

A Support Vector Regression Approach to Predict Carbon Dioxide Exchange

Juan F. De Paz, Belén Pérez, Angélica González, Emilio Corchado,
and Juan M. Corchado

Abstract. In this study, a new monitoring system for carbon dioxide exchange is presented. The mission of the intelligent environment presented in this work, is to globally monitor the interaction between the ocean's surface and the atmosphere, facilitating the work of oceanographers. This paper proposes a hybrid intelligent system integrates case-based reasoning (CBR) and support vector regression (SVR) characterised for their efficiency for data processing and knowledge extraction. Results have demonstrated that the system accurately predicts the evolution of the carbon dioxide exchange.

Keywords: Carbon dioxide, Support Vector Regression, Case-based Reasoning.

1 Introduction

One of the factors of greatest concern in climactic behaviour is the quantity of carbon dioxide (CO₂) present in the atmosphere. Carbon dioxide is one of the greenhouse gases that helps to make the earth's temperature habitable, so long it maintains certain levels [6]. Traditionally, it has been considered that the main system regulating carbon dioxide in the atmosphere is the photosynthesis and respiration of plants. However, thanks to tele-detection techniques it has been shown that the ocean plays a highly important role in the regulation of carbon quantities, the full significance of which still needs to be determined [7]. Current technology allows us to obtain data and make calculations that were unimaginable some time ago. This data gives us an insight into carbon dioxide's original source, its decrease and the causes for this decrease [1], which allow predictions on its behaviour in the future.

This paper proposes a hybrid intelligent system that integrates case-based reasoning (CBR) and support vector regression (SVR) characterised for their efficiency for data processing and knowledge extraction. CBR is a type of reasoning

Juan F. De Paz, Belén Pérez, Angélica González, Emilio Corchado, and Juan M. Corchado
Departamento Informática y Automática
Universidad de Salamanca
Plaza de la Merced s/n, 37008, Salamanca, Spain
University of Salamanca, Spain
e-mail: {fcofds, lancho, angelica, escorchado, corchado}@usal.es

that uses past experiences to resolve new problems, and is very appropriate for use in scenarios where adaptation and learning abilities are necessary. In order to acquire intelligent behaviours, it is necessary to provide the systems with learning capabilities. One of the possibilities is learning from past experiences, which can facilitate cognitive knowledge. CBR systems are aimed at providing learning and adaptation capacities [3, 8, 9, 10]. The use of past experiences allows these systems to resolve new problems [8, 11]. SVR is a variation of support vector machines, able to provide regression models for non-linear datasets. The combination of CBR and SVR provides an added value to the prediction of the CO₂ exchange. This proposal is a step in this direction and the first step toward the development of predictive models based on non-linear data. The model presented within this work provides great capacities for learning and adaptation to the characteristics of the problem in consideration by using novel algorithms in each of the stages of the CBR cycle that can be easily configured and combined. It also provides results that notably improve those provided by the existing methods for CO₂ analysis.

Section 2 presents the problem that motivates this research. Then, in Section 3 the related work is presented. Section 4 describes the approach proposed in this research. Finally, in section 5 some preliminary results and the conclusions will be presented.

2 Carbon Dioxide Exchange

The oceans contain approximately 50 times more CO₂ in dissolved forms than the atmosphere, while the land biosphere including the biota and soil carbon contains about 3 times as much carbon (in CO₂ form) as the atmosphere [7]. The CO₂ concentration in the atmosphere is governed primarily by the exchange of CO₂ with these two dynamic reservoirs. Since the beginning of the industrial era, about 2000 billion tons of carbon have been released into the atmosphere as CO₂ from various industrial sources including fossil fuel combustion and cement production. It is important, therefore, to fully understand the nature of the physical, chemical and biological processes, which govern the oceanic sink/source conditions for atmospheric CO₂ [7, 4].

The need to quantify the carbon dioxide valence, and the exchange rate between the oceanic water surface and the atmosphere, has motivated us to develop the distributed system, presented here, that incorporates a CBR model capable of estimating such values using accumulated knowledge and updated information. The CBR model receives data from satellites, oceanographic databases and oceanic and commercial vessels. The case-based reasoning system incorporated is able to optimize tasks such as the interpretation of images using various strategies [5]. The information received is composed of satellite images of the ocean's surface, wind direction and strength, and other parameters such as water temperature, salinity and fluorescence. An improvement of the forecasting methods presented in [0, 1, 2] is incorporated in the CBR model presented in this paper.

It is possible to find different systems in literature aimed at predicting CO₂ exchange rates [15, 16, 1]. These works propose an approach based on obtaining

regression models that are generated manually by experts. The works presented in [15, 16] focus on the variation of the exchange of CO₂ produced during the day and during the night, while the work presented in [1] prioritizes the difference of pressures that exists between the ocean surface and the air. The regression models proposed in these works have, in general, a high level of complexity and sometimes require the incorporation of new variables once the model has been generated, which means recalculating the equations of the model. In this sense, the estimation of the CO₂ exchange rate obtained by means of manual models presents deficiencies when working in dynamic environments, where the system needs to automatically adapt itself to the changes that occur in its surroundings and evolve over time.

3 Support Vector Regression

SVR comes from Support Vector Machine (SVM) and is specialized in obtaining regression models by means of a change in the dimensionality of the data. SVM is a supervised learning technique that is applied to the classification and regression of different elements. SVM facilitates working with data that cannot be adjusted to linear models [12], initially conceived to obtain classifications in linear separable problems, by means of finding a hyperplan able to separate the elements of a set. One of the advantages of SVM is that it also allows separation of non-linear data. To obtain non-linear separation, SVM performs a mapping of the initial data into a high dimensionality space, where the data can be linearly separable using specific functions. Given that the dimensionality of the new space can be very high, most of the time it is not viable to use hyperplans to obtain linear separation. As a solution, non-linear functions called kernels are used. SVR is a variation of SVM to generate regressions [12, 13, 14]. The aim is to adjust the data. As in the case of SVM there is a mapping of the input data into a high dimensionality space. In this new space the regression can be carried out without the initial limitations. Equation (1) shows the linear regression obtained by means of $g_j(x)$ functions that transform the input vectors from their initial coordinates to a high dimensionality space.

$$f(\vec{x}, \vec{w}) = \sum_{j=1}^m w_j g_j(\vec{x}) + b \quad (1)$$

4 System Description

The model proposed in this paper presents a case-based reasoning systems, which models the air-sea CO₂ exchange rate. The CBR system has two aims. The first one is to generate models which are capable of predicting the atmospheric/oceanic interaction in a particular area of the ocean in advance. The second one is to permit the use of such models.

Moreover, the reasoning cycle is one of the activities carried out by the system. We can see how the reasoning cycle of a case-based reasoning system is included among the activities, composed of stages of retrieval, reuse, revise and retain. Also, an additional stage that introduces expert's knowledge is used.

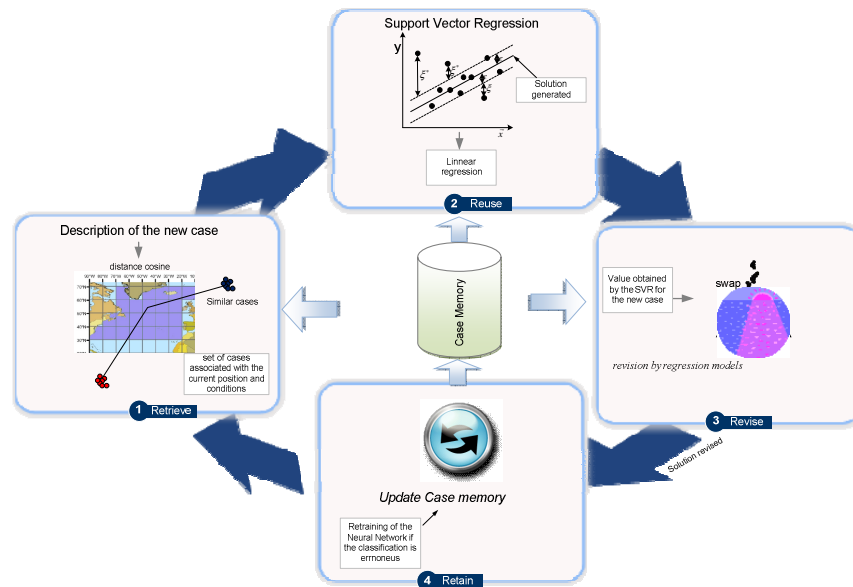


Fig. 1 Internal structure of CBR-System

Figure 1 shows the internal structure of the proposed CBR. Problem description (initial state) and solution (situation when final state is achieved) are represented as a set of values related to the oceanic and atmospheric status, the final state is the solution achieved for the problem (the predicted flux of CO₂), and the sequences of actions are the steps carried out in each of the stages of the CBR cycle. The structure of a case for the CO₂ exchange problem can be seen in Table 1. Table 1 shows the description of a case: DATE, LAT, LONG, SST, S, WS, WD, Fluo_calibrated, SW pCO₂ and Air pCO₂. Flux of CO₂ is the value to be identified. DATE represents the date of the case, LAT represents the latitude of the location where the data has been obtained and LONG, the longitude in decimal degrees. SST represents the temperature of the ocean and S, the salinity. WS is the wind strength and WD is the wind direction. Fluo_calibrated represents the fluorescence calibrated with chlorophyll.

4.1 Retrieve

The prediction for the CO₂ exchange rate is obtained from the parameters shown in Table 1. The prediction is carried out taking into consideration different regions

Table 1 Case Attributes.

Case Field	Measurement
DATE	Date (dd/mm/yyyy)
LAT	Latitude (decimal degrees)
LONG	Longitude (decimal degrees)
SST	Temperature (°C)
S	Salinity (unitless)
WS	Wind strength (m/s)
WD	Wind direction (unitless)
Fluo_calibrated	Fluorescence calibrated with chlorophyll
SW pCO ₂	Surface partial pressure of CO ₂ (micro Atmospheres)
Air pCO ₂	Air partial pressure of CO ₂ (micro Atmospheres)
Flux of CO ₂	CO ₂ exchange flux (Moles/m ²)

of the Atlantic Ocean and, in order to obtain an effective prediction, the system needs to recover the appropriated past experiences. That is, those cases that contain problem descriptions for similar latitudes and longitudes. In order to establish this first filter in the retrieve stage, the oceanic region taken into consideration for this study was divided into grids of 10° for the latitudes and longitudes. The predictions and estimations are provided for the complete grid as a set. Once a region has been selected, the selection of the most similar case study is performed according to the cosine distance applied to the following set of variables SST, S, WS, WD, Fluo_calibrated, and Air pCO₂. The cosine distance is used to avoid data normalization and corresponding problems with the data units.

4.2 Reuse

Once the most similar cases have been retrieved, the regression model is generated. As indicated in Section 4, the technique that will be used to create the regression model is Support Vector Regression (SVR). The input vector x represents a dataset with the structure presented in Table 1. The input vector can be represented as $x=(\text{DATE}, \text{LAT}, \text{LONG}, \text{SST}, \text{S}, \text{WS}, \text{WD}, \text{Fluo_calibrated}, \text{SW pCO}_2 \text{ and Air pCO}_2)$. The regression is obtained making use of all the vectors provided by the most similar cases retrieved in the previous stage of the CBR cycle, and the SVR is calculated following the algorithm presented in Section 4. The regression model is used to estimate the swap of the new case, which is used to generate the prediction value.

4.3 Revise

This phase is performed in an automatic fashion, and takes into account the error rate provided by the SVM. The error rate is calculated from the previous existing data using the coefficient of variation, in such a way that if the value obtained is minor than a pre-fixed value, then the prediction can be considered as successful. It is necessary to take into account that once the real data are obtained, the

predicted exchange values are eliminated. The estimated values are only used to obtain prediction models under different conditions.

Moreover, during the revision stage an equation (F) is used to validate the proposed solution p^* .

$$F = k\alpha(pCO_2SW - pCO_2AIR) \quad (2)$$

Where: F: is the flux of CO_2 and k: is the gas transfer velocity. Then

$$k = (-5,204 Lat + 0,729 Long + 2562,765) / 3600 \quad (3)$$

5 Results and Conclusions

In order to make evident the need to carry out a separation of the data in latitudes and longitudes, Figure 2 shows the results obtained after calculating the predictions using SVR with a dataset of 365 cases distributed in a homogeneous manner along the North Atlantic Ocean. The kernel function used for the experiments was polynomial and the loss function was \mathcal{E} -insensitive. The blue lines in Figure 2 represent the real value of the data and the red lines represent the predicted values. As can be seen in Figure 2, the error rate obtained in this experiment is very high compared the error rate obtained in Figure 3. The numerical values represent the millions of Tonnes of carbon dioxide that have been absorbed (negative values) or generated (positive values) by the ocean during each of the three months.

To evaluate the prediction capacities of the systems presented in this study, different tests were performed along the North Atlantic oceanic region with data obtained during 2009. In each of the tests, when a case containing the description of an oceanic area was introduced to the system, the most similar cases in the grid with the same latitude and longitude as the new case were taken into consideration

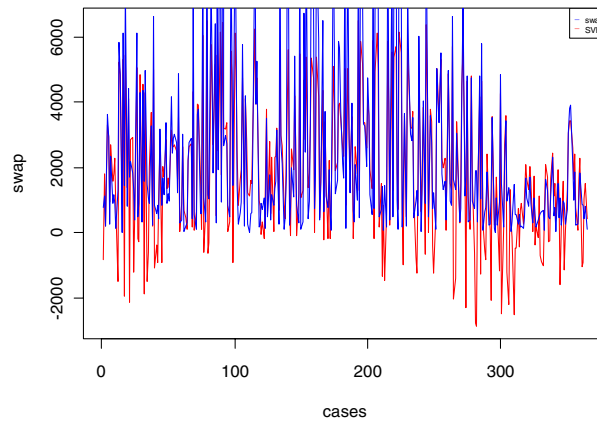


Fig. 2 Prediction if previous similar cases are selected

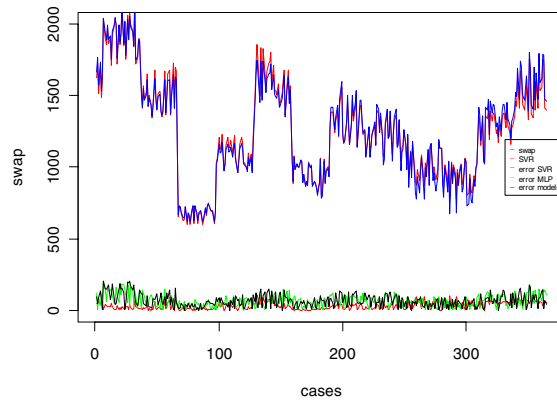


Fig. 3 Comparison between the real values and the prediction values for the CO₂ exchange rate.

to obtain the prediction. Figure 3 shows the results obtained from the experiment. The blue line represents the real value and the red line represents the predicted value. Moreover, Figure 3 shows the absolute error rate obtained for the predicted value (red line) provided by the SVR. The absolute error rate obtained was 31.43, with an error deviation of 39.63. The error percentage obtained was 2.5%.

The absolute error rate obtained with the SVR has been compared to the error rate provided by alternative techniques, such as the multilayer perceptron and the oceanographers' manual models. Figure 3 shows the absolute error rate obtained for each of these predictions. The green line represents the error introduced in the system when the prediction is carried out using a multilayer perceptron. The multilayer perceptron used 27 neurons in the hidden layer and the final error percentage obtained was 5.1%. Finally, the error rate introduced in the system when the manual models are considered was 6.7%.

This study has presented a CBR intelligent system to predict and monitor the CO₂ exchange rate in the North Atlantic Ocean. It applies a hybrid reasoning system specifically designed to analyze data from satellite images and vessels and predict potential CO₂ fluxes in order to provide an innovative method for exploring the CO₂ exchange prediction process and extract knowledge. This knowledge helps human experts to understand the prediction process and to obtain conclusions about the relevance of the situation of the oceanic environment.

Acknowledgements. This work has been supported by the MICINN TIN 2009-13839-C03-03 project.

References

1. Bajo, J., Corchado, J.M.: Evaluation and monitoring of the air-sea interaction using a CBR-Agents approach. In: Muñoz-Ávila, H., Ricci, F. (eds.) ICCBR 2005. LNCS (LNAI), vol. 3620, pp. 50–62. Springer, Heidelberg (2005)

2. Bajo, J., Corchado, J.M.: Multiagent architecture for monitoring the North-Atlantic carbon dioxide Exchange rate. In: Marín, R., Onaindía, E., Bugarín, A., Santos, J. (eds.) CAEPIA 2005. LNCS (LNAI), vol. 4177, pp. 321–330. Springer, Heidelberg (2006)
3. Corchado, J.M., Aiken, J., Corchado, E., Lefevre, N., Smyth, T.: Quantifying the Ocean's CO₂ Budget with a CoHeL-IBR System. In: Funk, P., González Calero, P.A. (eds.) ECCBR 2004. LNCS (LNAI), vol. 3155, pp. 533–546. Springer, Heidelberg (2004)
4. Kolodner, J.: Case-based reasoning. Morgan Kaufmann, San Francisco (1993)
5. Lefevre, N., Aiken, J., Rutllant, J., Daneri, G., Lavender, S., Smyth, T.: Observations of pCO₂ in the coastal upwelling off Chile: Sapatial and temporal extrapolation using satellite data. *Journal of Geophysical research* 107(6), 8.1–8.15 (2002)
6. Perner, P.: Different Learning Strategies in a Case-Based Reasoning System for Image Interpretation. In: Smyth, B., Cunningham, P. (eds.) EWCBR 1998. LNCS (LNAI), vol. 1488, pp. 251–261. Springer, Heidelberg (1998)
7. Sarmiento, J.L., Dender, M.: Carbon biogeochemistry and climate change. *Photosynthesis Research* 39, 209–234 (1994)
8. Takahashi, T., Olafsson, J., Goddard, J.G., Chipman, D.W., Sutherland, S.C.: Seasonal Variation of CO₂ and nutrients in the High-latitude surface oceans: a comparative study. *Global biochemical Cycles* 7(4), 843–878 (1993)
9. Kolodner, J.: Maintaining organization in a dynamic long-term memory. Morgan Kaufmann, San Francisco (1993)
10. Kolodner, J.: Maintaining organization in a dynamic long-term memory. *Cognitive Science* 7, 243–280 (1983)
11. Kolodner, J.: Reconstructive memory, a computer model. *Cognitive Science* 7(4), 281–328 (1983)
12. Leake, D., Kendall-Morwick, J.: Towards Case-Based Support for e-Science Workflow Generation by Mining Provenance. In: Althoff, K.-D., Bergmann, R., Minor, M., Hanft, A. (eds.) ECCBR 2008. LNCS (LNAI), vol. 5239, pp. 269–283. Springer, Heidelberg (2008)
13. Vapnik, V.N.: An overview of statistical learning theory. *IEEE Transactions on Neural Networks* 10, 988–999 (1999)
14. Vapnik, V.: *The Nature of Statistical Learning Theory*. Springer, Heidelberg (1995)
15. Smola, A., Scolköpf, B.: A tutorial on support vector regression. *Statistics and Computing* (2003)
16. Jeffer, C.D., Woolf, D.K., Robinson, I.S., Donlon, C.J.: One-dimensional modelling of convective CO₂ exchange in the Tropical Atlantic. *Ocean Modelling* 19(3-4), 161–182 (2007)
17. Jeffery, C.D., Robinson, I.S., Woolf, D.K., Donlon, C.J.: The response to phase-dependent wind stress and cloud fraction of the diurnal cycle of SST and air-sea CO₂ exchange, vol. 23(1-2), pp. 33–48 (2008)
18. Rutgersson, A., Smedman, A.: Enhanced air-sea CO₂ transfer due to water-side convection. *Journal of Marine Systems* 80(1-20), 125–134 (2010)