Proceedings of the International Conference
on Computational and Mathematical Methods
in Science and Engineering, CMMSE2009
Gijón, Spain, June 30, July 1-3 2009

A summarization of SOM ensembles algorithm to boost the performance of a forecasting CBR system applied to forest fires.

Emilio Corchado[1], Aitor Mata[2] and Bruno Baruque[2]

[1] *Department of Civil Engineering, University of Burgos, Spain*
[2]*Department of Computing Science and Automatic, University of Salamanca, Spain*

emails: escorchado@ubu.es, aitor@usal.es, bbaruque@ubu.es

Abstract

A new forecasting system is presented in this paper. The Case-Based Reasoning methodology combined with a summarization of SOM ensembles algorithm has been used to face this problem. CBR represents a knowledge extraction methodology, where past information is used to generate new solutions to new problems. The new summarization algorithm (WeVoS-SOM) organizes the stored information to make it easier to retrieve the most useful information from the case base. The developed system has been checked with forest fires data. Forest fires represent an environmental risk that should be predicted in order to avoid further damages. The WeVoS-CBR system has been able to predict the future situation of geographic areas after a forest fire has been originated.

Key words: Case-Based Reasoning, forest fires, Self Organizeing Memory, summarization.

1. Introduction

Case-Based Reasoning systems have the potential to use past information in order to generate useful knowledge that may be used to solve new problems. Those systems should organise the information handled in order to improve the way that information is used. When the amount of information stored in a CBR system grows, the results obtained are normally better. But the growth of the *case base* (internal structure where the data is accumulated) also implies a more difficult retrieval process, where more information has to be considered in order to obtain the best possible collection of data.

The summarization algorithm presented here, WeVoS-SOM (Weighted Voting Summarization of SOM ensembles) represents the organizing system of the internal structure of the data in the CBR system. With such an inner organization, it is easier to locate the new data that is introduced in the system and to retrieve the needed information to solve new problems.

The combination of both, the generalization power of the CBR methodology and the organizational capabilities of the WeVoS-SOM algorithm generates a system that has been used to generate predictions in a natural environment such as forest fires. Historical data have been used to check the correction of the system, where the predictions generated were compared with the actual past data.

In this paper, and after this introduction, the Case-Based Reasoning methodology is briefly explained. Then the summarization algorithm is presented. Next, the developed system is described. Finally, some results of the application of the system to the forest fire case study are shown.

2. Case-Based Reasoning: the core methodology

Case-Based Reasoning [1] origins are in knowledge based systems. CBR systems solve new problems acquiring the needed knowledge from previous situations [2]. The main element of a CBR system is the case base, a structure that stores the information used to generate new solutions.

The learning capabilities of CBR systems are due to its own structure, composed of four main stages [3]: *retrieve*, *reuse*, *revision* and *retain*. These stages are depicted in Fig. 1. The first stage is called *retrieve*, and consists in finding the cases (from the case base) that are most similar to the new problem. Once a set of cases is extracted from the case base, they are *reused* by the system. In this second stage (*retain*), the selected cases are adapted to fit in the new problem. After applying the new solution to the problem, that solution is *revised* to check its performance. If it is an acceptable solution, then it is *retained* by the system and could eventually serve as a solution to future problems.

As a methodology [1], CBR has been used to solve a great variety of problems. It is a cognitive structure that can be easily applied to solve problems such as those related with soft computing, since the procedures used by CBR are quite easy to assimilate by soft computing approaches. CBR has also helped to create applications related to quite different environments. Different kinds of neural networks such as ART-Kohonen [4] or Growing Cell Structures [5] have been combined with CBR to automatically create the inner structure of the case base. Some effort has also been devoted to the case-based maintenance issue [6].

It is easy to understand that the *case base* is one of the key elements of a CBR system. It is crucial to dispose of a great amount of data, but properly organized.

The quantity is important, but if there is no order within the stored cases, it could become impossible to obtain all the knowledge that such an accumulation of data may offer. This is why an algorithm like the WeVoS-SOM that is going to be explained next, is so useful to CBR systems.

3. WeVoS: Weighted Voting Summarization of SOM ensembles

Case-Based Reasoning systems are highly dependent on stored information. The new algorithm presented here, *Weighted Voting Summarization of Som ensembles (WeVoS-SOM)* is used to organize the data that is accumulated in the case base. It is also used to recover the most similar cases to the proposed problem.

The main objective of the new fusion of an ensemble of topology preserving maps [7] algorithm presented here, WeVoS-SOM, is to generate a final map processed unit by unit. Instead of trying to obtain the best position for the units of a single map trained over a single dataset, it aims to generate several maps over different parts of the dataset. Then, it obtains a final summarized map by calculating by consensus which is the best set of characteristics vector for each unit position in the map. To do this calculation, first this meta-algorithm must obtain the "quality" [8] of every unit that composes each map, so that it can relay in some kind of informed resolution for the fusion of neurons.

The final map obtained is generated unit by unit. The units of the final map are first initialized by determining their centroids in the same position of the map grid in each of the trained maps. Afterwards, the final position of that unit is recalculated using data related with the unit in that same position in every of the maps of the ensemble. For each unit, a sort of voting process is carried out as shown in Eq. 1:

$$V(p,m) = \frac{|x_{p,m}|}{\sum_1^M |x_p|} \cdot \frac{q_{p,m}}{\sum_1^M q_p} \qquad (1)$$

The final map is fed with the weights of the units as it is done with data inputs during the training phase of a SOM, considering the "homologous" unit in the final map as the BMU. The weights of the final unit will be updated towards the weights of the composing unit. The difference of the updating performed for each "homologous" unit in the composing maps depends on the quality measure calculated for each unit. The higher quality (or the lowest error) of the unit of the composing map, the stronger the unit of the summary map will be updated towards the weights of that neuron. The summarization algorithm will consider the weights of a composing unit "more suitable" to be the weights of the unit in the final map according to both the number of inputs recognized and the quality of adaptation of the unit (Eq. 1). With this new approach it is expected to obtain more faithful maps to the inner structure of the dataset.

This algorithm generates a structure where similar data is placed close together, with a clear relationship between the distribution of the initial data and the structure obtained by the algorithm. This relation between *realty* and *inner structure* is quite useful to a CBR system, where the stored information must be used in future situations to obtain future solutions.

Next, the system developed combining the CBR methodology and the WeVoS-SOM algorithm is explained, focusing on the main phases of the CBR cycle.

4. WeVos-CBR: an hybrid forecasting system

CBR have already been used to generate predictions in complicated environments where different parameters were [9]. In this occasion, the CBR methodology is used in combination with a summarization of SOM ensembles algorithm, in order to improve its results. The WeVos-CBR system presented here is able to generate predictions using past information as a source of knowledge to solve new problems. The available information is divided into cases that are stored in the case base. Those cases are structured, using the WeVoS-SOM algorithm.

When a new problem should be solved by the system, then the most similar data to the problem is retrieved from the case base. Then, the inner organization generated by the WeVoS-SOM algorithm is useful to recover those elements more similar to the one that is introduced in the system as a problem. The retrieved cases are used to generate the new solution, by feeding a neural network, trained to generate solutions.

If the solution generated is good enough to be proposed to the user, then it is also stored in the case base, to serve as new data to solve new problems. All this process, covering the four main phases of the CBR cycle explained before, is covered by the next sub-sections.

4.1 Pre-processing and retrieval

When the case base is created the WeVoS-SOM algorithm is used to structure it. The graphical capabilities of this novel algorithm are used in this occasion to create a model that represents the actual variability of the parameters stored in the cases. At the same time, the inner structure of the case base will make it easier to recover the most similar cases to the problem cases introduced in the system.

The WeVoS-SOM algorithm is also used to recover the most similar cases to the problem introduced in the system. That process if performed once the case base is structured keeping the original distribution of the available variables. When a new problem comes into the system, then its *virtual* position into the inner structure of the case base is calculated. The system tries to store the problem into the case base like if it was a solution. That virtual allocation serves to calculate the

position of the problem into the case base, and to recover those elements that are located close to that virtual position. Those retrieved cases are used in the next stage to generate the solution.

4.2 Reuse

After recovering the most similar cases to the problem from the case base, those cases are used to obtain a solution. *Growing RBF networks* [10] are used to generate the predicted solution corresponding to the proposed problem. The selected cases are used to train the GRBF network. This adaptation of the RBF network lets the system grow during the training phase in a gradual way increasing the number of elements (prototypes) which work as the centres of the radial basis functions. The error definition for every pattern is shown below:

$$e_i = {}^{1}\!/_{p} \cdot \sum_{k=1}^{p} \left\| t_{ik} - y_{ik} \right\| . \quad (2)$$

Where t_{ik} is the desired value of the k^{th} output unit of the i^{th} training pattern, y_{ik} the actual values ot the k^{th} output unit of the i^{th} training pattern. After the creation of the GRBF network, it is used to generate the solution to the introduced problem. The solution will be the output of the network using as input data the retrieved cases.

4.3 Revision and Retain

In order to verify the precision of the proposed solution, *Explanations* are used [11]. To justify and validate the given solution, the retrieved cases are used once again. The selected cases have their own future associated situation. Considering the case and its solution as two vectors, a distance between them can be measured by calculating the evolution of the situation in the considered conditions. If the distance between the proposed problem and the solution given smaller than the distances obtained from the selected cases, then the proposed solution considered as a good one.

Once the proposed prediction is accepted, it can be stored in the case base in order to serve to solve new problems. It will be used equally than the historical data previously stored in the case base. The *WeVoS-SOM* algorithm is used again to introduce new elements in the case base.

5. Case study – forest fires

The WeVoS-CBR system presented here has been applied to generate predictions in a forest fire scenario. Forest fires represent a great environmental risk. The main approaches that have been used to solve this problem come first from the detection of the fires [12], where different techniques have been applied. Once the fire is detected, it is important to generate predictions that should help to take decision in those contingency response situations [13]. Finally, there are complex

models that face the forest fire problem trying to forecast its evolution and to minimize its associated risks [14].

The data used to check the WeVoS-CBR system was a subset of the available data that has not been previously used in the training phase. The predicted situation was contrasted with the actual future situation as it was known (historical data was used to train the system and also to test its correction). The proposed solution was, in most of the variables, close to 90% of accuracy.

To create the cases, the geographical area analyzed was divided into small squares, each of which were considered a case, with all its associated parameters (longitude, latitude, wind, pressure, temperature, etc.). The squares determine the area to be considered in every case. The problem is represented by the current situation of the area (all its parameters and the presence or not of fire). The solution is represented by the situation in that area in a future moment (same location but parameters changed to next day, or next step –if less than a day is considered in every step-).

Table 1. Percentage of good predictions obtained with different techniques.

Number of cases	RBF	CBR	RBF + CBR	WeVoS-CBR
500	43 %	42 %	46 %	46 %
1000	46 %	48 %	53 %	61 %
3000	54 %	58 %	62 %	75 %
5000	63 %	66 %	76 %	86 %

In *table 2* a summary of the results obtained is shown. In this table different techniques are compared. The table shows the evolution of the results along with the increase of the number of cases stored in the case base. All the techniques analyzed improve its results when increasing the number of cases stored. Having more cases in the case base, makes easier to find similar cases to the proposed problem and then, the solution can be more accurate.

The *"RBF"* column represents a simple Radial Basis Function Network that is trained with all the data available. The network gives an output that is considered a solution to the problem. The *"CBR"* column represents a pure CBR system, with no other techniques included, the cases are stored in the case bases and recovered considering the Euclidean distance. The most similar cases are selected and after applying a weighted mean depending on the similarity, a solution is proposed. The *"RBF + CBR"* column corresponds to the possibility of using a RBF system combined with CBR. The recovery from the CBR is done by the Manhattan distance and the RBF network works in the reuse phase, adapting the selected cases to obtain the new solution. The results of the *"RBF+CBR"* column are, normally, better than those of the *"CBR"*, mainly because of the elimination

of useless data to generate the solution. Finally, the "*WeVoS-CBR"* column shows the results obtained by the proposed system, obtaining better results than the three previous analyzed solutions.

6. Acknowledgments.

This research has been partially supported through projects BU006A08 and SA071A08 of the JCyL and project CIT-020000-2008-2 of the Spanish Ministry of Education and Innovation. The authors would also like to thank the vehicle interior manufacturer, Grupo Antolin Ingenieria, S.A., within the framework of the project MAGNO2008 - 1028.- CENIT Project funded by the Spanish Ministry.

7. References

[1] WATSON, I. (1999) Case-based reasoning is a methodology not a technology, *Knowledge-Based Systems*, 12 (5-6), 303-308.

[2] AAMODT, A. (1991) A Knowledge-Intensive, Integrated Approach to Problem Solving and Sustained Learning, *Knowledge Engineering and Image Processing Group. University of Trondheim*.

[3] AAMODT, A. AND PLAZA, E. (1994) Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches, *AI Communications*, 7 (1), 39-59.

[4] YANG, B.S., HAN, T. AND KIM, Y.S. (2004) Integration of ART-Kohonen neural network and case-based reasoning for intelligent fault diagnosis, *Expert Systems With Applications*, 26 (3), 387-395.

[5] DIAZ, F., FDEZ-RIVEROLA, F. AND CORCHADO, J.M. (2006) Gene-CBR: A case-based reasoning tool for cancer diagnosis using microarray data sets, *Computational Intelligence*, 22 (3/4), 254-268.

[6] LIU, C.-H., CHEN, L.-S. AND HSU, C.-C. (2008) An Association-based Case Reduction Technique for Case-based Reasoning, *Information Sciences*, 178 (17), 3347-3355.

[7] KOHONEN, T. (1998) The self-organizing map, *Neurocomputing*, 21 (1-3), 1-6.

[8] PÖLZLBAUER, G. (2004) Survey and Comparison of Quality Measures for Self-Organizing Maps. In Rauber, J.P.a.G.P.a.A., *Fifth Workshop on Data Analysis (WDA'04)*. Elfa Academic Press, 67--82.

[9] CORCHADO, J.M. AND FDEZ-RIVEROLA, F. (2004) FSfRT: Forecasting System for Red Tides, *Applied Intelligence*, 21, 251-264.

[10] KARAYIANNIS, N.B. AND MI, G.W. (1997) Growing radial basis neural networks: merging supervised andunsupervised learning with network growth techniques, *Neural Networks, IEEE Transactions on*, 8 (6), 1492-1506.

[11] SØRMO, F., CASSENS, J. AND AAMODT, A. (2005) Explanation in Case-Based Reasoning–Perspectives and Goals, *Artificial Intelligence Review*, 24 (2), 109-143.

[12] MAZZEO, G., MARCHESE, F., FILIZZOLA, C., PERGOLA, N., *et al.* (2007) A Multi-temporal Robust Satellite Technique (RST) for Forest Fire Detection. *Analysis of Multi-temporal Remote Sensing Images, 2007*. 1-6.

[13] ILIADIS, L.S. (2005) A decision support system applying an integrated fuzzy model for long-term forest fire risk estimation, *Environmental Modelling and Software*, 20 (5), 613-621.

[14] SÉRO-GUILLAUME, O., RAMEZANI, S., MARGERIT, J. AND CALOGINE, D. (2008) On large scale forest fires propagation models, *International Journal of Thermal Sciences*, 47 (6), 680-694.