly, J.D., Davis, L.: Hybridizing the Genetic Algorithms and the K Nearest Neighbors. Proceedings of the Fourth International Conference on Genetic Algorithms (1991)
8. Koza, J.R.: Genetic Programming. On the programming of computers by means of natural selection. Massachusetts Institute of Technology Press (1992)
9. Koza, J.R.: Genetic Programming II. Automatic Discovery of Reusable Programs. Massachusetts Institute of Technology Press (1994)
10. Martí, J., Español, J., Golobardes, E., Freixenet, J., García, R., Salamó, M.: Classification of microcalcifications in digital mammograms using case-based reasonig. Proceedings of the 5th International Workshop on Digital Mammography, Medical Physics Publishing (2000)285–294
11. Martí, J., Cufí, X., Regincós, J., et al.: Shape-based feature selection for microcalcification evaluation. Imaging Conference on Image Processing **3338** (1998) 1215–1224
12. Raymer, M.L., Punch, W.F., Goodman, E.D., Kuhn, L.A.: Genetic Programming for Improved Data Mining: An Application to the Biochemistry of Protein Interactions. Genetic Programming 1996: Proceedings of the First Annual Conference (1996) 375–380
13. Riesbeck, C.K., Schank, R.C.: Inside Case-Based Reasoning. Lawrence Erlbaum Associates (1989)
14. Salzberg, S.: A nearest hyperrectangle learning method. Machine Learning **6** (1991) 277–309

This article was processed using the LaTeX macro package with LLNCS style