INTERNATIONAL JOURNAL OF COMPUTATIONAL INTELLIGENCE AND APPLICATIONS, IMPERIAL COLLEGE, LONDON, UK

Hybrid Instance-based System for Predicting Ocean Temperatures

Juan M. Corchado
Artificial Intelligence Research Group
Department of Languages and Computing Systems
University of Vigo, Campus Universitario, 32004, Ourense, Spain
email: corchado@ei.uvigo.es

Brian Lees
Applied Computational Intelligence Research Unit
University of Paisley, Paisley, PA1 2BE, U.K.
email: lees-ci0@paisley.ac.uk

Jim Aiken
Plymouth Marine Laboratory
Prospect Place, West Hoe
Plymouth, PL1 - 3DH
email: JA@wpo.nerc.ac.uk

Abstract

An instance-based problem solving model is presented in which the aim is to forecast, in real time, the physical parameter values of a complex and dynamic environment: the ocean. 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 an instance-based reasoning system can provide a more effective means of performing such predictions than other connectionist or symbolic techniques. The instance-based reasoning system incorporates a radial basis function artificial neural network for the instance adaptation. The results obtained from experiments, in which the system operated in real time in the oceanographic environment, are presented.

Key words: instance-case reasoning, neural adaptation, forecasting, real-time modelling

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 artificial intelligence (AI) model to a real time forecasting problem. 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 (Tomczak *et al*., 1994). Although some statistical models have been formulated to describe partial oceanographic water masses, there are, as yet, no accurate general models. It is well known that the behaviour and characteristics of the oceans change seasonally and spatially. However, current knowledge of the ocean structure is still too weak to create a comprehensive model. Ocean water masses are extremely heterogeneous; each water mass has certain properties that differentiate it from other water masses.

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 (Lees *et al*., 1992). Several types of standard artificial neural network (ANN) have been used to forecast time series (Corchado *et al*., 1997a, 1997b, 1998, 1999), in these experiments it has been discovered that it is very difficult to train a neural network to forecast successfully over the whole time series, especially if the data relate to a dynamic system. Statistical models such as Auto-Regressive Integrated Moving Averages (ARIMA) have been applied, but the results so far obtained have indicated that neural networks (although they are not enough accurate) have a greater facility for forecasting, using this type of time series data, than statistical models.

Following earlier experiments using case-based reasoning (CBR) as a problem solving strategy in other domains (Lees, 1997; Adams *et al*., 1997), an instance-based approach to the forecasting problem was considered worthy of investigation. Consequently a prototype system based on this approach was developed in the belief that an instance-based reasoning (IBR) mechanism might, as a data mining strategy, make better use of the vast database of oceanographic data held at the Plymouth Marine Laboratory (PML). The results of subsequent experiments employing an IBR approach indicated that instance or case-based reasoning methods could facilitate the organisation of data, the recovery of relevant data necessary to make an accurate forecast and incremental system learning. The adaptation of the recovered data is a crucial factor in obtaining an accurate result. With the aim of providing a more effective instance or case adaptation procedure it was decided to investigate a hybrid systems approach in which a neural network would be employed as a means of instance adaptation. As a result, a *Radial Basis*

Function (RBF) network has been found to be effective in the instance/case adaptation stage of the IBR system.

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 on the surface of the ocean, and at any time of the year without human intervention. The results obtained to date suggest that the approach to be described in this paper appears to fulfil this aim. The structure of the paper is as follows. First a brief overview of the basic concepts of case-based reasoning is given. Then the oceanographic problem domain is briefly outlined. The hybrid neural network enhanced IBR system is then explained, and, finally, an outline of some of the results obtained to date is presented.

2. CBR Systems Overview

Although knowledge-based systems (KBS) represent one of the commercial successes of the outcome of artificial intelligence research, developers of these systems have encountered several problems (Watson *et al.*, 1994). 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, may also be difficult to maintain. With the aim of overcoming these problems Schank (1982) proposed a revolutionary approach, *case-based reasoning*, which is in fact a model of human reasoning (Joh, 1997). 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 (Klein *et al.*, 1988). 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 (Kolodner, 1993).

A case-based reasoning system solves new problems by adapting solutions that were used to solve previous problems (Riesbeck *et al.*, 1989). 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, in which human interaction may possibly be needed. Case-base 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 that a new problem is to be solved (Kolodner, 1993; Aamodt, 1994; Watson, 1997). Figure 1 outlines the basic CBR cycle.

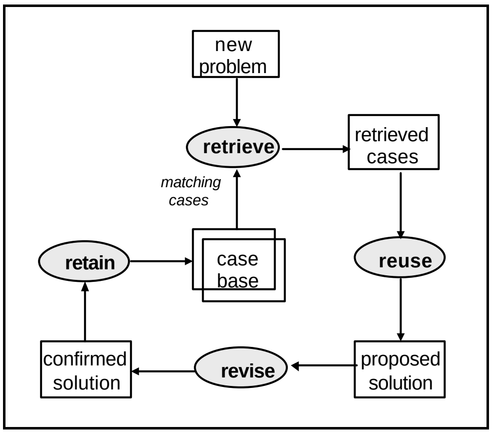

Figure 1. CBR Cycle

This cyclic process of CBR involves four major steps, represented by the ellipses in Figure 1:

- ***Retrieve*** the most relevant case(s),
- ***Reuse*** the case(s) to attempt to solve the problem,
- ***Revise*** the proposed solution if necessary, and
- ***Retain*** the new solution as a part of a new case.

The purpose of the retrieval step is to search the case base and to select from it 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 remembered 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 that 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 of automating the process of case adaptation (revision) is presented for the solution of problems in which the cases are characterised predominantly by numerical information.

The Instance-based reasoning systems are highly syntactic CBR-approaches (Aamodt *et al*., 1994). In cases where there is a lack of guidance from general background knowledge, a relatively large number of instances are needed in order to obtain a concept definition or solution. The representation of the instances are usually simple (e.g. feature vectors), since a major focus is to study automated learning without user intervention (Aha, 1991).

2.1 CBR Systems for Forecasting

Several researchers (Nakhaeizadeh, 1994; Lendaris *et al*., 1994) 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 (1994) uses a relatively complex hybrid CBR-ANN system. In contrast, Lendaris and Fraser (1994) forecast a data set just by searching in a given sequence of data values for segments that closely match the pattern of the last *n* measurements and then by supposing that similar antecedent segments are likely to be followed by similar consequent segments.

In most of the cases the CBR systems used in forecasting problems have a flat memories with simple data representation structures. 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 of a small case-base. A survey of such forecasting CBR systems can be found in Corchado *et al*., 1998.

3. The Oceanographic Environment

Oceanography is a science that concentrates on the understanding of the physical principles that drive the oceans, and which uses the tools of mathematics and theoretical fluid dynamics to forecast their behaviour (Tomczak *et al*. 1994). Oceanic waters are divided into provinces (also called water masses) which are moderately homogenous. The boundaries between provinces or water masses are known as *fronts*. These areas are very dynamic and their properties depend on the nature of the two water masses converging to create each front. Some frontal areas, e.g. the Arctic and Antarctic convergence zones, are extremely heterogeneous and are very variable; therefore forecasting the temperature of the water in these areas can be difficult.

The movement of water masses makes the ocean's water temperature change in a complex manner in both spatial and temporal domains (Tomczak *et al*. 1994). The analysis and interpretation of large volumes of oceanographic data have traditionally been achieved through the use of statistical software tools. However the processing speed when using such tools is limited by the need for frequent user intervention.

Knowledge-based systems have been developed to assist in weather prediction. However, apart from the work of Lybanon (1986), there appears to be little evidence of the application of knowledge based approach as a means of predicting the location and movements of large water masses.

With the aim of providing an improved method for analysing the large masses of available oceanographic data, the application of knowledge-based (in particular, rule-based) methods was investigated in an earlier research project involving collaboration between the Plymouth Marine Laboratory and the University of Paisley (Lees *et al*., 1992). The motivation for that work was derived from the oceanographers' need for a better understanding of the ocean environment, for which, because of the complexities of the ocean, it is very difficult to formulate adequate and complete mathematical models. The results, using rule-based methods, provided some insight into the nature and complexities of the ocean. Building on this experience, it was decided to carry out further investigations into the application of alternative artificial intelligence problem solving approaches. Detecting oceanographic features, their boundaries and being able to predict their evolution was the goal of the *Knowledge Based Oceanographic System* (KBOS) (Rees *et al*., 1991) and subsequent projects: the *On line Real time Knowledge based Analysis* (ORKA) system (Rees *et al*., 1995), and the *Simulated Tactical Environmental Bubble* (STEB) system (Corchado *et al*., 1997a, 1997b). The AI methods which were applied in these projects include neural networks (Corchado, 1995) and case-based reasoning (Lees, 1997).

The focus in recent research has been to investigate ways to forecast the thermal structure of the water ahead of an ongoing vessel, in real time. Forecasting in such an environment is a difficult task due to the nature and behaviour of the ocean waters, which are in a continuous state of movement. The scales of physical motion of the oceans and the atmosphere range from being ocean-wide down to tiny eddies, which are present in the neighbourhood of fronts.

In order to obtain acceptable predictions an autonomous universal methodology, capable of forecasting changes in the water temperature as an expert oceanographer might do, is desirable. In addition, the system should be able to analyse, in real time, the variation in the temperature of the water on a local basis, to analyse and select the most relevant local knowledge from the huge database of information available (in the form of satellite images and thermal data profiles) to produce a forecast, taking into account factors such as the season of the year and the geographical location of the vessel from which the forecast is made.

4. Forecasting Systems

In the current work the aim is to develop a system for forecasting as a methodology for predicting the values of physical parameters (in particular, sea temperature) at a given depth around a sea going vessel from data acquired in real time, and also from past records of sea temperature (and possibly other oceanographic parameters) surrounding the vessel at some point ahead which will be reached in the immediate future. This information may also then be used to provide a forewarning of an impending oceanographic front. The approach builds on the methods and expertise previously developed in the earlier research referred to above.

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

Given: a sequence of data values (which may be obtained either in real-time, or from stored records) relating to some physical parameters,

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

The raw data (on sea temperature, salinity, density and other physical characteristics of the ocean) which are measured in real time by sensors located on the vessel consist of a number of discrete sampled values of a parameter in the form of a time series. These data values are supplemented by additional data derived from satellite images, which are received weekly. In the present work the parameter used is the temperature of the water mass at a fixed depth. Values are sampled along a single horizontal dimension, thus forming a set of data points.

4.1 The Hybrid IBR Forecasting System

In order to produce a forecast, in real-time, of ocean temperature a certain distance ahead of a vessel as it traverses the ocean, a *problem instance* is generated every 2 km. A problem instance consists of a sequence of the *N* sampled data values (after suitable filtering and pre-processing) immediately preceding the data value corresponding to the current position of the vessel. A value of 40 for *N* (corresponding to 40 km) has been found empirically to produce satisfactory results when forecasting the temperature of the water 5 km ahead of an ongoing vessel. The problem instance also includes various other numerical values, including the current geographical location of the vessel and the time and date when the case was recorded. To forecast at other distances, different values of *N* are required.

The set of *N* data values forms an *input vector*, which is then used to produce a forecast of the ocean temperature, several km ahead. In outline, this process is depicted in Figure 2. Note that, in practice, it is the set of differences ($T_i - T_{i-1}$, $T_i - T_{i-2}$ etc.) between the temperature T_i at the current point and the temperature at successive earlier points which is used as the input vector.

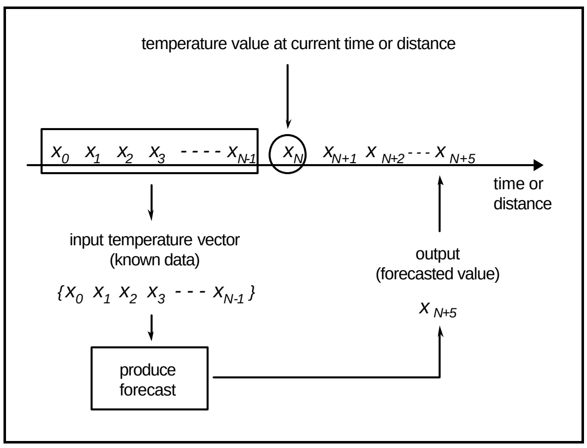

Figure 2. Forecasting operation

The forecasted values are created using a neural network enhanced instance-base reasoning system. The IBR mechanism allows the experience recorded in previous forecasting situations to be reused. The role of the neural network lies in the instance adaptation process. The relationships between the processes and components of the hybrid system are illustrated in Figure 3.

The cyclic IBR process shown in the figure has been inspired by work of Aamodt and Plaza (1994). The four basic phases in the IBR cycle are shown as ellipses. Superimposed on the fundamental IBR cycle is a cycle of neural network operations during which the network parameters are retrieved from a neural network knowledge base, employed in instance adaptation, and then are revised, with their updated values being stored back in the knowledge base. The full cycle of operations of the hybrid system is explained in the following section.

The particular type of neural network of interest in the current research is the *Radial Basis Function* (Haykin, 1994), 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.

The hidden neurons form a basis for the input vectors; the output neurons merely calculate a linear combination of the hidden neurons' outputs. Activation is fed forward from the input layer to the hidden layer where a Basis Function is calculated. The weighted sum of the hidden neurons' activations is calculated at the single output neuron.

Radial Basis Functions (RBF) are better at interpolating that at extrapolating. Furthermore, RBFs are less sensitive to the order in which data is presented to them than is the case with other neural network models, such as Multi-Layer Perceptrons. Radial Basis Functions are of potential use in hybrid systems because of their fast learning capability.

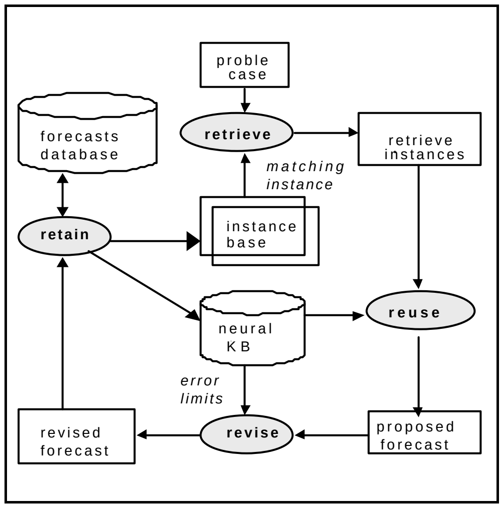

Figure 3. Modified instance/case-base reasoning cycle

4.2 IBR System Operation

The forecasting system uses data from two sources: (i) the real-time data are used to create a succession of problem instances, characterising the current forecasting situation; (ii) data derived from satellite images are stored in a database (which, for clarity, is not shown in Figure 3). The satellite image data values are used to generate instances, which are then stored in the instance base and subsequently updated during the IBR operation.

The cycle of forecasting operations (which is repeated every 2 km) proceeds as follows. First a new problem instance is created from the pre-processed real-time data. A set of *k* instances, which most closely matches this current problem instance, is then obtained from the instance base during the IBR *retrieve* phase, using nearest neighbour matching.

In the *reuse* phase, the values of the weights and centres of the neural network used in the previous forecast are retrieved from the neural network knowledge base. These network parameters together with the *k* closest matching instances are then used to create a forecast of the temperature a distance 5 km, say, ahead. At this point the parameters of the network are modified by taking into account the information contained in the retrieved instances. The effect of this is to allow the system to learn from all these *k* instances (rather than simply using the single adjudged closest matching instance) in making a new forecast.

During each forecasting cycle the RBF network is retrained, using the retrieved weights and centres, with the input vectors contained in the *k* matching instances applied as inputs to the network. This process adapts the network, by accommodating the retrieved instances, thus updating the values of the network parameters (empirically, a value for *k* of between 500 to 1000 has been found to be appropriate). The input vector from the problem instance is then fed into the trained network to produce a proposed forecast.

In the *revise* phase, the proposed forecast is modified by taking into account the accuracy of the previous forecasts, which were reused in obtaining the new forecast. Each instance has associated with it a measure of the average error over the previous forecasts for which that particular instance was used to train the neural network. Error limits are calculated by averaging the average error of the *k* instances used to train the ANN in producing the current forecast. The revised forecast is then expressed, using the error limits, as an interval between upper and lower limits rather than as a single value.

The revised forecast is then retained in a temporary store – the forecasts database. When the vessel has travelled a further 5 km the actual value of the water temperature at that point is measured. The forecasted value for the temperature at this point can then be evaluated, by comparison of the actual and forecasted values, and the error obtained. A new instance, corresponding to this forecasting operation, is then entered in the instance base. Knowledge of the forecasting error is also, at this point, used to update the average error of all the *k* instances that were reused to obtain that forecast.

4.3 Radial Basis Function Operation

The RBF network uses nine input neurons, between twenty and thirty five neurons in the hidden layer and a single neuron in the output layer. Input vectors (explained earlier) form the input to the network; the output of the network is the difference between the temperature at the present point and the temperature a fixed distance ahead. Initially, twenty vectors are randomly chosen from the first training data set and used as centres in the middle layer of the RBF network. All the centres are associated with a Gaussian function, the width of which, for all the functions, is set to the mean value of the Euclidean distance between the two centres that are separated the most from each other.

Training of the network is done by presenting pairs of corresponding input and desired output vectors. After an input vector activates every 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 centre to each particular input vector is moved toward the input vector by a percentage α of the present distance between them. By using this technique the centres are positioned close to the highest densities of the input vector data set. The aim of this adaptation is to force the centres to be as close as possible to as many vectors from the input space as possible. The value of α is initialised to a value of 20 each time that the network is retrained, and its value is linearly decreased with the number of iterations until its value becomes zero; then the network is trained for a number of iterations (between 10 and 30 iterations for the whole training data set, depending on the time left for the training) in order to obtain the best possible weights for the final value of the centres.

A new centre is inserted into the network when the average error in the training data set does not fall by more than 10% after 10 iterations (using the whole training set). In order to determine the most distant centre *C*, the Euclidean distance between each centre and each input vector is calculated and the centre whose distance from the input data vectors is largest is chosen. A new centre is inserted between *C* and the centre closest to it. Centres are also eliminated when they do not contribute significantly to the output of the neural network. Thus, a neuron is eliminated if the absolute value of the weight associated with that neuron is smaller than twenty per cent of the average value of the absolute value of the five smallest weights. The number of neurons in the middle layer is maintained above 20.

5. Results

The hybrid forecasting system has been tested in the Atlantic Ocean in September 1997 on a research cruise going from the UK to the Falkland Islands (also known as Malvinas). The cruise crossed several water masses and oceanographic fronts (areas where two water masses with different characteristics converge). Although, the system was monitored and improved from a computing point of view, over a month, the forecasting method explained in previous sections remained unchanged. The prototype system used in this experiment was set up to forecast the temperature 5 km ahead. Figure 4 illustrates the error in the forecasts over a total distance traversed of 10500 km. The strategy adopted was to create an accurate successful method which was able to forecast a short distance ahead, and then to extend it so as to produce forecasts a further distance ahead.

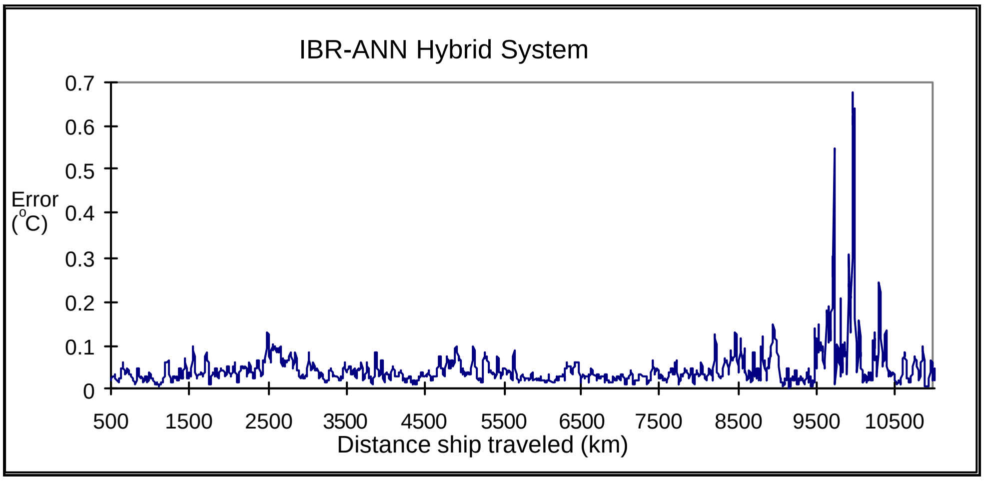

Figure 4. Absolute value of the error using the hybrid system

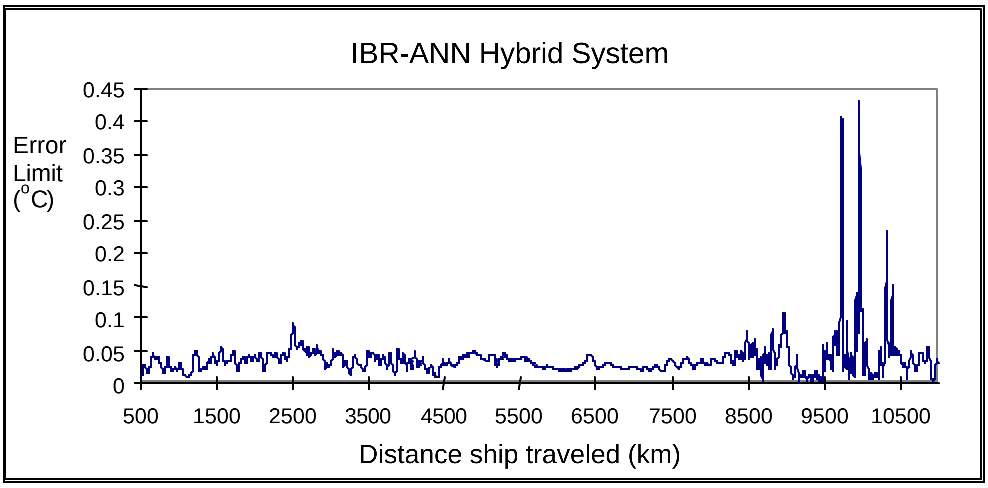

Figure 5. Error limit values for a forecast 5 km ahead using the hybrid system

The average error in the forecast was found to be 0.02 °C. Only 4.5% of the forecasts have an error higher than 0.5 °C, 8.3% higher than 0.04 °C, 32% higher than 0.02 °C. These figures indicate that the hybrid system is able to produce a forecast with an average error of 0.02 °C and with a probability of 0.96 that the error in the forecast is smaller than 0.05 °C. Although the experiment was carried out using a limited data set (11000 km between the latitudes 50° North and 50° South), eleven water masses with different characteristics were crossed, six fronts were traversed. The Falkland Front (km 10000) in particular is one of the most chaotic oceanographic areas in the world. It is believed that these error value results are significant enough to be extrapolated over the whole Atlantic Ocean.

Figure 4 shows the absolute value of the difference between the actual temperature value of the water and the forecast value obtained using the RBF neural network for the instance adaptation in the hybrid system. This graph does not take into account any improvement that may be obtained using error limits during the review phase of the CBR cycle.

The use of error limits can substantially improve the accuracy of the forecast. The error limit values are determined and modified dynamically from information relating to the past forecasting performance of the system and which is contained in the stored instances. These error limits indicate the range of forecast values than may be expected to be produced through the adaptation of particular stored instances.

Figure 5 shows the value of the error limits used during the experiment and Figure 6 the *forecast error outside the error limits*. (If, using error limits, the forecasted temperature value at a particular location is, say, 6 ± 0.5 °C - i.e. from 5.5 °C to 6.5 °C - then, if the actual temperature value is found to be 6.8 °C, the value of the forecast error outside the error limits will be 6.8 - 6.5, i.e. 0.3 °C).

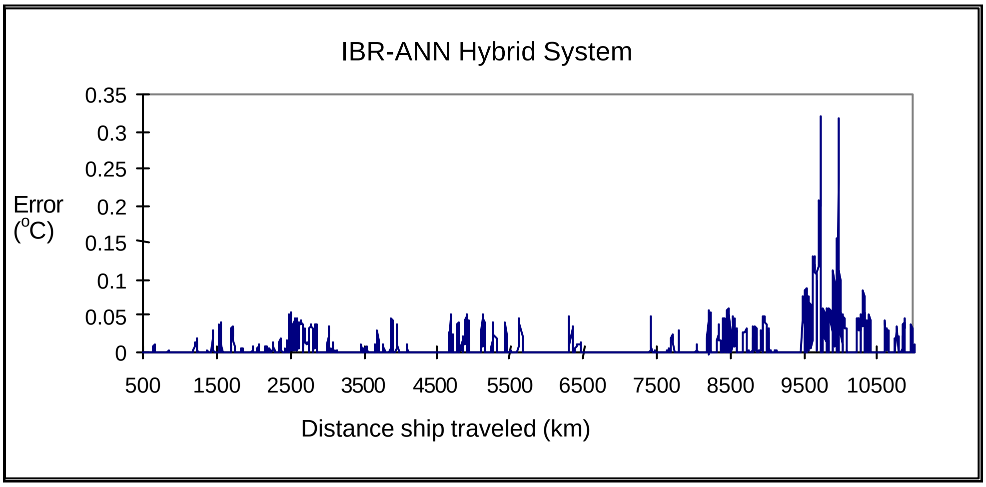

Figure 6. Absolute error values using the hybrid system (with error limits)

Time between the picture being recorded and the real time data being recorded (number of weeks)	Average Error (°C)	Error Limits (°C)	Average error (outside error limits) (°C)
1	0.020	0.0198	0.00100
2	0.024	0.0255	0.00173
3	0.034	0.0332	0.00212
4	0.048	0.0459	0.00333
52	0.033	0.0318	0.00252

Table 1. Average error in the forecast outside the error limits with the hybrid system

Although 45.5% of the predictions were outside the limits of the error band, only 3.4% of the predictions were more than 0.005 °C outside the error limits. The average error of the predictions using error limits is 0.001 °C. The shape of the error limit plot (Figure 6) is very similar to the error in the forecast presented in Figure 4; this means that the error limits adapt themselves to the pattern of temperatures in the different water masses. Both the error and the error limits are, on average, higher in frontal water masses than in homogeneous water masses. This was to be expected due to the higher dynamic nature and heterogeneity of such areas.

A similar experiment was carried out using the data recorded by the vessel during the cruise, but this time using instances obtained from satellite images recorded more than one week previously. Table 1 shows how the average error, the average error limits and the average error outside the error limits were found to vary when satellite images of different ages were used. Table 1 shows that the forecasting error is only slightly changed when using satellite images which are one or even two weeks old. The table also shows that when using satellite pictures collected exactly one year back the error in the forecast may be similar or smaller than the error obtained using pictures that are three or more weeks old. This is the reason why data up to one year old is kept in the database and in the instance base; if for technical reasons (e.g. clouds covering a certain area or problems with the data telecommunications) recent satellite images can not be obtained, data recorded one year back can be used by the system and may in fact produce better results that data recorded three or four weeks previously. This is due to the annual cycle of most of the water masses. However, such results can not be guaranteed as there are also other factors that determine the pattern of ocean temperature variations.

Further experiments have been carried out to compare the performance of the IBR-ANN hybrid forecasting system with several other forecasting approaches. These include standard statistical forecasting algorithms and the application of several neural network methods. The results obtained from these experiments are listed in Table 2.

The table shows the average error obtained with a Finite Impulse Response ANN (Corchado *et al*., 1999), a standard Radial Basis Function network (Corchado *et al*., 1997a), a Linear Regression model, an ARIMA (Auto-regressive Integrated Moving Averages) model (Box *et al*., 1976) and a CBR system without the neural network component which creates instances in real time using the temperature recorded from the sea surface (Rees *et al*., 1997).

Table 2 shows how the forecasting error generated by the hybrid system is less than 20% of the corresponding value produced by any of the other forecasting methods. 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 in each of the individual water masses crossed by the vessel when travelling from the UK to the Falkland Islands. For a complete descriptions of the results obtained in the framework of this investigation refer to Corchado (2000).

6. Conclusions

This paper has presented a problem solving method that combines an instance/case-based reasoning system integrated with an artificial neural network, which is able to identify trends present in a large data set in order to create forecasts in real time. As such, it could be regarded as an application of data mining in a real time situation. The forecasting task, which is addressed in this work, is difficult for two reasons: the unpredictable and complexity of the media in which the forecast must to be done, and the fact that the forecast must be done in real time. Essentially, the instances stored in the instance base are created through the selection of data values, which correspond to the current location, and track of the ship. The time series data values obtained in real time are then matched by the IBR mechanism against the data patterns of the stored instances in order to produce the required forecast. The data are derived from the extensive database of earlier historical values, or from satellite images. The evaluation of the forecasts produced and the consequent modifications of the stored instances enable the system to learn and to improve its performance over time.

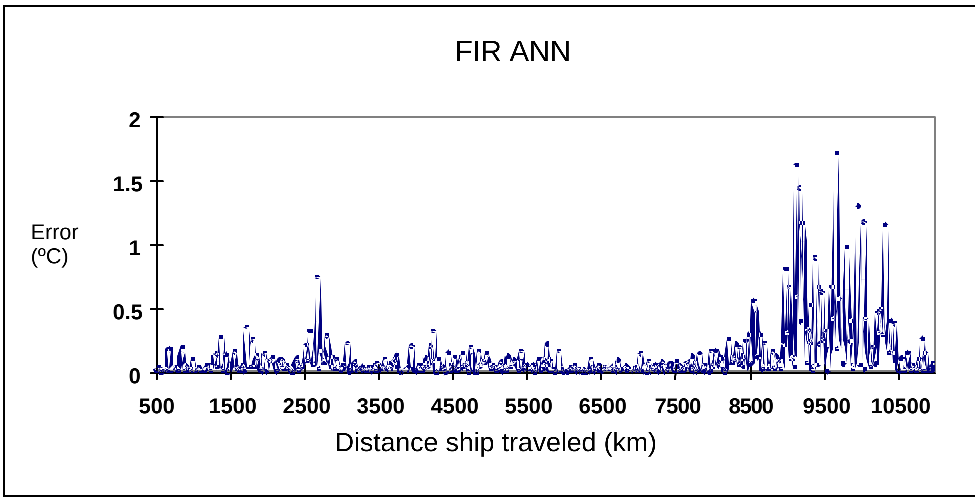

Table 7. Absolute value of the error using the FIR ANN

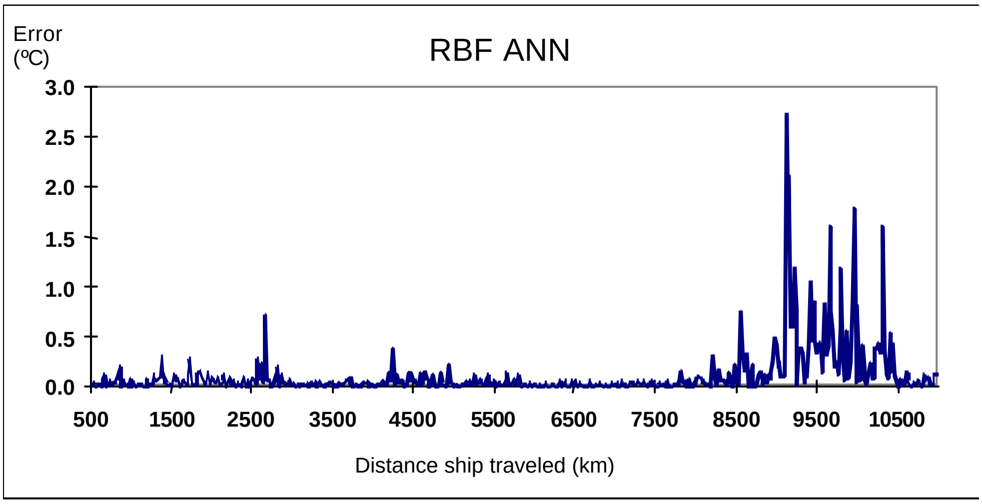

Table 8. Absolute value of the error using the RBF ANN

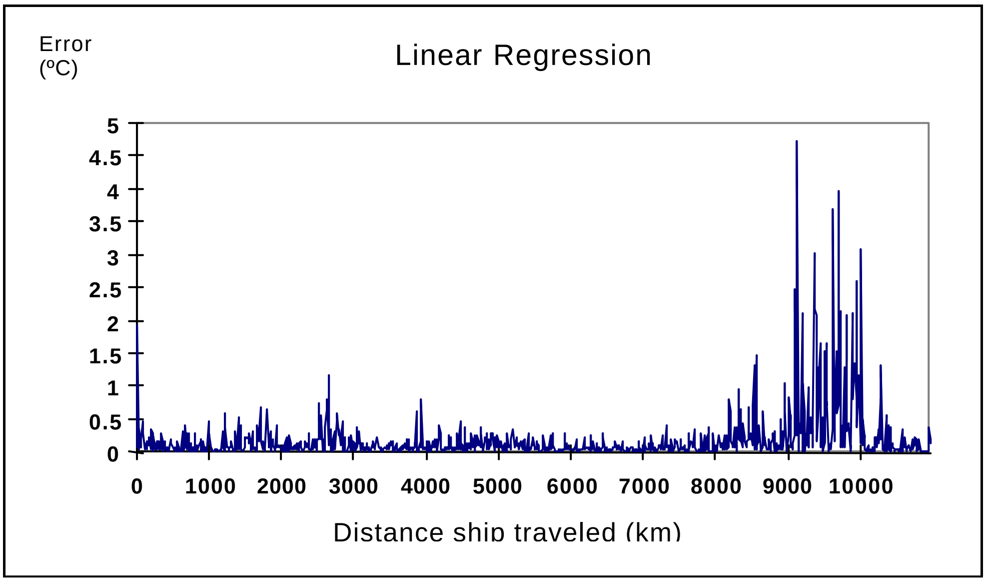

Table 9. Absolute value of the error using the Linear Regression

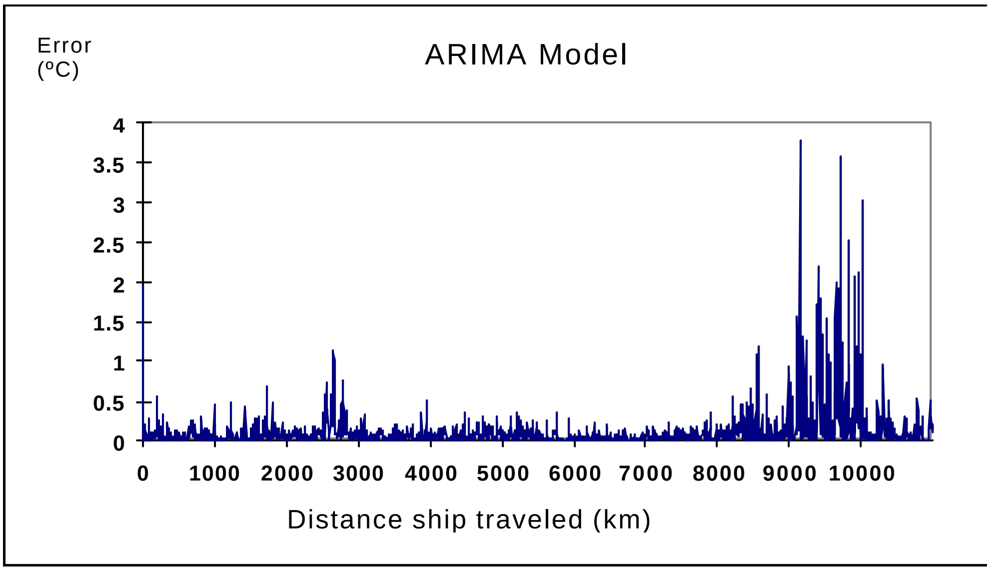

Table 10. Absolute value of the error using the ARIMA model

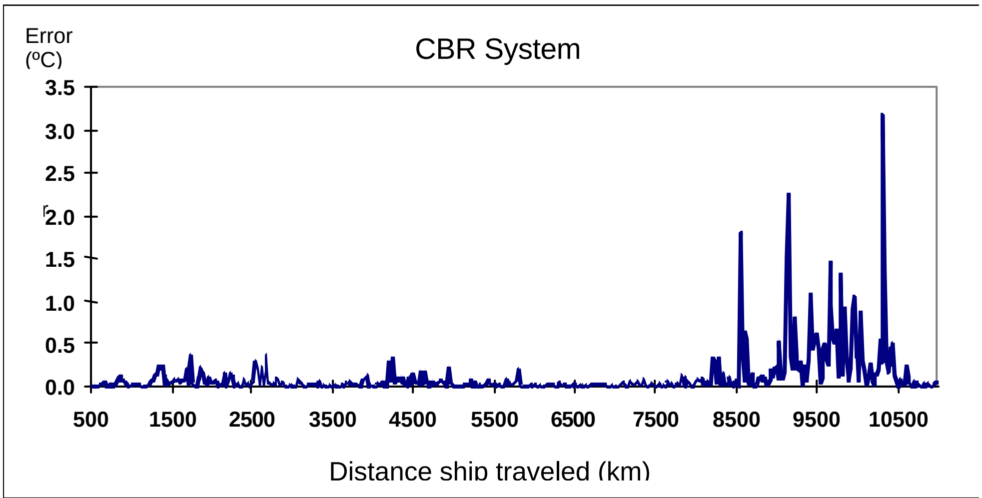

Table 11. Absolute value of the error using a CBR system

Method	Type	Average Error (°C)
FIR	ANN	0.096
RBF	ANN	0.114
Linear Regression	Statistics	0.174
ARIMA	Statistics	0.129
CBR	CBR	0.120
IBR-ANN hybrid	IBR – ANN	0.020
IBR-ANN hybrid (outside error limits)	IBR – ANN	0.001

Table 2: Average forecasting error using various forecasting methods

The forecasting system is able to produce a forecast with an acceptable degree of accuracy and within the time constraints imposed by the real time nature of the problem. Although the accuracy of the forecast depends, to a great extent, on the quality of the instances and on the actual date when the data from which the instances were created was collected, it has been demonstrated that good quality forecasts may be obtained even with data collected one year before the forecast was made.

The method combines the ability of instance-based reasoning to index, organise and retrieve relevant data with the generalisation, learning and adaptation capabilities of the radial basis function neural network. The resulting hybrid system thus combines complementary properties of both connectionist and symbolic AI methods. The neural network plays an important role in the system; it adapts the instances selected during the instance-based operations, combines aspects of the knowledge contained in several instances and supports the generation of the prediction. The Radial Basis Function network adapts its structure, without human intervention, to the characteristics of the environment in which the system is operating and acts as a function that facilitates system learning by extracting the relevant characteristics of a number of closely matching instances and combining them in the form of a representative instance. 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 limitations of this method of forecasting in its present form are as follows.

- Forecasts can only be produced while the vessel is proceeding in a straight line. The present system is not able to forecast while the vessel is changing its direction (or has changed it in the last 40 km). However, it would be possible to adapt the way in which the data is extracted from the database during the instance creation process, to enable it to overcome this limitation.
- The present system operates satisfactorily only if there are no discontinuities in the data greater than 2 km in length. However, after a discontinuity the forecast can be resumed by interpolating the missing data with data from both ends of the instance vector. This strategy has been found to be successful if the discontinuity is no longer than 5 km.
- The system can not function in a particular area if there are no stored instances from that area. In this situation, the only solution is to use a back-up mechanism to prime the system; for this, the experimental results obtained in comparing neural network methods suggest that a Finite Impulse Response network may be the most appropriate method to use (Corchado *et al*., 1999). Once the system is in operation and is producing forecasts, a succession of instances will be generated, thus enabling the hybrid forecasting mechanism to function autonomously.

In conclusion, the instance-based reasoning problem solving approach provides an effective strategy for forecasting in an environment in which the raw data is derived from three distinct sources: a large database of historical data, satellite image data, and time series data obtained in real-time.

References

Aamodt A. and Plaza E. (1994), Case-Based Reasoning: foundational Issues, Methodological Variations, and System Approaches, AICOM, Vol. 7 No. 1, March 1994, pp. 39-59.

Adams D., Lees B., Irgens C. and MacArthur E. W. (1997). Using case outcome to integrate customer feedback into the quality function deployment process, *Procs 5th German Workshop on Case Based Reasoning - Foundations, Systems and Applications*, eds. R Bergmann and W Wilke, LSA 97-01E, Centre for Learning Systems & Applications, University of Kaiserslautern, March 1997, pp. 1-9.

Aha D., Kibler D., and Albert M. K. (1991). Instance-Based Learning Algorithms. Machine Learning, vol.6, pp 37-66.

Bull, M., Kundt, G. and Gierl, L. (1997). Discovering of Health Risks and Case-Based Forecasting of Epidemics in a Health Surveillance System. Proc. of PKDD97, Springer LNAI 1263, pp. 68-77.

Corchado J. M. (1995). Application of masks, filters and artificial neural networks to oceanographic forecasting. BSc Honours Software Engineering Dissertation, University of Paisley, 1995.

Corchado J. M., Rees N., Lees B. and Fyfe C. (1997a). Study and Comparison of Multi-Layer Perceptron NN and Radial Basis Functions NN in Oceanographic Frontal Warning. Proceedings of Application and Science of Artificial Neural Networks, SPIE's International Symposium an Aerospace/Defense Sensing and Control. April 21-25, Orlando, Florida. pp. 550-560.

Corchado J. M., Lees B.,. Fyfe C. and Rees N. (1997b). Adaptive Agents: Learning from the Past and Forecasting the Future, Proceedings of the First International Conference on Knowledge Discovery and data mining, London, April. pp. 109-123.

Corchado J. M., Lees B., Fyfe C., Rees N., Aiken J.(1998). Neuro-Adaptation Method for a Case Based Reasoning System, IJCNN'98 Anchorage, Alaska, May 5-9. pp 304-312.

Corchado J. M., Fyfe C. (1999) Unsupervised neural method for temperature forecasting, Artificial Intelligence in Engineering, ISSN: 0954-1810, Volume : 13, Issue : 4, 02-Nov-1999, pp 351-357.

Corchado J. M. (2000) Neuro-symbolic Model for Real-time Forecasting Problems. University of Paisley, UK.

Haykin S. (1994) Neural Networks: A Comprehensive Foundation, Macmillan, 1994.

Joh D. Y. (1997). CBR in a Changing Environment. Case Based Reasoning Research and Development. ICCBR-97. Providence, RI, USA. pp 145-156.

Klein G. and Whitaker L. (1988). Using Analogues to Predict and Plan. Proceedings of a Workshop on Case-Based Reasoning, pp. 224-232.

Kolodner J. (1993). Case-Based Reasoning, San Mateo, CA: Morgan Kaufmann, 1993.

Lees B., Rees N. and Aiken J. (1992). Knowledge Based Oceanographic Data Analysis, Procs. EXPERSYS 92, IITT Paris, pp 561-566.

Lees B and Corchado J. (1997). Case based reasoning in a hybrid agent-oriented system, *Procs 5th German Workshop on Case Based Reasoning - Foundations, Systems and Applications,* eds. R Bergmann and W Wilke, LSA 97-01E, Centre for Learning Systems & Applications, University of Kaiserslautern, March 1997, pp. 139-144.

Lees, B. (1997). Engineering Design Support through Case Based Reasoning, *Procs. Colloquium on Intelligent Design Systems*, (Digest No: 97/016), IEE, London, 25th February 1997, pp. 8/1-8/3.

Lendaris G. G. and Fraser A. M. (1994). 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.

Rees N., Aiken J. and Lees B. (1991). Artificial Intelligence Techniques Applied to the Analysis of Oceanographic Data Sets. OSATES, Brest, France, September, 1991. pp. 24-27.

Rees N. and Aiken J. (1995). Detection of Oceanographic Fronts: A knowledge Based. Procs Expert Systems 95, British Computer group on Expert Systems, Cambridge, pp. 135-146.

Rees N., Lees B., and Corchado J. M. (1997). Oceanographic Technical Report II, 1997, PML.

Riesbeck C. K. and Schank R. C. (1989). Inside Case-Based Reasoning. Lawrence Erlbaum Ass, Hillsdale.

Schank R. C. (1982). Dynamic Memory, Cambridge University Press, Cambridge, UK.

Tomczak M. and Godfrey J. S. (1994). Regional Oceanography: An Introduction. Pergamon. New York, 1994.

Watson I. and Marir F. (1994). Case-Based Reasoning: A Review. Cambridge University Press, 1994. The knowledge Engineering Review, Vol 9. No. 3. pp. 34-41.

Watson I. (1997). Applying Case-Based Reasoning: Techniques for Enterprise Systems. Morgan Kaufmann.