CBR based system for forecasting red tides — **Source link** ⧉

Florentino Fdez-Riverola, Juan M. Corchado

Institutions: University of Vigo, University of Salamanca

Related papers:

- Hybrid artificial intelligence methods in oceanographic forecast models

- Unsupervised neural method for temperature forecasting

- FSfRT: Forecasting System for Red Tides

- Maximum likelihood hebbian learning based Retrieval method for CBR systems

- A forecasting solution to the oil spill problem based on a hybrid intelligent system

CBR based System for Forecasting Red Tides

Florentino Fdez-Riverola
Dpto. de Informática, University of Vigo
Campus Universitario As Lagoas s/n., 32004, Ourense, Spain
riverola@uvigo.es

Juan M. Corchado
Dpto. de Informática y Automática, University of Salamanca
Plaza de la Merced s/n., 37008, Salamanca, Spain
corchado@gugu.usal.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. The system employs a case-based reasoning 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 case-based reasoning system to retrieve, adapt and review the proposed solution to the present problem. This system has been used to predict the red tides that appear in the coastal waters of the north west of the Iberian Peninsula. The results obtained from experiments 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, which is discussed, is capable of producing satisfactory results in situations in which neither artificial neural network nor statistical models have been sufficiently successful.

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 microscopic sea plants, known as 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 coast of the Iberian Peninsula in late summer and autumn [2]. The prevailing southerly winds cause cold, nutrient-rich water to rise up from the deeper regions of the ocean to the surface, a process known as *upwelling*. Swept along with this upwelled water are dinoflagellate cysts, the resting stages of the organism, which lie dormant in the sediments on the sea floor. The high nutrient concentrations in the upwelled water, together with ideal conditions of temperature, salinity and light, trigger the germination of the cysts, so that the dinoflagellates begin to grow and divide. The rapid increase in dinoflagellate numbers, sometimes to millions of cells per liter of water, is described as a *bloom* of phytoplankton (concentration levels above the 100,000 cells per liter). Concentration of the bloom by wind and currents, as well as the dinoflagellates' ability to swim to the surface, combine to form a red tide. This study focusses on the pseudo-nitzschia spp diatom dinoflagellate, which causes amnesic shellfish poisoning (or ASP).

An artificial intelligence 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 networks (ANN) have been used to forecast the evolution of different oceanographic parameters [3, 4 & 5]. The reported work shows how difficult it is to train neural networks to successfully forecast time series of oceanographic and/or biological parameters such as the temperature, chlorophyll and salinity of the water. Statistical models such as Auto-Regressive Integrated Moving Averages (ARIMA) have been applied, but the results obtained so far have indicated that neural networks (although not accurate enough) have a greater facility for forecasting such parameters than statistical models [6].

An important aim in the current work is to develop a universal forecasting mechanism, in the sense that it might operate effectively anywhere, at any point, in coastal waters, and at any time of the year without human intervention. The results obtained to date suggest that the approach described in this paper appears to fulfil these aims.

The study is based on the successful results obtained with the hybrid case-based reasoning system reported [4, 5 & 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 afore-mentioned problem and to completely automate the reasoning process of the proposed forecasting mechanism.

The structure of the paper is as follows: first the red tide problem domain is briefly outlined, then the hybrid neuro-symbolic system is explained, and finally, the results obtained to date with the proposed forecasting system are presented and analyzed.

2. The Red Tides Problem Domain

Recently red tides have been very much in the news. Dinoflagellates are usually regarded as the causative organisms, but not all red tides are caused by dinoflagellates and not all dinoflagellates cause red tides. Even the colour factor is variable: so-called *red tides* may be brown, yellow, green, etc. Some red tides may be very extensive and several square kilometers of ocean may be affected, even to the extent that satellites have been used to track blooms. Surface waters of these blooms are associated with the production of toxins, resulting in mortality of fish and other marine organisms. Toxic blooms of dinoflagellates fall into three categories: (1) blooms that kill fish but few invertebrates; (2) blooms that kill primarily invertebrates; (3) blooms that kill few marine organisms, but whose toxins are concentrated within the siphons, digestive glands, or mantle cavities of filter-feeding bivalve mollusc such as clams, oysters, and escallops.

What causes such blooms? A range of factors seem to be involved, but very little definite information is available. In some places there seems to be a strong correlation between the occurrence of upwelling (nutrient-rich waters coming in from deep water) and such blooms [7]. But, in other areas, the blooms have been found to be associated with tidal turbulence or they seem to be set off by heavy rainfall on the land, the runoff washing phosphates into the sea and also lowering the salinity, all factors which seem to favour dinoflagellate growth. It is also thought that Vitamin B_{12}, which is required by most dinoflagellates, may also be washed into the sea from the soil and salt-marsh areas, where it is produced by bacteria and bluegreen algae. Humic substances have also been suggested as possible causative agents.

2.1 Recent Trends

The nature of the red tides problem has changed considerably over the last two decades around the world. Where formerly a few regions were affected in scattered locations, now virtually every coastal state is threatened, in many cases over large geographic areas and by more than one harmful or toxic algal species [8]. Few would argue that the number of toxic blooms, the economic losses from them, the types of resources affected, and the number of toxins and toxic species have all increased dramatically in recent years in all over the world. Disagreement only arises with respect to the reasons for this expansion. Possible explanations include: a) species dispersal through currents, storms, or other natural mechanisms; b) nutrient enrichment of coastal waters by human activities, leading to a selection for, and proliferation of, harmful algae; c) increased aquaculture operations which can enrich surrounding waters and stimulate algal growth; d) introduction of fishery resources (through aquaculture development) which then exposes itself to the presence of indigenous harmful algae in the surrounding waters; e) dispersal of the species via ship ballast water or shellfish seeding activities; f) long-term climatic trends in temperature, wind speed, or insolation and g) increased scientific and regulatory scrutiny of coastal waters and fishery products and improved chemical analytical capabilities that lead to the discovery of new toxins and toxic events [9].

2.2 Models

Models of dinoflagellate blooms have been developed fom several different perspectives. [10] examined the response of a swimming dinoflagellate to internal waves and showed that accumulation of motile and non-motile cells may occur due to an internal wave field, with the accumulation of vertically migrating cells being most significant. These models consider only the physics of the wave field and the swimming behavior of the phytoplankton, without regard to the phytoplankton response to nutrients or light. Others have examined the response of phytoplankton to the flow field of Langmuir cells [11] or to 2-dimensional, cross-frontal circulation [12], to name just two of many physical systems that have been studied in this theoretical context. The growth and accumulation of individual harmful algal species in a mixed planktonic assemblage are exceedingly complex processes involving an array of chemical, physical, and biological interactions. Our level of knowledge about each of the many species varies significantly, and even those most widely studied remain poorly characterized with respect to bloom or population dynamics. Resolution of various rate processes integral to the population dynamics (i.e., input and losses due to growth, grazing, encystment, excystment, and physical advection) has not been accomplished, but is fundamental to the long-term management of fishery resources or marine habitats affected by harmful algae. Many of the processes are difficult to quantify in the field because harmful species often represent only a small fraction of the biomass in natural samples. The end result is that despite the proven utility of models in so many oceanographic disciplines, there are no predictive models of population development, transport, and toxin accumulation. There is thus a clear need to develop models for regions subject to red tides, and to incorporate biological behavior and population dynamics into those simulations [9].

3.   Forecasting Red Tides

In the current work, the aim is to develop a system for forecasting one week in advance the concentrations (in cells per liter) of the pseudonitzschia spp, the diatom that produces the most harmful red tides, at different geographical points. The problem of forecasting, which is currently being addressed, may be simply stated as follows:

- **Given**: a sequence of data values (representative of the current and immediately previous state) relating to some physical and biological parameters,

- **Predict**: the value of a parameter at some future point(s) or time(s).

The raw data (sea temperature, salinity, PH, oxygen and other physical characteristics of the water mass) which is measured weekly by the monitoring network for toxic proliferations in the CCCMM (Centro de Control da Calidade do Medio Marino, *Oceanographic environment Quality Control Centre*, Vigo, Spain), consists of a vector of discrete sampled values (at 5 meters' depth) of each oceanographic parameter used in the experiment, in the form of a time series. These data values are complemented by additional data derived from satellite images,

which is received and processed daily, and other data belonging to ocean buoys that record data on a daily basis. In the present study, the parameter for prediction is the concentration of pseudo-nitzschia spp in a given water mass one week in advance.

3.1 The Hybrid Forecasting System

In order to forecast the concentration of pseudo-nitzschia spp at a given point a 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 pre-processed) recorded from the water mass to which the forecast will be applied. The problem descriptor also contains various other numerical values, including the current geographical location of the sensor buoys 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 over two weeks. The prediction or output of the system is the concentration of pseudo-nitzschia spp one week later, as indicated in Table I.

Table I. 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	cell/liter	W_{n-1}, W_n
Pseudo-nitzschia spp	cell/liter	W_{n-1}, W_n
Pseudo-nitzschia spp (future)	*cell/liter*	W_{n+1}

The forecast values are obtained using a neural network enhanced hybrid case-base reasoning system. Figure 2 illustrates the relationships between the processes and components of the hybrid CBR system. The cyclic CBR process shown in the figure has been inspired by the work of [4] and [5]. The diagram shows the technology used at 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. The GCS facilitates the indexation of cases and the selection of those that are most similar to the problem descriptor. The reuse of cases is carried out with a Radial Basis Function (RBF) ANN, which generates an initial solution creating a model with the retrieved cases. 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, and updating the knowledge structure of

the whole system. The cycle of operations of the hybrid system is explained in the following section in detail.
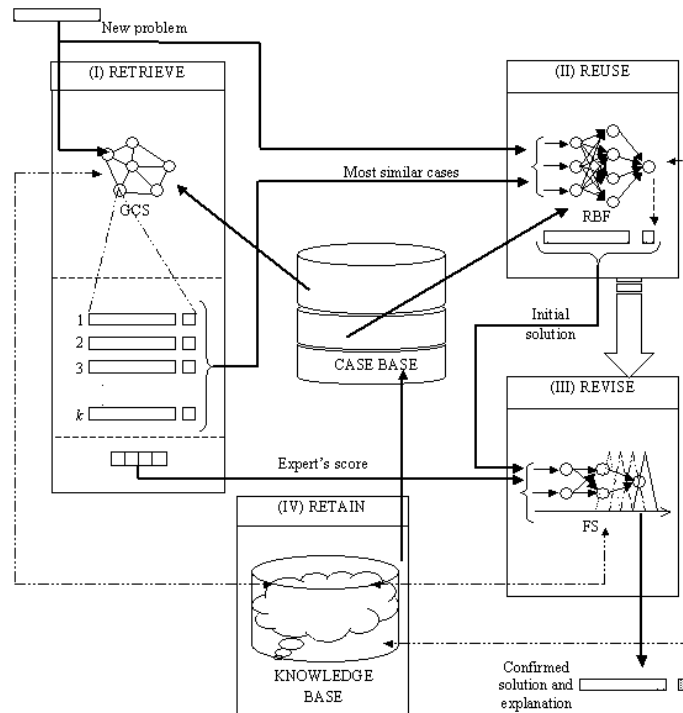

Figure 2. Hybrid neuro-symbolic system.

The neural networks used in the current study are: a) Growing Cell Structures (GCS) [13] which are a variation of Kohonen's Self-Organising Maps and provide the basis for powerful information retrieval applications and similarity visualization tools offering several advantages over both non-self-organising neural networks and the Kohonen self-organising maps cited above, and b) the Radial Basis Function (RBF) [14], in which the input layer is a receptor for the input data, whilst the hidden layer performs a non-linear transformation from the input space to the hidden layer space.

3.2   System Operation

The forecasting system uses data from two main sources: (i) the data (coming from the buoys and monitoring net) used to create a succession of problem descriptors, characterizing the current forecasting situation; (ii) data derived from satellite images stored on a database. The satellite image data values are used to generate cloud and superficial temperature indices which are then stored with the problem descriptor and subsequently updated during the CBR operation. Table I shows the variables that characterise the problem. Data from the previous 2 weeks (W_{n-1}, W_n)

is used to forecast the concentration of pseudo-nitzschia spp one week ahead (W_{n+1}).

Several experiments have been carried out over a testing data set in order to identify the optimum number of weeks for constructing a case. Table II shows a summary of the results using the hybrid system to predict the concentration of pseudo-nitzschia spp a week ahead. Each row shows the results obtained when forecasting the concentration of pseudo-nitzschia spp using data of the last 1, 2 and 3 weeks to construct the cases. The best results were obtained using data of 2 weeks previous (W_{n-1}, W_n).

Table II. Summary of results using a RBF with information coming from several weeks.

Week	MAE	Incorrect predictions	Not detect.	False alarms
1	36553,06	15	8	4
2	**32573,88**	**9**	**5**	**7**
3	46798,66	15	5	10

Two situations of special interest are those corresponding to the *false alarms* and the *blooms not detected*. The former refers to predictions of bloom (concentration of pseudo-nitzschia = 100,000 cell/liter) which don't actually materialize (real concentration = 100,000 cell/liter). The latter, more significant occurrence arises when a bloom exists but the model fails to detect it. Another unwelcome situation occurs when the number of predictions exceeds an absolute error of 100,000 cell/liter (labelled as incorrect predictions).

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.

Figure 3. Summary of technologies employed by the hybrid system.

The cycle of forecasting operations (which is repeated every week) proceeds as follows:

First a new problem instance is created from the pre-processed data cited above. When a new problem is presented to the system, the GCS neuronal network is used to obtain k more similar cases to the given problem (identifying the class to which the problem belongs, see Figure 3).

In the reuse phase, the values of the weights and centers of the neural network used in the previous forecast are retrieved from the knowledge base. These 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 (see Figure 3). 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 [15]. For each class of the GCS neural network a vector of four values is maintained (see Figure 3). This "importance" vector is initialised with a value of (0.25, 0.25, 0.25, 0.25) and represents the accuracy of each revision subsystem with respect to a class. The sum of the four values of the vector should be one. During revision, the class-associated "importance" vector associated to which the problem case belongs is used to ponder the outputs of each fuzzy revision system. Each vector value is associated with 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 in order 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 forecast value for the variable can then be evaluated, though comparison of the actual and forecast value and the error obtained (see Figure 3). A new case, corresponding to this forecasting operation, is then stored in the case base. The forecasting error value is also used to update the importance vector associated with the revision subsystems of the retrieved class.

3.3 Running Growing Cell Structures

The GCS used in this work is characterized by a two-dimensional space, where the cells (neurons) are connected and organized into triangles. Each cell in the network is associated with a weight vector, w, of the same dimension than the input data. At the beginning of the learning process, the weight vector of each cell is initialized with random values [16]. The basic learning process in a GCS network consists of topology modification and weight vector adaptations carried out in three steps. The training vectors of the GCS network are the cases stored in the CBR case-base, as indicated in Figure 2.

In the first step of each learning cycle, the cell c, with the smallest euclidean distance between its weight vector, w_c, and the actual input vector, x, is chosen as the *winner cell* or best-match cell.

The second step consists in the adaptation of the weight vector of the winning cells and their neighbours.

In the third step, a *signal counter* is assigned to each cell, wich reflects how often a cell has been chosen as winner.

Growing cell structures also modify the overall network structure by inserting new cells into those regions that represent large portions of the input data, or removing cells that do not contribute to the input data representation.

Each cell of the CGS neural network has an associated weighted vector. These weighted vectors are used by the fuzzy systems during the revision stage, as will be shown later.

Figure 4 provides a more concise description of the GCS-based case retrieval regime described above, where v_x is the value feature vector describing the new query case X, GCS is the set of cells describing the GCS topology after the training and K is the retrieved set of most relevant cases.

```
procedure retrieve (in vₓ, GCS, out K)
{
01  CD ← ∅ /* vector of pairs(cell, distance) */
02  for each cell c ∈ GCS
03          compute distance: d_c ← DIS(vₓ,w_c)
04          assign cell-distance pair: CD ← (c,d_c)
05  order_by_distance(CD) /* ascending */
06  for each pair p ∈ CD
07          K ← get_cases_from_cell(p)
08          if |K| > 0 then
09                  go to line 10 /* non-empty cell */
10  end.
}
```

Figure 4. GCS-based case retrieval.

The neural network topology of a GCS network is incrementally constructed on the basis of the training data presented to the network. Effectively, such a topology represents the result of the basic clustering procedure (see Figure 4). Such a topology has the added advantage that inter-cluster distances can be precisely quantified. Since such networks contain explicit distance information, they can be used effectively in CBR to represent an *indexing structure* which indexes sets of cases in the case base and a *similarity measure* between case sets.

3.4　Running Radial Basis Function

The RBF network used in the framework of this experiment, uses 18 input neurons (see Table I), between three and fifty neurons in the hidden layer and a

single neuron in the output layer. Input vector is presented to the network; the output of the network is the concentration of pseudo-nitzschia spp for a given water mass. Initially, three vectors are randomly chosen from the training data set and used as centers in the middle layer of the RBF network. All the centers are associated with a Gaussian function, the width of which, for all the functions, is set to the value of the distance to the nearest center multiplied by 0.5 (see [14] for more information about RBF network).

Training of the network is carried out by presenting pairs of corresponding input and desired output vectors. After an input vector has activated each Gaussian unit, the activations are propagated forward through the weighted connections to the output units, which sum all incoming signals. The comparison of actual and desired output values enables the mean square error (the quantity to be minimized) to be calculated.

The closest center to each particular input vector is moved toward the input vector by a percentage a of the present distance between them. By using this technique the centers are positioned close to the highest densities of the input vector data set. The aim of this adaptation is to force the centers to be as close as possible to as many vectors from the input space as possible. The value of a is linearly decreased by the number of iterations until its value becomes zero; then the network is trained for a number of iterations (1/4 of the total of established iterations for the period of training) in order to obtain the best possible weights for the final value of the centers.

A new center is inserted into the network when the average error in the training data set does not fall by more than 15% after n iterations (where n is calculated dividing the value that corresponds to the 3/4 parts of the total of iterations among the maximum number of centers of the hidden layer, 50). To calculate the place where the new center will be inserted, the center C, with the greatest accumulated error is selected. A new center is then inserted near C with an average of the input data vectors of the two near centers.

3.5 Running Fuzzy System

The construction of the revision subsystem is carried out in two main steps.

(i) First, a Sugeno-Takagi fuzzy model [17] is generated using the trained RBF network configuration (center and weights). In order to transform a RBF neural network to a well interpretable fuzzy rule system, the following conditions should be satisfied:

- The basis functions of the RBF neural network have to be Gaussian functions.

- The output of the RBF neural network has to be normalized. The basis functions may have different variances.

- A certain number of basis functions for the same input variable should share a mutual center and a mutual variance.

(ii) A measure of similarity is applied to the fuzzy system with the purpose of reducing the number of fuzzy sets describing each variable in the model. Similar fuzzy sets for one oceanographic parameter are merged to create a common fuzzy set to replace them in the rule base. If the redundancy in the model is high, merging similar fuzzy sets for each variable might result in equal rules that also can be merged, thereby reducing the number of rules as well.

In our model, four fuzzy inference subsystems have been created, starting from the first (with no generalization at all), with different generalization degrees for carrying out the revision of the initial prediction. When similar fuzzy sets are replaced by a common fuzzy set representative of the originals, the system's capacity for generalization increases. Four fuzzy sets are associated with each case class. The importance value of the fuzzy set that best suits a particular class is increased and the other three are proportionally decreased. This process is carried out because it is difficult to ascertain in advance the optimum level of generalisation for a given data set. Given a problem descriptor and a forecast proposed for it, each of the four fuzzy inference subsystems generate a solution that is pondered according to the importance vector associated GCS class to which it belongs, as previously before.

The value generated by the revision subsystem is compared with the prediction carried out by the RBF and its difference (in percentage) is calculated. If the initial forecast doesn't differ by more than 10% of the solution generated by the revision subsystem, this prediction is supported and its value is considered as the final forecast. If, on the contrary, the difference is greater than 10%, the average value between the value obtained by the RBF and that obtained by the revision subsystem is calculated, and this revised value adopted as the final output of the system. This threshold has been identified after carrying out several experiments and following the advice of human experts.

The exposed revision subsystem improves the generalization ability of the RBF network. Fuzzy models, especially if acquired from data, may contain redundant information in the form of similarities between fuzzy sets. As similar fuzzy sets represent compatible concepts in the rule base, a model with many similar fuzzy sets becomes redundant, unnecessarily complex and computationally demanding. The simplified rule bases allow us to obtain a more general knowledge of the system and gain a deeper insight into the logical structure of the system to be approximated.

3.6   Retain

As mentioned before, when the real value of the concentration of pseudo-nitzschia spp is known, a new case containing the problem descriptor and the solution is stored in the case base. The importance vector associated with the retrieved class is updated in the following way: The error percentage with respect to the real value is calculated. The revision subsystem that has produced the most accurate prediction is identified and the error percentage value previously calculated is added to the degree of importance associated with the fuzzy subsystem in question. As the sum of the four importance values associated to a class has to be one, the four values

are normalized, the sum dividing up accordingly between them. When the new case is added to the case base, its class is identified. The class is updated and the new case is incorporated into the network for future use.

4.   Results

The average error in the forecast was found to be 26,043.66 cell/liter and only 5.5% of the forecasts had an error higher than 100,000 cell/liter. Although the experiment was carried out using a limited data set (geographical area A0 ((42º28.90' N, 8º57.80' W) 61 m)), it is believed that these error value results are significant enough to be extrapolated along the whole coast of the Iberian Peninsula.

Table III shows the predictions carried out with success (in absolute values and %) and the erroneous predictions differentiating the *not detected blooms* from the *false alarms*.

Table III. Summary of results using the CBR-ANN-FS Hybrid System.

OK	OK (%)	Not detect.	False alarms
191/200	95.5%	8	1

Further experiments have been carried out to compare the performance of the CBR-ANN-FS hybrid system with several other forecasting approaches. These include standard statistical forecasting algorithms and the application of several neural networks methods. The results obtained from these experiments are listed in Table IV.

Table IV shows the number of successful predictions (in absolute value and %) as well as the blooms not detected and false alarms for each method. As it indicates, 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 and any of the other techniques studied during this investigation. This is due to the effectiveness of the revision subsystem and the re-training of the RBF neural network with the cases recovered by the GCS network. The performance of the hybrid system is better than the other methods at each of the individual geographical monitoring points.

Table IV. Summary of results using the CBR-ANN-FS Hybrid System.

Method	OK	OK (%)	Not detect.	False alarms
RBF	185/200	92.5%	8	7
ARIMA	174/200	87.0%	10	16
Quadratic Trend	184/200	92.0%	16	0
Moving Average	181/200	90.5%	10	9
Simple Exp. Smoothing	183/200	91.5%	8	9
Brown's Lin. Exp. Smooth.	177/200	88.5%	8	15

Table V shows the average error obtained with the hybrid model, a standard RBF network, an ARIMA model, a Quadratic Trend, a Moving Average, a Simple Exp. Smoothing, a Brown's Linear Exp. Smoothing and a Finite Impulse Response ANN [3], which was not able to converge for this type of problem.

Table V. Average error in the forecast with other techniques and the CBR-ANN-FS Hybrid System.

Method	Type	Average error (cell/liter)
CBR-ANN-FS	Hybrid System	26,043.66
RBF	ANN	45,654.20
FIR	ANN	-
ARIMA	Statistics	71,918.15
Quadratic Trend	Statistics	71,918.15
Moving Average	Statistics	71,918.15
Simple Exp. Smoothing	Statistics	41,943.26
Brown's Lin. Exp. Smooth.	Statistics	41,943.26

5. Conclusions and Future Work

In summary, this paper has presented an automated hybrid CBR system that combines a case-based reasoning system integrated with two artificial neural networks and a set of fuzzy inference systems in order to create a real time autonomous forecasting system. The model employs a case-based reasoning model that incorporates a growing cell structures network (for the index tasks to organize and retrieve relevant data), a radial basis function network (that contributes 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 both connectionist and symbolic AI methods.

The developed prototype is able to produce a forecast with an acceptable degree of accuracy. 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. The developed prototype can not be used in a particular geographical area if there are no stored cases from that area. Once the system is in operation and it is forecasting, a succession of cases will be generated, enabling the hybrid forecasting mechanism to evolve and to work autonomously.

In conclusion, the hybrid reasoning problem solving approach provides an effective strategy for forecasting in an environment in which the raw data is derived from different sources and it can be represented by means of a vector of numeric values. This model may be used to forecast in complex situations where the problem is characterized by a lack of knowledge and where there is a high degree of dynamism. The model presented here will be tested in different water masses and a

distributed forecasting system will be developed based on the model in order to monitor 500 km. of the North West coast of the Iberian Peninsula.

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, Departamento de Ecología y Biología Animal. Universidad de Vigo, 1998.
3. Corchado, J.M. & Fyfe, C. Unsupervised Neural Network for Temperature Forecasting. Artificial Intelligence in Engineering, 1999, 13(4):351–357.
4. Corchado, J.M. & Lees, B. A Hybrid Case-based Model for Forecasting. Applied Artificial Intelligence, 2001, 15(2):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, 2001, 1(1):35–52.
6. Corchado, J.M. Aiken, J. & Rees, N. Artificial Intelligence Models for Oceanographic Forecasting. Plymouth Marine Laboratory, U.K., 2001.
7. Fraga, S. Anderson, D.M. Bravo, I. Reguera, B. Steidinger K.A. & Yetsch, C.M. Influence of upwelling relaxation on dinoflagellates and shellfish toxity in Ria de Vigo. Es t. Coast and Shelf Sci., 1988, 27:349–361.
8. Hallegraeff, G.M. A review of harmful algal blooms and their apparent global increase. Phycologia, 1993, 32:79–99.
9. Anderson, D.M. Toxic algal blooms and red tides: a global perspective. RedTides: Biology Environmental Science and Toxicology. Edited by Okaichi, T. Anderson, D.M & Nemoto, T., Elsevier, 11-16, New York, 1989.
10. Kamykowski, D. The simulation of a southern California red tide using characteristics of a simultaneously-measured internal wave field. Ecol. Model, 1981, 12:253–265.
11. Watanabe, M. & Harashima, A. Interaction between motile phytoplankton and Langmuir circulation. Ecol. Model, 1986, 31:175–183.
12. Franks, P.J.S. & Anderson, D.M. Toxic phytoplankton blooms in the southwestern Gulf of Maine: testing hypotheses of physical control using historical data. Marine Biology, 1992, 112:165–174.
13. Fritzke. B. Growing Self-Organizing Networks - Why?. ESANN'96, Brussels, 1996, 61–72.
14. Fritzke, B. Fast learning with incremental RBF Networks. Neural Processing Letters, 1994, 1(1):2–5.
15. Jin, Y. Seelen, W von. & Sendhoff, B. Extracting Interpretable Fuzzy Rules from RBF Neural Networks. Technical Institut für Neuroinformatik. Ruhr-Universität Bochum, January 2000.

16. Fritzke, B. Growing Cell Structures - A Self-organizing Network for Unsupervised and Supervised Learning. Technical Report, International Computer Science Institute. Berkeley, 1993.
17. Takagi, T. & Sugeno, M. Fuzzy identification of systems and its applications to modeling and control. IEEE Transactions on Systems, Man, and Cybernetics, 1985, 15:116–132.