A WeVoS-CBR Approach to Oil Spill Problem

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

[1] University of Burgos, Spain
[2] University of Salamanca, Spain
escorchado@ubu.es, bbaruque@ubu.es, corchado@usal.es, aitor@usal.es

Abstract. The hybrid intelligent system presented here, forecasts the presence or not of oil slicks in a certain area of the open sea after an oil spill using Case-Based Reasoning methodology. The proposed CBR includes a novel network for data classification and data retrieval. Such network works as a summarization algorithm for the results of an ensemble of Visualization Induced Self-Organizing Maps. This algorithm, called Weighted Voting Superposition (WeVoS), is mainly aimed to achieve the lowest topographic error in the map. The system uses information obtained from various satellites such as salinity, temperature, pressure, number and area of the slicks. WeVoS-CBR system has been able to accurately predict the presence of oil slicks in the north west of the Galician coast, using historical data.

Keywords: Case-Based Reasoning; Oil Spill; Topology Preserving Maps; Ensemble summarization; Self Organizing Memory; RBF.

1 Introduction

When an oil spill is produced, the natural risks are evident, and complicated decisions must be taken in order to avoid great natural disasters. Predicting if an area is going to be affected by the slicks generated after an oil spill will provide a great aid to take those decisions.

The ocean is a highly variable environment where accurate predictions are difficult to achieve. The complexity of the system is increased if external elements are introduced in the analysis. In this case, oil spills are added, generating a rough set ofelements.

The solution given by the system presented in this paper is a probability of finding oil slicks in a certain area. The proposed system is a forecasting Case-Based Reasoning system [1]. A CBR system has the ability to learn from past situations, and to generate solutions to new problems based in the past solutions given to past problems.

The hybrid system combines the efficiency of the CBR systems with artificial intelligence techniques in order to improve the results and to better generalize from past data.

The developed system has been constructed using historical data and the knowledge generated after the Prestige accident, from November 2002 to April 2003. The WeVoS-ViSOM algorithm [2] is applied for the first time under the frame of a CBR to perform classification tasks, when creating the case base. The algorithm also facilitates the retrieval phase, and makes it faster by properly grouping together those cases that are actually similar.

After explaining the oil spill problem, both the CBR methodology and the WeVOS-ViSOM algorithm are explained. Then, the developed system is described, ending with the results, conclusions and future work.

E. Corchado, A. Abraham, and W. Pedrycz (Eds.): HAIS 2008, LNAI 5271, pp. 378–384, 2008.

2 Description of the Oil Spill Problem

Once an oil spill occurs, the evolution of the oil slicks generated must be supervised or even predicted, in order to know if an area is going to be contaminated or even better, to avoid contamination in some critical areas. To get an accurate prediction it is necessary to know how the oil slicks behave or, at least, the probability of finding oil slicks in an area.

First, position, shape and size of the oil slicks must be identified. Data related with the oil slicks, like their positions and sizes, has been obtained by treating SAR (*Synthetic Aperture Radar*) satellite images [3, 4]. The satellite images show certain areas where it seems to be nothing, like zone with no waves; that are the oil slicks. With these kind of images it is possible to distinguish between normal sea variability and slicks. It is also important to distinguish between oil slicks and look-alikes.

Once the slicks are identified, it is also essential to know the atmospheric and maritime situation that is affecting the slick in the moment that is being analysed. Information collected from satellites is used to obtain the atmospheric data needed. That is how different variables such as temperature, sea height and salinity are measured in order to obtain a global model that can explain how slicks evolve.

3 Case-Based Reasoning Systems

Case-Based Reasoning origins are in knowledge based systems. CBR systems solve new problems acquiring the needed knowledge from previous situations [5]. The principal element of a CBR system is the case base, a structure that stores the information used to generate new solutions. In the case base, data is organized into cases, where problems and its solutions are related. A case base can then be seen as a kind of database where a series of problems are stored, as long as their solutions and the relation between them.

The learning capabilities of the CBR systems are due to its own structure, composed of four main phases [6]: *retrieval, reuse, revision* and *retention*.

Applying CBR to solve a problem generally implies using other artificial intelligence techniques to solve the problems related with the different phases of the CBR cycle. In this study, a new algorithm is used to structure the case base and to easily and fast recover the most similar cases from the case base. That algorithm is the WeVoS-ViSOM algorithm, which will be explained next.

4 WeVoS-ViSOM for CBR

In this section, a novel classification algorithm for CBR is presented. It is used to sort out all the information that is stored in the case base and to retrieve the most similar cases to the one introduced in the system as a problem to solve.

4.1 Visualization Induced SOM (ViSOM)

The Self-Organizing Map (SOM) algorithm [7] and the Visualization Induced Self-Organizing Map (ViSOM) [8] are different types of Topology Preserving Mappings with a common target: to provide a low dimensional representation of multi-dimensional datasets while preserving the topological properties of the input space.

The ViSOM, aims to directly preserve the local distance information on the map, along with the topology. It constrains the lateral contraction forces between neurons and hence regularises the interneuron distances so that distances between neurons in the data space are in proportion to those in the input space [8].

Update of neighbourhood neurons in ViSOM:

$$w_k(t+1) = w_k(t) + \alpha(t)\eta(v,k,t)\left([x(t) - w_v(t)] + [w_v(t) - w_k(t)]\left(\frac{d_{vk}}{\Delta_{vk}\lambda} - 1 \right) \right) \quad (1)$$

where w_v is the winning neuron, α the learning rate of the algorithm, $\eta(v,k,t)$ is the neighbourhood function where v represents the position of the winning neuron in the lattice and k the positions of the neurons in the neighbourhood of this one, x is the input to the network and λ is a "resolution" parameter, d_{vk} and Δ_{vk} are the distances between the neurons in the data space and in the map space respectively.

4.2 Weighted Voting Summarization (WeVoS)

The idea behind the novel fusion algorithm, WeVoS, is to obtain a final map keeping one of the most important features of this type of algorithms: its topological ordering. WeVoS is an improved version of an algorithm presented in several previous works: [2, 9, 10] and in this study is applied for the first time to the ViSOM in the frame of a CBR.

It is based in the calculation of the "quality of adaptation" of a unit in the same position of different maps, in order to obtain the best characteristics vector in each of the units of the final one. This calculation is performed as follows:

$$V(p) = \frac{|x_p|}{\sum_{m=1}^{M} |x_{mp}|} \cdot \frac{q_p}{\sum_{m=1}^{M} q_{mp}} \quad (2)$$

The model, called WeVoS is described in detail in the *algorithm 1*.

Algorithm 1. Weighted Voting Superposition (WeVoS)

1: train several networks by using the bagging (re-sampling with replacement) meta-algorithm
2: **for each map** (m) **in the ensemble**
3: **for each unit position** (p) **of the map**
4: calculate the quality measure/error chosen for the current unit
5: **end**
6: **end**
7: calculate an accumulate total of the quality/error for each position Q(p) in all maps
8: calculate an accumulate total of the number of data entries recognized by a position in all maps D(p)
9: **for each unit position** (p)
10: initialize the fused map (*fus*) by calculating the centroid (w') of the units of all maps in that position (p)
11: **end**
12: **for each map** (m) **in the ensemble**
13: **for each unit position** (p) **of the map**
14: calculate the vote weight of the neuron (p) in the map (m) by using Eq. 2
15: feed, to the fused map (*fus*), the weights vector of the neuron (p) as if it was an input to the network, using the weight of the vote calculated in Eq. 2 as the learning rate and the index of that same neuron (p) as the index of the BMU.
This causes the unit of the final map (w') to approximate the unit of the composing ensemble (w_p) accordingly to its adaptation.
16: **end**
17: **end**

5 WeVoS-CBR: A New Oil Spill Forecasting System

There have already been CBR systems created to solve maritime problems [11] in which different maritime variables have been used. In this occasion, the data used have been previously collected from different observations from satellites, and then pre-processed, and structured to create the case base. The created cases are the main elements to obtain the correct solutions to future problems, through the CBR system. The developed system determines the probability of finding oil slicks in a certain area after an oil spill has been produced. To properly obtain the solutions, the system uses square divisions of the area to analyze of approximately half a degree side. Then the system calculates the number of slicks in every square as long as the surface covered by them.

The structure of a case in the presented system is composed by the values obtained in each square for the following 14 variables: longitude, latitude, date, sea height, bottom pressure, salinity, temperature, meridional wind, zonal wind, wind strength, meridional current, zonal current, current strength and area of slicks.

The described system includes different AI techniques to achieve the objectives of every CBR phase. Every CBR phase uses one or more AI techniques in order to obtain its solution. Those phases with its related techniques are going to be explained next.

5.1 Creating the Case Base and Recovering the Most Similar Cases

The data used to train the system has been obtained after the Prestige accident, between November 2002 and April 2003, in a specific geographical area to the north west of the Galician coast (longitude between 14 and 6 degrees west and latitude between 42 and 46 degrees north).

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

The WeVos algorithm is also used to recover the most similar cases to the problem introduced in the system. That process if performed once the case base is structured keeping the original distribution of the available variables.

5.2 Adaptation of the Recovered Cases

After recovering the most similar cases to the problem from the case base, those cases are used to obtain a solution. *Growing RBF networks* [12] are used to generate the predicted solution corresponding to the proposed problem. The selected cases are used to train the GRBF network. This adaptation of the RBF network lets the system grow during the training phase in a gradual way increasing the number of elements (prototypes) which work as the centres of the radial basis functions. In this occasion the creation of the GRBF is automatically done, which implies an adaptation of the original GRBF system. The error definition for every pattern is shown below:

$$e_i = 1/p \sum_{k=1}^{p} \| t_{ik} - y_{ik} \| \quad (3)$$

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

5.3 Revision and Retention of the Proposed Solution

In order to verify the precision of the proposed solution, *Explanations* are used [13]. To justify and validate the given solution, the retrieved cases are used once again. To create an explanation, different possibilities have been compared. The selected cases have their own future associated situation. Considering the case and its solution as two vectors, a distance between them can be measured by calculating the evolution of the situation in the considered conditions. If the distance between the proposed problem and the solution given is not bigger than the distances obtained from the selected cases, then the proposed solution considered as a good one, according to the structure of the case base.

Once the proposed prediction is accepted, it can be stored in the case base in order to serve to solve new problems. It will be used equally than the historical data previously stored in the case base. The *WeVoS* algorithm is used again to introduce new elements in the case base. After introducing a new case in the case base, the structure formed by the information stored in the case base, also change, to be adapted to the new situation created.

6 Results

To create the case base, data obtained from different satellites have been used. Temperature, salinity, bottom pressure, sea height, number and area of the slicks, as long as the date have been involved in the case base creation. All these data define the problem case and also the solution case. The problem solution is defined by the variables related to an area, but with the values of the different variables changed to the prediction obtained from the CBR system related to a future point.

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

Table 1 shows a summary of the obtained results. In this table different techniques are compared. The table shows the evolution of the results along with the increase of the number of cases stored in the case base. All the techniques analyzed show an improvement in their results when the number of cases stored in the case base is increased. The "*RBF*" column corresponds to a simple Radial Basis Function Network trained with all the available information. The RBF network generates an output that is considered a solution to the problem giving it, as input, the problem to be solved. The "*CBR*" column represents a simple CBR system, with no artificial intelligence techniques included. The cases are stored in the case bases and recovered considering the Euclidean distance. The most similar cases are selected and after applying a weighted mean depending on the similarity, a solution is proposed. It is a *mathematical* CBR. The "*RBF + CBR*" column corresponds to an approximation that uses a RBF system combined with a CBR methodology. The cases are recovered from the case base using the Manhattan distance to retrieve the most similar cases to the introduced problem. The RBF network is applied in the reuse phase, modifying the selected cases to generate the new solution. The results of the "*RBF+CBR*" column are, normally, better than those of the "*CBR*", mainly due to the removal of worthless data to obtain the solution. Last, the "*WeVoS-CBR*" column shows the results obtained by the proposed system, being better than the three previous solutions analyzed. The solution proposed do not generate a trajectory, but a series of probabilities in different areas, what is far more similar to the real behaviour of the oil slicks.

Table 1. Percentage of good predictions obtained with different techniques

Number of cases	RBF	CBR	RBF + CBR	WeVoS-CBR
100	45 %	39 %	42 %	43 %
500	46 %	41 %	44 %	47 %
1000	49 %	46 %	55 %	63 %
2000	56 %	55 %	65 %	72 %
3000	58 %	57 %	66 %	79 %
4000	60 %	63 %	69 %	84 %
5000	60 %	62 %	73 %	88 %

7 Conclusions and Future Work

A new hybrid intelligent predicting system, based on the CBR methodology, to forecast the probability of finding oil slicks in an area after an oil spill is presented in this paper. The explained system uses information recovered from different orbital satellites. All that information has been used to create the CBR system. The available data is first structure and classified to create the case base, where the knowledge used by the system to generate its predictions is stored. To achieve the different tasks related with the phases of the CBR cycle, the developed system uses diverse artificial intelligence techniques. A new voting algorithm, the Weighted Voting Superposition algorithm is used both to organize the case base and to retrieve the most similar cases to the one introduced as a problem to the system. The great organization capabilities of that new algorithm allow the system to create a valid structure to the case base and to easily recover similar cases from the case base.

Growing Radial Basis Function Networks has been used to generate a prediction using the cases retrieved from the case base. Using this evolution of the RBF networks a better adaptation to the structure of the case base is obtained. The results using Growing RBF networks instead of simple RBF networks are about a 4% more accurate, which is a quite good advance.

It has been proved that the system can predict in already known conditions, showing better results than previously used techniques. The use of the blend of techniques

integrated in the CBR structure makes possible to obtain better results than using the CBR alone (17% better), and also better than using the techniques isolated, without the integration feature produced by the CBR (11% better).

The next step is generalising the learning, acquiring new data to create a base of cases big enough to have solutions for every season. Another improvement is to create an on-line system that can store the case base in a server and generate the solutions dynamically to different requests. This on-line version will include real time connection to data servers providing weather information of the current situations in order to predict *real future* situations.

Acknowledgements

This research has been partially supported by projects BU006A08 and SA071A08 of the JCyL and TIN2006-14630-C03-03 of the MEC.

References

1. Watson, I.: Case-Based Reasoning Is a Methodology Not a Technology. Knowledge-Based Systems 12, 303–308 (1999)
2. Baruque, B., Corchado, E.: WeVoS: A Topology Preserving Ensemble Summarization Algorithm. Data Mining and Knowledge Discovery (submitted)
3. Palenzuela, J.M.T., Vilas, L.G., Cuadrado, M.S.: Use of Asar Images to Study the Evolution of the Prestige Oil Spill Off the Galician Coast. International Journal of Remote Sensing 27, 1931–1950 (2006)
4. Solberg, A.H.S., Storvik, G., Solberg, R., Volden, E.: Automatic Detection of Oil Spills in Ers Sar Images. IEEE Transactions on Geoscience and Remote Sensing 37, 1916–1924 (1999)
5. Aamodt, A.: A Knowledge-Intensive, Integrated Approach to Problem Solving and Sustained Learning. Knowledge Engineering and Image Processing Group. University of Trondheim (1991)
6. Aamodt, A., Plaza, E.: Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. AI Communications 7, 39–59 (1994)
7. Kohonen, T.: The Self-Organizing Map. Neurocomputing 21, 1–6 (1998)
8. Yin, H.: Data Visualisation and Manifold Mapping Using the Visom. Neural Networks 15, 1005–1016 (2002)
9. Baruque, B., Corchado, E., Yin, H.: Visom Ensembles for Visualization and Classification. In: Sandoval, F., Gonzalez Prieto, A., Cabestany, J., Graña, M. (eds.) IWANN 2007. LNCS, vol. 4507, pp. 235–243. Springer, Heidelberg (2007)
10. Corchado, E., Baruque, B., Yin, H.: Boosting Unsupervised Competitive Learning Ensembles. In: de Sá, J.M., Alexandre, L.A., Duch, W., Mandic, D.P. (eds.) ICANN 2007. LNCS, vol. 4668, pp. 339–348. Springer, Heidelberg (2007)
11. Corchado, J.M., Fdez-Riverola, F.: Fsfrt: Forecasting System for Red Tides. Applied Intelligence 21, 251–264 (2004)
12. Karayiannis, N.B., Mi, G.W.: Growing Radial Basis Neural Networks: Merging Supervised Andunsupervised Learning with Network Growth Techniques. IEEE Transactions on Neural Networks 8, 1492–1506 (1997)
13. Sørmo, F., Cassens, J., Aamodt, A.: Explanation in Case-Based Reasoning–Perspectives and Goals. Artificial Intelligence Review 24, 109–143 (2005)

Clustering Likelihood Curves: Finding Deviations from Single Clusters

Claudia Hundertmark[1] and Frank Klawonn[2]

[1] Department of Cell Biology, Helmholtz Centre for Infection Research, Inhoffenstr. 7, D-38124 Braunschweig, Germany
[2] Department of Computer Science, University of Applied Sciences Braunschweig/Wolfenbuettel, Salzdahlumer Str. 46/48, D-38302 Wolfenbuettel, Germany

Abstract. For systematic analyses of quantitative mass spectrometry data a method was developed in order to reveal peptides within a protein, that show differences in comparison with the remaining peptides of the protein concerning their regulatory characteristics. Regulatory information is calculated and visualised by a probabilistic approach resulting in likelihood curves. On the other hand the algorithm for the detection of one or more clusters is based on fuzzy clustering, so that our hybrid approach combines probabilistic concepts as well as principles from soft computing. The test is able to decide whether peptides belonging to the same protein, cluster into one or more group. In this way obtained information is very valuable for the detection of single peptides or peptide groups which can be regarded as regulatory outliers.

Keywords: Clustering, iTRAQ™, LC-MS/MS, likelihood curve.

1 Introduction

Comparative analyses between normal and pathological states of biological systems are an important basis for biological and medical research. Therefore, quantitative analyses of biological components, e.g. proteins, are of particular importance. Proteins are the basic components of cells and responsible for most of the processes in organisms. The basic structure of proteins are amino acid chains, which can be digested into smaller chains termed peptides. The totality of proteins, called the proteome, is highly dynamic and can vary significantly concerning its qualitative and quantitative composition due to changed conditions. A common technology for the analysis of the proteome is called liquid-chromatography mass spectrometry (LC-MS/MS).

In the meantime besides protein identification, mass spectrometry enables relative peptide quantitation by LC-MS/MS. One of the most popular technologies for this purpose is called iTRAQ™, which is based on chemical labelling of peptides. iTRAQ™ allows comparative analyses of up to eight proteinogenic samples in parallel (see [1], [2]). After iTRAQ™-labelling of peptides from different samples and subsequently LC-MS/MS analysis a so-called mass spectrum is available for every detected peptide. The obtained mass spectrum contains information on the