Applied Intelligence 21, 251–264, 2004

FSfRT: Forecasting System for Red Tides

FLORENTINO FDEZ-RIVEROLA
Dpto. de Informática, E.S.E.I., University of Vigo Edificio Politécnico, Campus Universitario As Lagoas, s/n., 32004, Ourense, Spain
riverola@uvigo.es

JUAN M. CORCHADO
Dpto. de Informática y Automática, Facultad de Ciencias, University of Salamanca, Plaza de la Merced, s/n., 37008, Salamanca, Spain
corchado@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. In such a situation, it has been found that a hybrid case-based reasoning system can provide a more effective means of performing such predictions than other connectionist or symbolic techniques. 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 at a different stage of the reasoning cycle of the case-based reasoning system to retrieve historical data, to adapt it to the present problem and to review the proposed solution. 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, 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, 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. If conditions at the water's surface become unfavourable for the dinoflagellates, for example, if the nutrients are depleted or the bloom is dispersed by wind and currents, the dinoflagellates will again form dormant cysts and sink to the sea floor. This study focusses on the pseudonitzschia 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 [3–5]. 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 [4–6] 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 a brief overview of the basic concepts that characterize the case-based reasoning model is presented; the red tide problem domain is briefly outlined; the hybrid neuro-symbolic system is explained; and finally, the results obtained to date with the proposed forecasting system are presented and analyzed.

2. CBR Systems Overview

Although knowledge-based systems (KBS) represent one of the commercial successes resulting from artificial intelligence research, their developers have encountered several problems [7]. Knowledge elicitation, a necessary process in the development of rule-based systems, can be problematic. The implementation of a KBS can also be complex, and, once implemented, it may also be difficult to maintain. With the aim of overcoming these problems [8] proposed a revolutionary approach, case-based reasoning, which is, in effect, a model of human reasoning. The idea underlying CBR is that people frequently rely on previous problem-solving experiences when solving new problems. This assertion may be verified in many day to day problem-solving situations by simple observation or by psychological experimentation [9]. Since the ideas underlying case-based reasoning were first proposed, CBR systems have been found to be successful in a wide range of application areas [7, 10, 11].

A case-based reasoning system solves new problems by adapting solutions that were used to solve previous problems [12]. The case base holds a number of cases, each of which represents a problem together with its corresponding solution. Once a new problem arises, a possible solution to it is obtained by retrieving similar cases from the case base and studying their recorded solutions. A CBR system is dynamic in the sense that, in operation, cases representing new problems together with their solutions are added to the case base, redundant cases are eliminated and others are created by combining existing cases.

A CBR system analyses a new problem situation, and by means of indexing algorithms, retrieves previously stored cases, together with their solution, by matching them against the new problem situation, then adapts them to provide a solution to the new problem by reusing knowledge stored in the form of cases, in the case base. All of these actions are self-contained and may be represented by a cyclic sequence of processes,

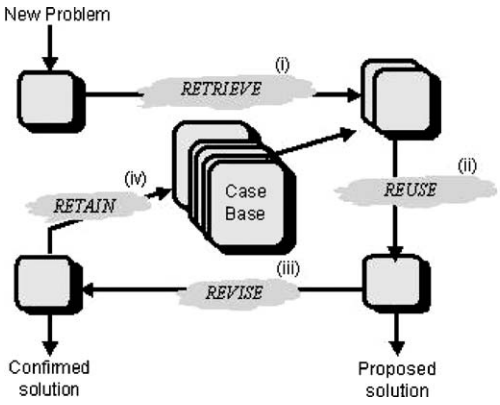

Figure 1. The classic CBR cycle.

in which human interaction may be needed. Case-based reasoning can be used by itself or as part of another intelligent or conventional computing system. Furthermore, case-based reasoning can be a particularly appropriate problem-solving strategy when the knowledge required to formulate a rule-based model of the domain is difficult to obtain, or when the number or complexity of rules relating to the problem domain is too great for conventional knowledge acquisition methods.

A typical CBR system is composed of four sequential steps which are called into action each time a new problem is to be solved [7, 10, 13]. Figure 1 outlines the basic CBR cycle.

The purpose of the retrieval step is to search the case base and select one or more previous cases that most closely match the new problem situation, together with their solutions. The selected cases are reused to generate a solution appropriate to the current problem situation. This solution is revised if necessary and finally, the new case (i.e. the problem description together with the obtained solution) is stored in the case base. Cases may be deleted if they are found to produce inaccurate solutions, they may be merged together to create more generalised solutions, and they may be modified, over time, through the experience gained in producing improved solutions. If an attempt to solve a problem fails and it is possible to identify the reason for the failure, then this information should also be stored in order to avoid the same mistake in the future. This corresponds to a common learning strategy employed in human problem-solving. Rather than creating general relationships between problem descriptors and conclusions, as is the case with rule-based reasoning, or relying on general knowledge of the problem domain, CBR systems are able to utilise the specific knowledge of previously experienced, concrete problem situations. A CBR system provides an incremental learning process because each time a problem is solved, a new experience is retained, thus making it available for future reuse.

In the CBR cycle there is normally some human interaction. Whilst case retrieval and reuse may be automated, case revision and retention are often undertaken by human experts. This is a current weakness of CBR systems and one of their major challenges. 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.

2.1. *CBR Systems for Forecasting*

Several researchers [14, 15] 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.Nakhaeizadeh [14] uses a relatively complex hybrid CBR-ANN system. In contrast, Lendaris and Fraser [15] forecast a data set simply by searching a given sequence of data values for segments that closely match the pattern of the last n measurements and then supposing that similar antecedent segments are likely to be followed by similar consequent segments. Other examples of CBR systems that carry out predictions can be found in [16–18].

In most cases, the CBR systems used in forecasting problems have flat memories with simple data

representation structures using k-nearest-neighbour metric 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 be no planned 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 [19].

Traditionally, CBR systems have been combined with other technologies like artificial neural networks, rule-based systems, constraint satisfaction problems and others, producing successful results [20] and [11], but the particularities of the problem described mean that these techniques are not the most appropriate for obtaining an accurate prediction.

3. 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 [21]. 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 blue-green algae. Humic substances have also been suggested as possible causative agents.

3.1. *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 [22]. 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 surroundingwaters; (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 [23].

3.2. *Models*

Models of dinoflagellate blooms have been developed from several different perspectives. Kamykowski [24] 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 [25] or to 2-dimensional, cross-frontal circulation [26], 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, 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 [23].

4. 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 pseudo-nitzschia spp, the diatom that produces the most harmful red tides, at different geographical points. The approach builds on the methods and expertise previously developed in earlier research [3–5]. 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.

4.1. *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 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

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	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}

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.

The neural networks used in the current study are: (a) Growing Cell Structures (GCS) [27] which are a variation of Kohonen's Self-Organising Maps and provide the basis for powerful information retrieval applications and similarity visualization tools offering serveral advantages over both non-self-organising neural networks and the Kohonen self-organising maps cited above, and (b) the Radial Basis Function (RBF) [28], 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.

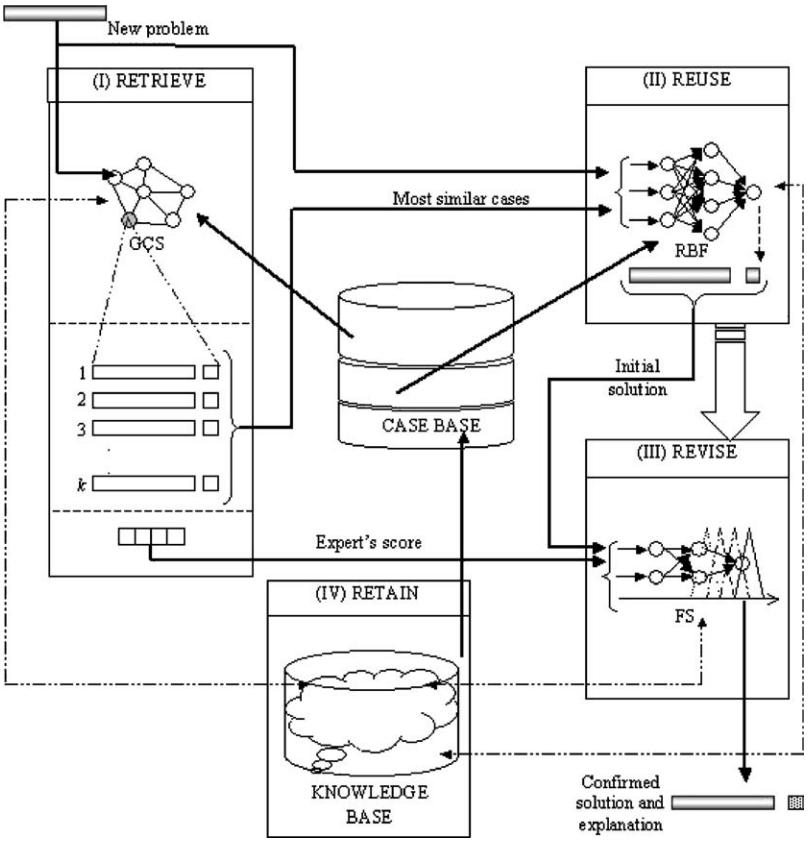

Figure 2. Hybrid neuro-symbolic system.

Table 2. Summary of results using a RFB with information coming from several weeks.

Weeks	MAE	Incorrect predictions	Not detect.	False alarms
1	36,553.06	15	8	4
2	32,573.88	9	5	7
3	46,798.66	15	5	10

4.2. 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 1 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 theoptimum number of weeks for constructing a case. Table 2 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).

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 $\geq$100,000 cell/liter) which don't actually materialize (real concentration $\leq$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).

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

First a new problem instance is created from the preprocessed 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 Table 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 Table 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 [29]. For each class

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

CBR-stage	Technology	Input	Output	Process
Retrieval	GCS network.	Problem descriptor.	k similar cases.	All the cases (k) that belong to the same class to which the GCS associates the problem case are retrieved.
Reuse	RBF network.	Problem descriptor. k similar cases.	Initial solution: concentration of pseudo-nitzschia spp.	The RBF network is retrained with the k retrieved cases.
Revision	4 Fuzzy systems.	Problem descriptor. Initial solution.	Confirmed solution: concentration of pseudo-nitzschia spp.	Four Fuzzy systems 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.

of the GCS neural network a vector of four values is maintained (see Table 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 to which the problem case belongs is used to ponder the outputs of each fuzzy revision system. Each vestor value is associated with one of the four revision subsystems. For each forecasting cycle, the value of the importance vector associated with 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, through comparison of the actual and forecast value and the error obtained (see Table 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.

4.3. Growing Cell Structures Operation

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 [30]. 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 Fig. 2.

In the first step of each learning cycle, the cell c, with the smallest 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, which reflects how often a cell has been chosen as winner.

```
   procedure RETRIEVE (input: vx, confGCS; output: K, P)
   {
00    begin.
01       CD ← ∅ /* vector of pairs (cell, distance) */
02       for each cell c ∈ confGCS do
03                compute_distance: dc ← DIS(vx, wc)
04                assign_cell-distance-pair: CD ← (c, dc)
05       order_by_distance(CD) /* ascending */
06       for each pair p ← CD do
07                K ← get_cases_from_cell(p)
08                if |K| > 0 then
09                         go_to_line 10 /* non-empty cell */
10    end.
   }
```

Figure 3. GCS-based case retrieval.

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 GCS 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 3 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.

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 Fig. 3). 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.

4.4. Radial Basis Function Operation

The RBF network used in the framework of this experiment, uses 18 input neurons (see Table 1), 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 pseudonitzschia 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 certers 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 [26] 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 certer 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 certers are positioned close to the highest densities of the input vector data set. The aim of this adaptation is to force the certers 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 certers.

A new certer 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 certer is then inserted near C with an average of the input data vectors of the two near certers.

4.5. *Fuzzy System Operation*

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

(i) First, a Sugeno-Takagi fuzzy model [31] 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. Figure 4 shows how the fuzzy set generalization is carried out given a variable (i.e. temperature).

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 (see Fig. 4). 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 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 mentioned.

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

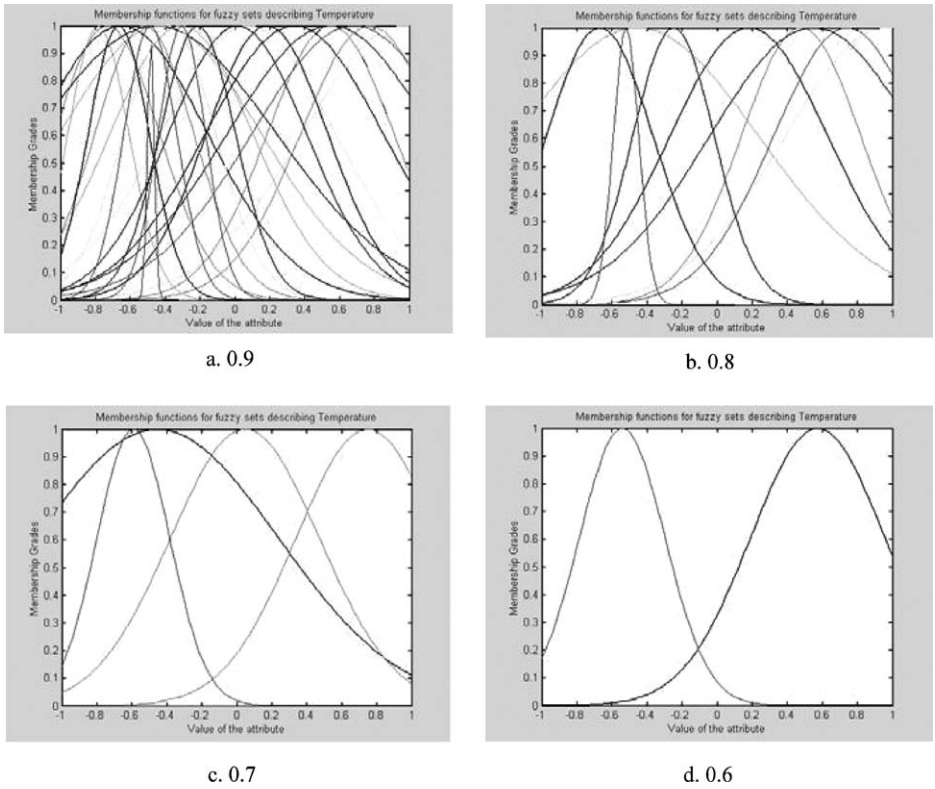

Figure 4. Different levels of generalization in a fuzzy set.

value adopted as the final output of the system. This threshold has been identified after carriying 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.

4.6. 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.

5. Results

The hybrid forecasting system has been tested along the north west coast of the Iberian Peninsula with data collected by the CCCMM from the year 1992 until the present. 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

Table 4. Summary of results using the CBR-ANN-FS hybrid system.

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

the coast of Vigo, a week in advance. Red tides appear when the concentration of pseudo-nitzschia spp is higher than 100,000 cell/liter. Although the aim of this experiment is to forecast the value of the concentration, the most important aspect is to identify in advance if the concentration is going to exceed this threshold.

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 ((42°28.90' N, 8°57.80' 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 4 shows the predictions carried out with success (in absolute values and %) and the erroneous predictions differentiating the not detected blooms from the false alarms.

Figure 5 shows the absolute value of the difference between the actual concentration of pseudo-nitzschia spp and the forecast value obtained using the hybrid system.

As it indicates, the combination of different techniques in the form of the hybrid CBR system previously presented, produces better results than a RBF neural network alone. 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.

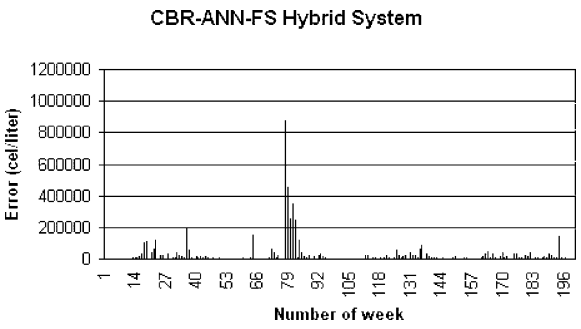

Figure 5. Absolute value of the error using a CBR-ANN-FS hybrid system.

Table 5. Summary of result using a RFB.

OK	OK (%)	Not detect.	False alarms
185/200	92.5	8	7

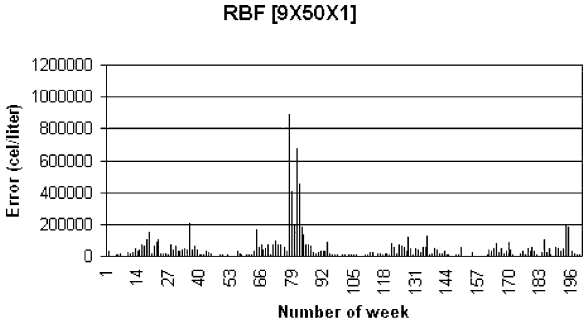

Figure 6. Absolute value of the error using a radial basis function network.

Table 5 shows the same information as the table above but with a RBF neural network. The best results were obtained with a configuration of 50 neurons in the hidden layer, maintaining the input layer (with 18 neurons) and output layer (with 1 neuron) output constant.

Figure 6 also shows the absolute value of the error with the RBF network. 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 6.

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

Table 6. Summary of results using statistical techniques.

Method	OK	OK (%)	Not detect.	False alarms
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 7. 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	70,354.35
Moving average	Statistics	51,969.43
Simple exp. smoothing	Statistics	41,943.26
Brown's Lin. exp. smooth.	Statistics	49,038.19

algorithms and the application of several neural networks methods. The results obtained from these experiments are listed in Table 6.

Table 6 shows the number of successful predictions (in absolute value and %) as well as the blooms not detected and false alarms for each method. From Table 6, it can be calculated that the forecasting error generated by the hybrid system is less than 37.9% of the corresponding value produced by the best statistical technique. In the same way, it improves, by 42.9% the results generated by the RBF neural network working alone. The hybrid system is more accurate than any of the other techniques studied during this investigation. The performance of the hybrid system is better than the other methods at each of the individual geographical monitoring points.

Table 7 shows the average error obtained with the hybrid model (Fig. 5), a standard RBF network (Fig. 6), an ARIMA model (Fig. 7), a Quadratic Trend (Fig. 8), a Moving Average (Fig. 9), a Simple Exp. Smoothing (Fig. 10), a Brown's Linear Exp. Smoothing (Fig. 11)

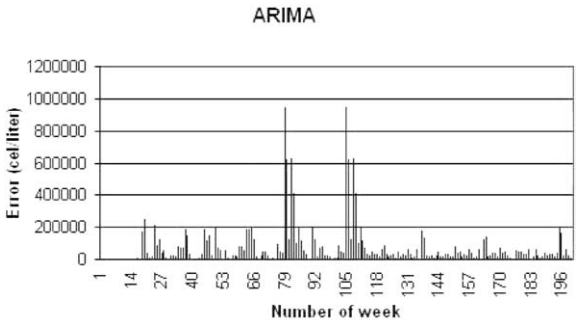

Figure 7. Absolute value of the error using ARIMA (1,0,0) × (0,1,0)26.

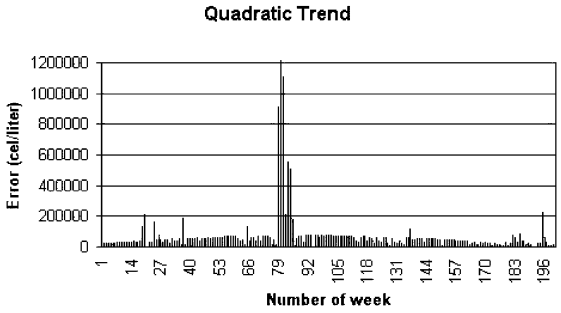

Figure 8. Absolute value of the error using quadratic trend (-8.407, $44 + 2.369$, $34t - 14$, $839t^2$).

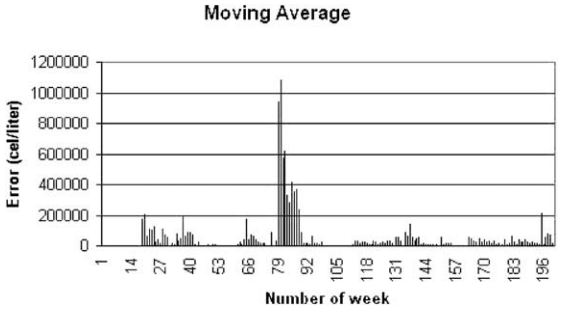

Figure 9. Absolute value of the error using moving average (order = 4).

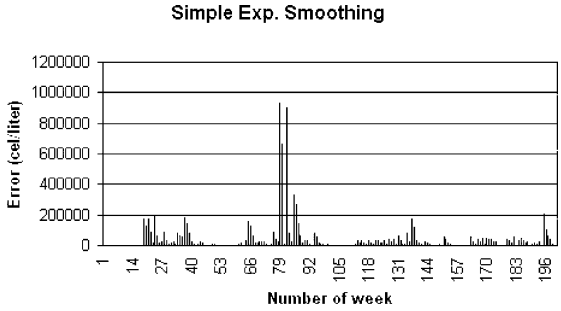

Figure 10. Absolute value of the error using simple exponential smoothing ($a = 0, 704$).

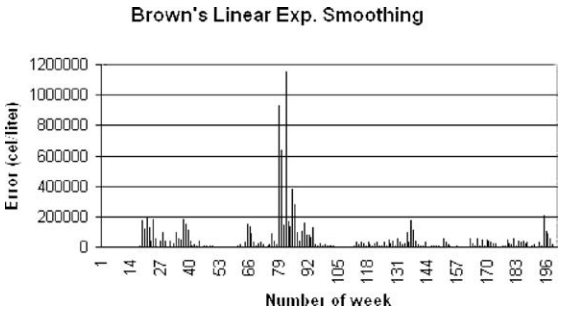

Figure 11. Absolute value of the error using Brown's Linear exponential smoothing ($a = 0, 3613$).

Table 8. Number of predictions with an error ≥100,000 cell/liter.

Method	Incorrect predictions
CBR-ANN-FS	12
RBF	17
ARIMA	38
Quadratic trend	14
Moving average	20
Simple exp. smoothing	18
Brown's Lin. exp. smooth.	25

and a Finite Impulse Response ANN [3], which was not able to converge for this type of problem.

Table 8 shows the number of predictions with an absolute error greater than 100,000 cell/liter. As it clearly shows, the hybrid system, once again, provides the best results.

Figures 7 to 11 show the absolute value of the difference between the actual concentration of pseudonitzschia spp and the forecast value.

6. Conclusion

In summary, this paper has presented a problem-solving method 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 forecasting system is able to produce a forecast with an acceptable degree of accuracy. Although the accuracy of the forecast depends, to a great extent, on the quality of the cases and the geographical monitoring point, it is believed that good quality forecasts may be obtained even with data collected several years before and belonging to other geographical points.

The method employs a case-based reasoning to wrap 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 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 system cannot 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 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 the different sources, mentioned previously. 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.

This work is financed by the project: *Development of techniques for the automatic prediction of the proliferation of red tides in the Galician coasts, PGIDT-00MAR30104PR*, inside the Marine Program of investigation of Xunta de Galicia. The authors want to thank the support lent by this institution, as well as the data facilitated by the CCCMM.

References

1. M. Tomczak and J.S. Godfrey, *Regional Oceanographic: An Introduction*, Pergamon: New York, 1994.
2. E. Fernández, "Las Mareas Rojas en las Rías Gallegas," Technical Report, Departamento de Ecología y Biología Animal, Universidad de Vigo, 1998.
3. J.M. Corchado and C. Fyfe. "Unsupervised neural network for temperature forecasting," *Artificial Intelligence in Engineering*, vol. 13, no. 4, pp. 351–357, 1999.
4. J.M. Corchado and B. Lees, "A hybrid case-based model for forecasting," *Applied Artificial Intelligence*, vol. 15, no. 2, pp. 105–127, 2001.
5. J.M. Corchado, B. Lees, and J. Aiken, "Hybrid instance-based system for predicting ocean temperatures," *International Journal of Computational Intelligence and Applications*, vol. 1, no. 1, pp. 35–52, 2001.
6. J.M. Corchado, J. Aiken, and N. Rees, *Artificial Intelligence Models for Oceanographic Forecasting*, Plymouth Marine Laboratory: UK, 2001.
7. I. Watson, *Applying Case-based Reasoning: Techniques for Enterprise Systems*, Morgan Kaufmann: San Mateo, CA, 1997.
8. R.C. Schank, *Dynamic Memory*, Cambridge University Press: Cambridge, UK, 1982.
9. G. Klein and L. Whitaker, "Using analogues to predict and plan," in *Proc. Workshop Case-Based Reasoning*, 1988, pp. 224–232.
10. J. Kolodner, *Case-based Reasoning*, Morgan Kaufmann: San Mateo, CA, 1993.
11. S.K. Pal, T.S. Dilon, and D.S. Yeung, "Soft computing in case based reasoning," *International Journal of Intelligent Systems*, Springer-Verlag: London, 2000.
12. C.K. Riesbeck and R.C. Schank, *Inside Case-Based Reasoning*, Lawrence Elrlbaum Ass: Hillsdale, 1989.

13. A. Aamodt and E. Plaza, "Case-based reasoning: Foundational issues, methodological variations, and system approaches," *AICOM*, vol. 7, pp. 39–59, 1994.
14. G. Nakhaeizadeh, "Learning prediction of time series. A theoretical and empirical comparison of CBR with some other approaches," in *Topics in Case-Based Reasoning, First European Workshop, EWCBR'93*, edited by S. Wess, K.D. Althoff, and M.M.Y. Richter, Kaiserslautern, Springer: Berlin, 1994.
15. G.G. Lendaris and A.M. Fraser, "Visual fitting and extrapolation," *Time Series Prediction, Forecasting the Future and Understanding the Past*, edited by A.S. Weigend and N.A. Gershenfield, Addison Wesley, 1994, pp. 35–46.
16. G.P. Lekkas, N.M. Arouris, and L.L. Viras, "Case-based reasoning in environmental monitoring applications," *Artificial Intelligence*, no. 8, pp. 349–376, 1994.
17. H.S. Mcintyre, D.D. Achabal, and C.M. Miller, "Applying case-based reasoning to forecasting retail sales," *Journal of Retailing*, vol. 69, no. 4, pp. 372–398, 1993.
18. R.H. Stottler, "Case-based reasoning for cost and sales prediction," *AI Expert*, pp. 25–33, 1994.
19. J.M. Corchado, B. Lees, C. Fyfe, N. Ress, and J. Aiken, "Neuro-adaptation method for a case based reasoning system," in *IJCNN'98*, Alaska, 1998, pp. 304–312.
20. J.M. Corchado and B. Lees, "Adaptation of cases for case-based forecasting with neural network support," in *Soft Computing in Case Based Reasoning*, edited by S.K. Pal, T.S. Dilon, and D.S. Yeung, Springer-Verlag: London, 2000.
21. S. Fraga, D.M. Anderson, I. Bravo, B. Reguera, K.A. Steidinger, and C.M. Yetsch, "Influence of upwelling relaxation on dinoflagellates and shellfish toxity in Ria de Vigo," *Est. Coast and Shelf Sci.*, no. 27, pp. 349–361, 1988.
22. G.M. Hallegraeff, "A review of harmful algal blooms and their apparent global increase," *Phycologia*, no. 32, pp. 79–99, 1993.
23. D.M. Anderson, "Toxic algal blooms and red tides: A global perspective," in *RedTides: Biology Environmental Science and Toxicology*, edited by T. Okaichi, D.M. Anderson, and T. Nemoto, Elsevier: New York, 1989, pp. 11–16.
24. D. Kamykowski, "The simulation of a southern California red tide using characteristics of a simultaneously-measured internal wave field," *Ecol. Model.*, vol. 12, pp. 253–265, 1981.
25. M. Watanabe and A. Harashima, "Interaction between motile phytoplankton and Langmuir circulation," *Ecol. Model.*, vol. 31, pp. 175–183, 1986.
26. P.J.S. Franks and D.M. Anderson, "Toxic phytoplankton blooms in the southwestern Gulf of Maine: Testing hypotheses of physical control using historical data," *Marine Biology*, vol. 112, pp. 165–174, 1992.
27. B. Fritzke, "Growing self-organizing networks—why?," in *ESANN'96*, Brussels, 1996, pp. 61–72.
28. B. Fritzke, "Fast learning with incremental RBF networks," *Neural Processing Letters*, vol. 1, no. 1, pp. 2–5, 1994.
29. Y. Jin, W. von Seelen, and B. Sendhoff, "Extracting interpretable fuzzy rules from RBF neural networks," Technical Report, Institut für Neuroinformatik. Ruhr-Universität Bochum, Jan. 2000.
30. B. Fritzke, "Growing cell structures—A self-organizing network for unsupervised and supervised learning," Technical Report, International Computer Science Institute, Berkeley, 1993.
31. T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 15, pp. 116–132, 1985.

F. Fdez-Riverola Ph.D. from the University of Vigo, Spain. He was born in Langen-Hessen, Germany, in 1973. He works as an Associate Professor for the Department of Computer Science of the University of Vigo, collaborating as investigator with the Group of Intelligent Computer Systems of the University of Salamanca and the research group SING (Computer Systems of New Generation) belonging to the University of Vigo. Talking about his investigation field, he is cantered in the study of hybrid methods of Artificial Intelligence and their application to real problems, although he has also worked in topics related with the development of communication protocols for wireless networks. He is joint author of several books and book's chapters, as well as author of different articles published by well-known editorials like Springer-Verlag, Ios Press, Kluwer, etc. (http://trevinca.ei.uvigo.es/∼riverola/).

Juan M. Corchado received a Ph.D. in Computer Science from the University of Salamanca in 1998 and a Ph.D. in Artificial Intelligence from the University of Paisley (UK) in 2000. At present he is Profesor Titular de Universidad at the University of Salamanca (Spain) and Director of the Postgraduate Programs at the Computing and Control Department, previously he was Sub-director of the Escuela Superior de Ingeniería Informática of the University of Vigo (Spain, 1998–00) and Researcher at the University of Paisley (UK, 1995–98). He has been a research collaborator with the Plymouth Marine Laboratory (UK) since 1993, working in the application of artificial intelligence models to oceanographic problems. He has worked on several Artificial Intelligence (AI) Research projects sponsored by Spanish and European public and private Institutions. He is the coauthor of over 100 books, book chapters, journal papers, technical reports, etc. published by organisations such us IEEE, IEE, ACM, AAAI, Kluwer Academic Publishers, Springer Verlag, Elsevier, Morgan Kaufmann, etc., most of these present practical and theoretical achievements of Hybrid AI and Distributed Systems. (http://gsii.usal.es/).