Neuro-symbolic System for Forecasting Red Tides *

Florentino Fdez-Riverola[1], Juan M. Corchado[2], and Jesús M. Torres[3]

[1] Dpto. de Informática, E.S.E.I., University of Vigo,
Campus Universitario As Lagoas s/n., 32004, Ourense, Spain
riverola@uvigo.es
[2] Dpto. de Informática y Automática, University of Salamanca,
Facultad de Ciencias, Plaza de la Merced, s/n., 37008, Salamanca, Spain
corchado@usal.es
[3] Dpto. de Física Aplicada, University of Vigo,
Facultad de Ciencias, Lagoas Marcosende, 36200, Vigo, Spain
jesu@uvigo.es

Abstract. A hybrid neuro-symbolic problem solving model is presented in which the aim is to forecast parameters of a complex and dynamic environment in an unsupervised way. In situations in which the rules that determine a system are unknown, the prediction of the parameter values that determine the characteristic behaviour of the system can be a problematic task. In such a situation, it has been found that a hybrid case-based reasoning (CBR) system can provide a more effective means of performing such predictions than other connectionist or symbolic techniques. The system employs a CBR model to wrap a growing cell structures network, a radial basis function network and a set of Sugeno fuzzy models to provide an accurate prediction. Each of these techniques is used in a different stage of the reasoning cycle of the CBR system to retrieve historical data, to adapt it to the present problem and to review the proposed solution. The results obtained from experiments, in which the system operated in a real environment, are presented.

1 Introduction

Forecasting the behaviour of a dynamic system is, in general, a difficult task, especially if the prediction needs to be achieved in real time. In such a situation one strategy is to create an adaptive system which possesses the flexibility to behave in different ways depending on the state of the environment. This paper presents the application of a novel hybrid artificial intelligence (AI) model to a forecasting problem over a complex and dynamic environment. The approach presented is capable of producing satisfactory results in situations in which neither artificial neural network nor statistical models have been sufficiently successful.

* This research was supported in part by PGIDT00MAR30104PR project of Xunta de Galicia, Spain

The oceans of the world form a highly dynamic system for which it is difficult to create mathematical models [1]. *Red tides* are the name for the discolourations caused by dense concentrations of the microscopic plants of the sea, the so-called phytoplankton. The discolouration varies with the species of phytoplankton, its pigments, size and concentration, the time of day, the angle of the sun and other factors. Red tides usually occur along the north west of the Iberian Peninsula in late summer and autumn [2]. The rapid increase in dinoflagellate numbers, sometimes to millions of cells per liter of water, is what is known as a *bloom* of phytoplankton (if the concentration ascends above the 100.000 cells per liter). The type of dinoflagellate on which we focus in this study is the pseudo-nitzschia spp diatom which is known to cause amnesic shellfish poisoning (ASP).

An AI approach to the problem of forecasting in the ocean environment offers potential advantages over alternative approaches, because it is able to deal with uncertain, incomplete and even inconsistent data. Several types of standard artificial neural network (ANN) have been used to forecast the evolution of different oceanographic parameters [3–5]. Our aim is to develop an autonomous and reliable forecasting mechanism. The results obtained to date suggest that the approach to be described in this paper appears to fulfil this aim.

The work presented in this paper is based on the successful results obtained with the hybrid CBR system reported in [4–6] and used to predict the evolution of the temperature of the water ahead of an ongoing vessel, in real time. The hybrid system proposed in this paper is an extension and an improvement of the previously mentioned research. The retrieval, reuse, revision and learning stages of the CBR system have been modified or changed for two reasons: to adapt the hybrid system to the previously mentioned problem and to automate completely the reasoning process of the proposed forecasting mechanism.

The structure of the paper is as follows. First a brief overview of the CBR systems for forecasting is presented. Then the red tide problem domain is briefly outlined. The hybrid neuro-symbolic system is then explained, and finally, the results obtained to date with the proposed forecasting system are presented and analyzed.

2 CBR Systems for Forecasting

Several researchers [7, 8], have used k-nearest-neighbour algorithms for time series predictions. Although a k-nearest-neighbour algorithm does not, in itself, constitute a CBR system, it may be regarded as a very basic and limited form of CBR operation in numerical domains. Other examples of CBR systems that carry out predictions can be found in [9–13].

In most cases, the CBR systems used in forecasting problems have flat memories with simple data representation structures using k-nearest-neighbour metrics in their retrieve phase. K-nearest-neighbour metric are acceptable if the system is relatively stable and well understood, but if the system is dynamic and the forecast is required in real time, it may not be possible to easily redefine the k-nearest-neighbour metrics adequately. The dominant characteristic of

the adaptation stage used in these models are similarity metrics or statistical models, although, in some systems, case adaptation is accomplished manually. If the problem is very complex, there may not be an adaptation strategy and the most similar case is used directly, but it is believed that adequate adaptation is one of the keys to a successful CBR paradigm. In the majority of the systems surveyed, case revision (if carried out at all) is performed by human expert, and in all the cases the CBR systems are provided with a small case-base. A survey of such forecasting CBR systems can be found in [14]. In this paper a method for automating the CBR reasoning process is presented for the solution of problems in which the cases are characterised predominantly by numerical information.

Traditionally, CBR systems have been combined with other technologies such as artificial neural networks, rule-based systems, constraint satisfaction problems and others, producing successful results [15]. Our proposal requires to embed two artificial neural networks and a set of fuzzy systems in the CBR life cycle.

3 Forecasting Red Tides

In the current work the aim is to develop a system for forecasting the concentrations of the pseudo-nitzschia spp, that it is the diatom that produces the most harmful red tides, at different geographical points one week in advance.

The problem of forecasting, which is currently being addressed, may be simply stated as follows:

- **Given**: a sequence of data values (representing the current state and the immediately previous one) about some physical and biological parameters,
- **Predict**: the value of a parameter at some future point(s) or time(s).

In order to forecast the concentration of pseudo-nitzschia spp at a given point one week in advance, a problem descriptor is generated on a weekly basis. A problem descriptor consists of a sequence of N sampled data values (filtered and preprocessed) recorded from the water mass for which the forecast is required, and the collection time and date. Every week the concentration of pseudo-nitzschia spp is added to a problem descriptor forming a new input vector. The problem descriptor is composed of a vector with the variables that characterise the problem recorded during two weeks. The prediction or output of the system is the concentration of pseudo-nitzschia spp one week after, as indicated in Table 1.

The forecasted values are obtained using a neural network enhanced hybrid case-base reasoning system. Figure 1 illustrates the relationships between the processes and components of the hybrid CBR system. The cyclic CBR process shown in Figure 1 the has been inspired by the work of [4] and [5]. The diagram shows the technology used in each stage, where the four basic phases of the CBR cycle are shown as rectangles.

The retrieval stage is carried out using a Growing Cell Structures (GCS) ANN [16]. The GCS facilitates the indexing of cases and the selection of those that are more similar to the problem descriptor. The GCS network groups similar cases and forms classes. When a new problem is presented to this network, it

Table 1. Variables that define a case

Variable	Unit	Week
Date	dd-mm-yyyy	W_{n-1}, W_n
Temperature	Cent. degrees	W_{n-1}, W_n
Oxygen	milliliters/liter	W_{n-1}, W_n
PH	acid/based	W_{n-1}, W_n
Transmitance	%	W_{n-1}, W_n
Fluorescence	%	W_{n-1}, W_n
Cloud index	%	W_{n-1}, W_n
Recount of diatoms	cel/liter	W_{n-1}, W_n
pseudo-nitzschia spp	cel/liter	W_{n-1}, W_n
pseudo-nitzschia spp (future)	*cel/liter*	W_{n+1}

is associated to its most representative class and all members of such class are retrieved. The reuse of cases is carried out with a Radial Basis Function (RBF) ANN [17], which generates an initial solution creating a model with the retrieved cases. The GCS network guarantees that these cases are homogeneous and can be modeled by the RBF network. The revision is carried out using a group of pondered Fuzzy systems that identify potential incorrect solutions. Finally, the learning stage is carried out when the real value of the concentration of pseudo-nitzschia spp is measured and the error value is calculated, updating the knowledge structure of all the system. The cycle of operations of the hybrid system is explained in detail in Section 3.1.

3.1 System Operation

The forecasting system uses data from two main sources: The raw data (sea temperature, salinity, PH, oxygen and other physical characteristics of the water mass) which are weekly measured by the monitoring net of toxic proliferations in the CCCMM (Centro de Control da Calidade do Medio Marino, *Oceanographic environment Quality Control Centre*, Vigo, Spain), consist of a vector of discrete sampled values (at 5, 10 and 15 meters deep) of each oceanographic parameter used in this experiment, in the form of a time series. These data values are complemented by additional data derived from satellite images, which are received and processed daily, and other data belonging to ocean buoys that record data on a daily bases. Table 1 shows the variables that characterise the problem. Data of the last 2 weeks (W_{n-1}, W_n) is used to forecast the concentration of pseudo-nitzschia spp one week ahead (W_{n+1}).

The cycle of forecasting operations (which is repeated every week) proceeds as follows. When a new problem is presented to the system, the GCS neuronal network is used to obtain the k most similar cases to the given problem (identifying the class to which the problem belongs).

In the reuse phase, the values of the weights and centers of the neural network [17] used in the previous forecast are retrieved from the knowledge base. These

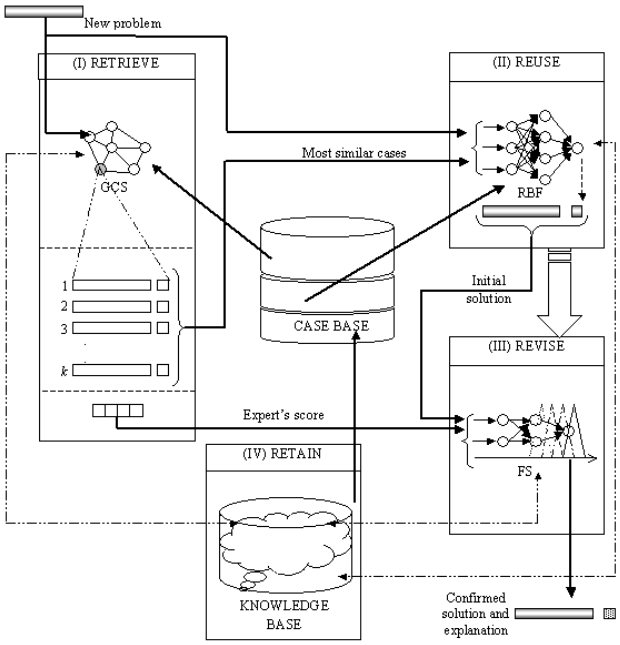

Fig. 1. Hybrid neuro-symbolic system

network parameters together with the k retrieved cases, are then used to retrain the RBF network and to obtain an initial forecast of the concentration of pseudo-nitzschia spp. During this process the values of the parameters that characterise the network are updated.

In the revision phase, the initial solution proposed by the RBF neural network is modified according to the responses of the four Fuzzy revision subsystems. Each revision subsystem has been created from the RBF network using neurofuzzy techniques [18]. For each class of the GCS neural network a vector of four values is maintained. This "importance" vector (see Figure 1) represents the accuracy of each revision subsystem with respect to a class. During the revision, the "importance" vector associated to the class to which the problem case belongs, is used to ponder the outputs of each of the fuzzy revision system. Each value of the vector is associated to one of the four revision subsystems. For each forecasting cycle, the value of the importance vector associated to the most accurate revision subsystem is increased and the other three values are proportionally decreased. This is done to give more relevance to the most accurate revision subsystem.

The revised forecast is then retained temporarily in the forecast database. When the real value of the concentration of pseudo-nitzschia spp is measured, the forecasted value for the variable can then be evaluated, by comparison of the actual and forecasted value, and the error obtained. A new case, corresponding to

CBR-STAGE	Technology	Input	Output	Process
Retrieval	GCS network.	Problem descriptor.	*K* cases.	All the cases that belong to the same class to which the GCS associates the Problem case.
Reuse	RBF network.	Problem descriptor.	Initial solution: concentration of pseudo-nitzschia spp.	The RBF network is retrained with the K retrieved cases.
Revision	4 Fuzzy systems.	Problem descriptor.	Confirmed solution: concentration of pseudo-nitzschia spp.	Four fuzzy sistems are created using the RBF network configuration with different degrees of generalization.
Retain	GCS network. RBF network. 4 Fuzzy systems.	Problem descriptor. Forecasting Error.	Configuration parameters of the GCS network, RBF network and 4 Fuzzy systems.	The configurations of the GCS network, the RBF network and the Fuzzy subsystems are updated according to the accuracy of the forecast.

Fig. 2. Summary of technologies employed by the hybrid model

this forecasting operation, is then stored in the case base. The forecasting error value is also used to update the importance vector associated to the revision subsystems of the retrieved class.

4 Results

The hybrid forecasting system has been proven in the coast of north west of the Iberian Peninsula with data collected by the CCCMM from the year 1992 until the present time. The prototype used in this experiment was set up to forecast the concentration of the pseudo-nitzschia spp diatom of a water mass situated near the coast of Vigo (geographical area A0 ((42°28.90' N, 8°57.80' W) 61 m)), one week in advance. Red tides appear when the concentration of pseudo-nitzschia spp is higher than 100.000 cel/liter. Although the aim of this experiment is to forecast the value of this concentration, the most important objective is to identify in advance if the concentration is going to be over this threshold.

The average error in the forecast was found to be 26,043 cel/liter and only 5.5% of the forecasts had an error higher than 100,000 cel/liter. Although the experiment was carried out using a limited data set, it is believed that these error value results are sufficiently representative to be extrapolated over the whole coast of the Iberian Peninsula.

Two situations of special interest are those corresponding to the *false alarms* and the *undetected blooms*. The first one happens when the model predicts bloom (concentration of pseudo-nitzschia $\geq$ 100,000 cel/liter) and this doesn't take place (real concentration $\leq$ 100,000 cel/liter). The second, more important, arise when bloom really exists and the model doesn't detect it.

Table 2 shows the predictions carried out with success (in absolute value and %) and the erroneous predictions differentiating the undetected blooms and the false alarms. This table also shows the average error obtained with all the techniques. As can be seen, the combination of different techniques in the form of the hybrid CBR system previously presented, produces better results that a RBF neural network working alone or any of the tested statistical techniques. This is due to the effectiveness of the revision subsystem and the retrained of the RBF neural network with the cases recovered by GCS network. The hybrid system is more accurate than any of the other techniques studied during this investigation.

Table 2. Summary of results forecasting pseudo-nitzschia spp

Method	Correct predictions	% Correct	Undetected blooms	False alarms	Average error (cel/liter)
CBR-ANN-FS	**191/200**	**95.5%**	**8**	**1**	**26,044**
RBF	185/200	92.5%	8	7	45,654
ARIMA	174/200	87%	10	16	71,918
Quadratic Trend	184/200	92%	16	0	70,354
Moving Average	181/200	90.5%	10	9	51,969
Simp. Exp. Smooth.	183/200	91.5%	8	9	41,943
Lin. Exp. Smooth.	177/200	88.5%	8	15	49,038

5 Conclusions

This paper has presented a problem solving method in which a CBR system is integrated with two artificial neural networks and a set of fuzzy inference systems in order to create a real-time, autonomous forecasting system. The forecasting system is able to produce a forecast with an acceptable degree of accuracy.

The method uses a CBR system to wrap a growing cell structures network (to index, organize and retrieve relevant data), a radial basis function network (that contributes with generalization, learning and adaptation capabilities) and a set of Sugeno fuzzy models (acting as experts that revise the initial solution) to provide a more effective prediction. The resulting hybrid system thus combines complementary properties of connectionist and symbolic AI methods. The results obtained may be extrapolated to provide forecasts further ahead using the same technique, and it is believed that successful results may be obtained. However, the further ahead the forecast is made, the less accurate the forecast may be expected to be. In conclusion, our hybrid approach to problem solving provides an effective strategy for forecasting in an environment in which the raw data is derived from the previously mentioned sources.

References

1. Tomczak, M., Godfrey, J. S.: Regional Oceanographic: An Introduction. Pergamon, New York, (1994)
2. Fernández, E.: Las Mareas Rojas en las Rías Gallegas. Technical Report, Department of Ecology and Animal Biology. University of Vigo, (1998)
3. Corchado, J. M., Fyfe, C.: Unsupervised Neural Network for Temperature Forecasting. Artificial Intelligence in Engineering, 13, num. 4, (1999) 351–357
4. Corchado, J. M., Lees, B.: A Hybrid Case-based Model for Forecasting. Applied Artificial Intelligence, 15, num. 2, (2001) 105–127
5. Corchado, J. M., Lees, B., Aiken, J.: Hybrid Instance-based System for Predicting Ocean Temperatures. International Journal of Computational Intelligence and Applications, 1, num. 1, (2001) 35–52
6. Corchado, J. M., Aiken, J., Rees, N.: Artificial Intelligence Models for Oceanographic Forecasting. Plymouth Marine Laboratory, U.K., (2001)
7. Nakhaeizadeh, G.: Learning prediction of time series. A theoretical and empirical comparison of CBR with some other approaches. Proceedings of First European Workshop on Case-Based Reasoning, EWCBR-93, Kaiserslautern, Germany, (1993) 65–76
8. Lendaris, G. G., Fraser, A. M.: Visual Fitting and Extrapolation. Weigend, A. S., Fershenfield, N. A. (Eds.). Time Series Prediction, Forecasting the Future and Understanding the Past. Addison Wesley, (1994) 35–46
9. Lekkas, G. P., Arouris, N. M., Viras, L. L.: Case-Based Reasoning in Environmental Monitoring Applications. Artificial Intelligence, 8, (1994) 349–376
10. Faltings, B.: Probabilistic Indexing for Case-Based Prediction. Proceedings of Case-Based Reasoning Research and Development, Second International Conference, ICCBR-97, Providence, Rhode Island, USA, (1997), 611–622
11. Mcintyre, H. S., Achabal, D. D., Miller, C. M.: Applying Case-Based Reasoning to Forecasting Retail Sales. Journal of Retailing, 69, num. 4, (1993), 372–398
12. Stottler, R. H.: Case-Based Reasoning for Cost and Sales Prediction. AI Expert, (1994), 25–33
13. Weber-Lee, R., Barcia, R. M., Khator, S. K.: Case-based reasoning for cash flow forecasting using fuzzy retrieval. Proceedings of the First International Conference on Case-Based Reasoning, ICCBR-95, Sesimbra, Portugal, (1995), 510–519
14. Corchado, J. M., Lees, B., Fyfe, C., Ress, N., Aiken, J.: Neuro-adaptation method for a case based reasoning system. Computing and Information Systems Journal, 5, num. 1, (1998), 15–20
15. Pal, S. K., Dilon, T. S., Yeung, D. S.: Soft Computing in Case Based Reasoning. Springer Verlag, London, (2000)
16. Azuaje, F., Dubitzky, W., Black, N., Adamson, K.: Discovering Relevance Knowledge in Data: A Growing Cell Structures Approach. IEEE Transactions on Systems, Man and Cybernetics, 30, (2000) 448–460
17. Fritzke, B.: Fast learning with incremental RBF Networks. Neural Processing Letters, 1, num. 1, (1994) 2–5
18. Jin, Y., Seelen, W. von., Sendhoff, B.: Extracting Interpretable Fuzzy Rules from RBF Neural Networks. Internal Report IRINI 00-02, Institut für Neuroinformatik, Ruhr-Universität Bochum, Germany, (2000)