Weighting methods for a Case-Based Classifier System

[a]Maria Salamó [a]Elisabet Golobardes [a]David Vernet [a]Mireya Nieto

[a]Computer Science Department
Enginyeria i Arquitectura La Salle, Ramon Llull University
Passeig Bonanova, 8, 08022 Barcelona, Catalonia, Spain
{mariasal, elisabet, dave, mireyan}@salleURL.edu
Telephone: +34 932 902 400
Fax: +34 932 902 416

Summary

Objective

This paper describes two methods for weighting the feature relevance in a Case-Based Reasoning system. The first weighting method used inside the Case-Based Reasoning is based on Rough Sets theory. The second weighting method studied is based on Sample Correlation. Experiments in a large number of domains from the UCI repository show that these weighting methods improve accuracy rate.

Case-Based Classifier Systems

Case-Based Reasoning (CBR) integrates in one system two different characteristics: machine learning capabilities and problem solving capabilities. CBR uses a similar philosophy to that which humans sometimes use: it tries to solve new cases (examples) of a problem by using old previously solved cases [Riesbeck:1990]. The process of solving new cases contributes with new information and new knowledge to the system. This new information can be used for solving other future cases. The basic method can be easily described in terms of its four phases [Aamodt:1994]. The first phase *retrieves* old solved cases similar to the new one. In the second phase, the system tries to *reuse* the solutions of the previously retrieved cases for solving the new case. The third phase *revises* the proposed solution. Finally, the fourth phase *retains* the useful information obtained when solving the new case. In a Case-Based Classifier System, it is possible to simplify the reuse phase. Classifying the new case with the same class as the most similar retrieved case can do reuse.

The kernel in a Case-Based Reasoning system is the *retrieval* phase (phase 1). Phase 1 retrieves the *most similar* case or cases to the new case. Obviously, the meaning of *most similar* will be a key concept in the whole system. Similarity between two cases is computed using different similarity functions.

For our purpose in this paper, we use the similarity functions based on the distance concept. The most used similarity function is the *Nearest Neighbour* algorithm, which computes the similarity between two cases using a global similarity measure [Aha:1998]. The practical implementation (used in our system) of this function is based on the *Minkowski's metric*.
Minkowski's metric is defined as:

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

Where *Case_x, Case_y* are two cases, whose similarity is computed; *F* is the number of features that describes the case; x_i, y_i represent the value of the *ith* feature of case *Case_x* and *Case_y* respectively; and w_i is the weight of the *ith* feature. In this study we test the Minkowsky's metric for three different values of r: *Hamming distance* (*r = 1*), *Euclidean distance* (*r = 2*), and *Cubic distance* (*r = 3*). This similarity function needs to compute the feature relevance *(w_i)* for each problem to be solved. Assuming an accurate weight setting, a case-based reasoning system can increase their prediction accuracy rate. We use also the Clark's and the Cosine distance [Llora:2000]; both are based on distance concept and also use weighting features.

Sometimes human experts can not adjust the feature relevance, automatic method can solve this limitation. Literature describes many methods for feature weighting [Aha:1998].

This paper presents a weighting method based on the Rough Sets theory introduced by Pawlak [Pawlak:1982, Pawlak:1991]. It is a single weighting method (RSWeight) that computes the feature weights from the initial set of train cases in the CBR system. We also introduce a weighting method that computes the Sample Correlation among the features and the classes that the cases may belong to. Next sections will describe briefly the main ideas of rough sets and Sample Correlation weighting methods.

Feature Selection based on Rough Sets theory

Zdzislaw Pawlak introduced Rough Sets theory in 1982 [Pawlak:1982, Pawlak:1991, Skowron:1995]. The idea of the rough set consists of the approximation of a set by a pair of sets, called the lower and the upper approximation of this set. In fact, these approximations are inner and closure operations in a certain topology generated by the available data about elements of the set.

The main research trends in Rough Sets theory which try to extends the capabilities of reasoning systems are: (1) the treatment of incomplete knowledge; (2) the management of inconsistent pieces of information; (3) the manipulation of various levels of representation, moving from refined universes of discourse to coarser ones and conversely.

The nature of Rough Sets theory made them useful for reducing the knowledge, extracting dependencies in knowledge, reasoning about knowledge, pattern recognition, etc.

We use Rough Sets theory for reducing and extracting the dependencies in the knowledge. These dependencies are the basis for computing the relevance of each attribute inside the Case-Based Reasoning system. How Rough Sets theory is incorporated into our Case-Based Classifier system?

First of all, we incorporate some briefly explanations in this article to explain how the dependencies we are looking forward are obtained. The mathematical formulas have been omitted in this extended abstract.

We compute from our *universe* (finite set of objects that describe our problem, *the case memory*) the *concepts* (objects or *cases*) that form partitions of that Universe. The union of all the concepts made the entire Universe. Using all the concepts we can describe all the *equivalence relations* (R) over the universe. Let an *equivalence relation* be *a set of features* that describe a specific concept. The universe and the relations form the knowledge base, defined as KB = (U; R). Every relation over the universe is an *elementary concept* in the knowledge base.

All the concepts are formed by a set of equivalence relations that describe them. So we search for the minimum set of equivalence relations that define the same concept as the initial set. The set of minimum equivalence relations is called *reduct*. A reduct is the essential part, which suffices to define the basic concepts occurring in the knowledge. The *core* is the set of all indispensable equivalence relations over the universe, in a certain sense the most important part of the knowledge. The core is defined as the intersection of all the reducts.

Reducts contain the dependencies from the knowledge. We can use this information to weigh the relevance of each feature in the system. An attribute that does not appear in the reduct has a feature weight value of 0.0, whereas an attribute that appears in the core has a feature weight value of 1.0. The rest has a feature weight value depending on the proportional appearance in the reducts. This is the weight feature information that we use in the case-based classifier system.

Sample Correlation

Sample Correlation computes the weights w_i computing the sample correlation which exists between each feature x_i and the class z (see equation [Sample_Correlation]).

The Sample Correlation is defined as:

$$Sample\_Correlation(x_i, z) = \frac{1}{N-1} \sum_{j=1}^{N} \left(\frac{x_{ij} - \bar{x}_i}{S_{x_i}} \right) \left(\frac{z_j - \bar{z}}{S_z} \right)$$

Where N is the number of cases; x_{ij} is the value of the *ith* feature for the case j; z_j is the class which belong to the case j; ξ_i is the mean of the *ith* feature; | is the mean of the classes; S_{xi} is the standard deviation of the feature x_i; and S_z is the standard deviation of the class z.

Experiment Study: UCI datasets

The experiment conducted use 4 data sets from the UCI repository, echocardiogram, iris, breast cancer Wisconsin, water-treatment and one data set from our own repository (mammogram problem) [see Table 1]. The mammogram problem consists of detecting breast cancer using the information found in a mammography [Llora:2000, Marti:2000, Salamo:1999]. A microcalcification (μCa) usually appears, in the mammographies, as small, bright, arbitrarily shaped regions on the large variety of breast texture background. Thus their analysis and characterisation are performed throughout the extraction of features and visibility descriptors by means of several image processing techniques [Shen:1994]. Each example contains the description of several (μCa) present in the image. For each of these microcalcifications there are 23 real valued features. In other words, the input information used is a set of *m x 23* real valued matrixes, where *m*[1] is the number of μCa present on the image. The data set contains 216 examples.

Domain	Examples	Features	Classes	Missing Values	Inconsistent
Echocardiogram	132	9	2	132	Yes
Iris	150	4	3	0	No
Breast cancer (Wisconsin)	699	9	2	9	Yes
Water-treatment	527	38	13	591	Yes
[2]Mammogram problem	216	23	2	0	Yes

Table 1 Data sets used in the experiment.

The examples of each data set have been grouped in two sets: the training set and the test set. We use the first set to train the system, and the second set to test the system. The training set and the test set are generated using different proportions of the examples: 10% of the examples for the training set and the rest (90%) for the test set, 20% of the examples for the training set and the rest (80%) for the test set, …, until 90% for the training set and 10% for the test set.

[1] We want to remark that can be different for each mammogram.
[2] Dataset from our own repository.

For each proportion 10 random divisions are generated, in such a way that several test sets (and training sets) for one proportion are different. So, the testbed contains 10 versions for each proportion. It means a total of 90 experiments for each version of the CBR system.

Preliminary Results

We present in Table 2 the results obtained during the execution of the proportion 90% training set and 10% test set. This proportion has been chosen for the accurate rate obtained, we want to notice that the results presented are the maximum value obtained during the executions.

Domain	Max Not weight	Max Weight RS	Max Weight Correlation
Echocardiogram	78.57 %	78.57 %	85.71 %
Iris	100 %	100 %	100 %
Breast cancer (Wisconsin)	98.71 %	100 %	98.71 %
Water treatment	77.35 %	79.20 %	79.20 %
Mammogram problem	77.27 %	81.81 %	81.81 %

Table 2 Preliminary results.

The results presented obtain a good accuracy rate. It is very important to mention that some maximum values obtained without weighting features appear few times in the different configurations made to the system, whereas the feature weighting methods maintain the maximum values obtained.

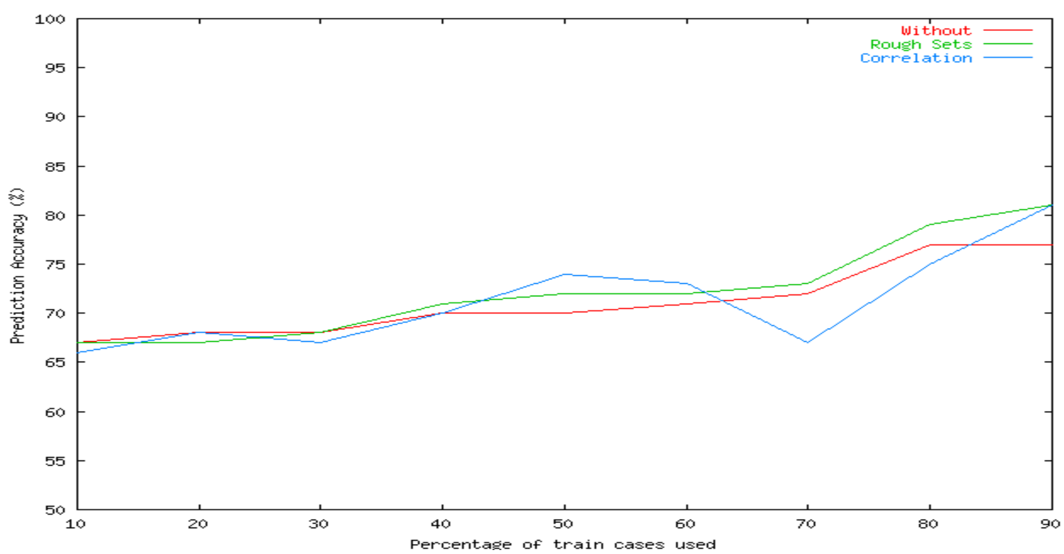

Figure 1. Preliminary results for all the proportions.

Figure 1 shows the results obtained for all the proportions in the mammogram problem. As it can be seen the weighting feature methods needs a huge amount of cases to develop a good weighting feature selection during the retrieval phase. However, the system accuracy rate increases when there are enough information in the system to develop a good weighting policy. Also, the system weighted computes a minus deviation as the system without weighting values. We can notice also that it is very important to select the correct case memory policy to achieve better results. Most of the best results obtained have been achieved with a previous train in the training set. This training set has been decreased following this policy, so the cases chosen were the best fits to explain the problem. The system weighted on average, including means values, outperform the prediction accuracy and the deviation obtained has been reduced.

Main References

[Aamodt:1994] *Case-Based Reasoning: Foundations issues, methodological variations, and systems approaches,* A. Aamodt and E. Plaza, AI Communications, Vol. 7. 39-59,1994

[Aha:1998] Feature weighting for lazy learning algorithms, David W. Aha, Technical Report AIC-98-003. Washington, DC: Naval Research, Laboratory, Navy Center for Applied Research in Artificial Intelligence, 1998

[Llora:2000] Diagnosis of Microcalcifications using Case-Based Reasoning and Genetic Algorithms, X. Llorà, E. Golobardes, M. Salamó, J. Martí, Proceedings Symposium of Engineering of Intelligent Systems (EIS2000). To Appear

[Marti:2000] *Classification of Microcalcifications in digital mammograms using Case-Based Reasoning,* J. Martí, J. Español, E. Golobardes, J. Freixenet, R. García, M. Salamó, International Workshop on Digital Mammography (IWDM2000). To Appear

[Pawlak:1982] *Rough Sets,* Zdzislaw Pawlak, International Journal of Information and Computer Science, Vol. 11, 1982

[Pawlak:1991] *Rough Sets: Theoretical Aspects of Reasoning about Data,* Zdzislaw Pawlak, Kluwer Academic Publishers, Vol. 9 of Series D: System Theory, Knowledge Engineering and Problem Solving, 1991

[Riesbeck:1989] *Inside Case-Based Reasoning,* C.K. Riesbeck and R.C. Shank, Lawrence Erlbaum Associates, Hilldale, NJ, USA, 1989

[Salamo:1999] *Clasificación de microcalcificaciones usando razonamiento basado en casos,* M. Salamó, E. Golobardes, J.M. Garrell, J. Martí, III Jornadas de Transferencia Tecnológica de Inteligencia Artificial (TTIA'99) Murcia, Spain, 1999

[Shen:1994] *Detection and classification of mammographic calcifications*, Shen, L., R.M. Rangayyan, and J.L. Dessautels, In *State of the Art in Digital Mammographic Image Analysis*, K. Bowyer and S. Astley, eds., pp. 198-212, World Scientific Publishing, 1994