Contemporary Engineering Sciences, Vol. 11, 2018, no. 95, 4717 - 4724
HIKARI Ltd, www.m-hikari.com
https://doi.org/10.12988/ces.2018.89517

Prediction of a Time Series Using Regression by Vector Support Machines

Pedro Pablo Cárdenas Alzate

Department of Mathematics and GEDNOL
Universidad Tecnológica de Pereira
Pereira, Colombia

Germán Correa Vélez

Department of Mathematics and GIMAE
Universidad Tecnológica de Pereira
Pereira, Colombia

Fernando Mesa

Department of Mathematics and GIMAE
Universidad Tecnológica de Pereira
Pereira, Colombia

Copyright © 2018 Pedro Pablo Cárdenas Alzate et al. This article is distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

The Vector Support Machines (VSM) have their greatest application in the area of pattern recognition, for the solution of biclass and multiclass classification problems, however, they can also be used for the solution of regression problems in order to predict time series. In this paper, the problem of linear and non-linear regression is developed for the prediction of a time series using Vector Support Machines (VSM), in order to show the advantages obtained with conventional regression techniques. For this, a UCI Machine Learning database was taken consisting of the variation of 8 design parameters for the energy efficiency of one.

Keywords: Vector Support Machines (VSM), regression problems, time series

1 Introduction

The theory of Vector Support Machines (VSM) is a recent technique traditionally used to solve classification problems in pattern recognition applications, however, this technique can also be used for the regression problem , which consists of the adjustment of curves by using a set of data that is available. The use of VSMs for the prediction of time series covers many practical applications in medicine, financial markets, electrical services, control systems, climate prediction, among other non-linear processes. The main difference of VSMs with conventional regression methods is the use of structural risk minimization (SRM) and not empirical risk minimization (ERM), which it is equivalent to the minimization of the superior error of the generalization instead of the minimization of the training error, therefore, it is expected that this technique works better than the conventional ones [1]. In this document the VSM is applied to a data set for the efficient design of the heating load of a building. The data set consists of 8 design parameters (characteristics) and 768 construction forms (samples) in order to perform the prediction of two heating load and cooling load output variables [2].

The VSM is a technique with few free parameters of which an exact way of calculating them is not available, therefore, an experimental study will be carried out for different values of these between reasonable limits. The work is organized as follows: In the next section we present the prediction of a time series in a general way, the theoretical part of how the VSM parameters are calculated to describe the system model is also developed. In section 3, describe the problem that you want to address using the VSMs to the data set worked on. Then, in section 4 the results obtained by simulations carried out in Matlab are shown. Finally, we conclude about the advantages that VSMs have over conventional regression techniques.

2 Prediction of a time series with VSM

2.1 Prediction of a time serie

When it is desired to analyze the dynamics of a system, it is assumed that the dynamic system is soft $F(t, X)$, that is, $F : \mathbb{R} \times S \to S$ where S is an open set of a Euclidean space. The system $F(t, X) = F_0(X)$ fulfills the initial condition $F_0(X) = X_0$. The objective of the dynamic system analysis is to

be able to describe the behavior of the trajectories $X(t)$ for any given initial value X_0. The interest that is had in this work is the inverse problem, where the dynamic model of the system is not known, that is to say, it only has some measurements of the variables that interact on the system. Mathematically, it is said that we have $x(t)$ finite such that $x(t)$ is a portion of $X(t)$, which is assumed to be maintained on a variety of dimension D [3]. The objective is to predict the future values of $x(t)$ that describes in a certain way the dynamics of the system. At first glance it seems an impossible task to perform the description of the dynamics with only a few variables, however the Takens immersion theorem [4] ensures that under certain conditions and for many values of τ and some values of $m \leq D + 1$, which are the delay time and the insertion dimension, a smooth map $f : \mathbb{R}^m \to \mathbb{R}$ can be constructed such that:

$$x\left(n\tau\right) = f(x\left(n-1\right)\tau, x\left(n-2\right)\tau, \ldots, x\left(n-m\right)\tau) \tag{1}$$

2.2 VSM for regression

The VSM was developed by Vapnik and other authors in the past decade (1995) [7], initially as a technique that performed classification tasks in pattern recognition problems, however, this can also be applied to the regression problems and it is common to find them with the name of SVR (for its acronym in English Support Vector Regression). The method of the VSM in general is based on static learning, that is, it is trained with a set of past and present data with which a function that will describe the dynamics of the system will be estimated, based on the data, since it does not his exact model is known.

Given a set of data $G = \{x(t), y(t)\}$, where t is the number of samples N that are available: $t : 0, 1, 2, \ldots, N-1$ and $\hat{y}(t)$ is some prediction of the output for values of $t \geq N$. An algorithm for prediction is to find the function $f(x)$ in such a way that:

$$\hat{y}\left(t\right) = f\left(x\left(t-a\right), x\left(t-b\right), x\left(t-c\right), \ldots\right) \tag{2}$$

The function $f(x)$ will be determined by some type of regression, the general forms being those indicated in the following equations

$$f\left(x\right) = \omega \cdot x + b \tag{3}$$

$$f\left(x\right) = \omega \cdot \phi(x) + b, \tag{4}$$

where x are the characteristics of the data set, w and b are coefficients that must be estimated from the data set.

3 Analysis and results - Description of the data base

The set of data used in this document is the result of simulations carried out in *AUTODESK ECOTEC by A. Tsanas, A. Xifara* in order to analyze the energy efficiency of buildings that have 8 constructive parameters, a total of 12 different construction forms with a constant volume of $771.75\,m^3$ with different surface areas and dimensions, each construction form consisting of 18 elementary cubes of size $3.5m \times 3.5m \times 3.5m$ [2]. The material used is the same for all constructions. In addition, the simulation assumes that the buildings are located in Athens, Greece, being residential homes for 7 people with sedentary activity ($70W$). In total, 768 building forms were generated.

To apply the regression technique by SVM to the data set, the model is trained with 80% and 20% will be the set of data from which the outputs are to be predicted. The following table indicates each of the variables involved in the problem.

Mathematical representation	Variable
x_1	Relative compactness
x_2	Superficial area
x_3	Wall area
x_4	Roof area
x_5	Total height
x_6	Orientation
x_7	Glazing area
x_8	Glazing area distribution
y_1	Glazing area distribution
y_2	Cooling load

Table 1: Summary of input and output variables.

3.1 Results

In this section the results obtained when the VSM is applied to the prediction problem of the variables of treated outputs will be presented. As mentioned above, the VSM have two free parameters which must be selected empirically, these are the values of C and $\in$. In addition, the Kernel function to be applied to project the data set must also be empirically selected. space of characteristics. In the following figures you can see how the variation of these parameters affects the regression using a Gaussian kernel, for the output y_1.

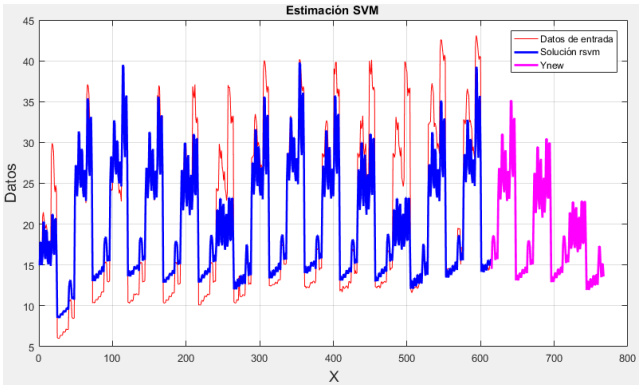

Figure 1: Regression VSM with $C = 5$ y $\in= 10^{-4}$.

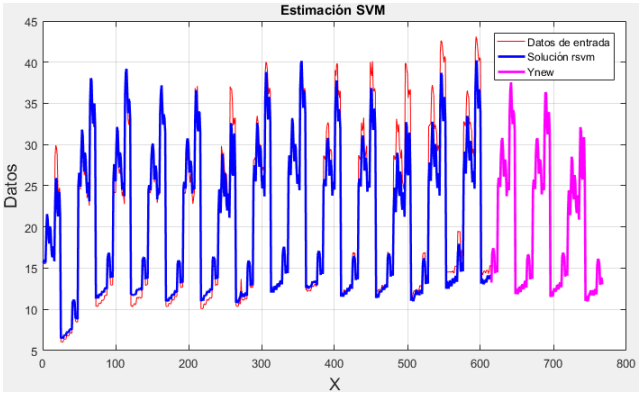

Figure 2: Regression VSM with $C = 10$ y $\in= 10^{-8}$.

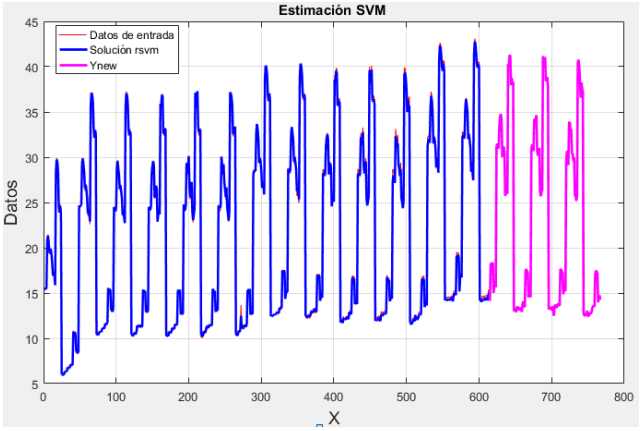

Figure 3: Regression VSM with $C = 100$ y $\in= 10^{-3}$.

The variation of the parameter $\in$ indicates the complexity of the optimization problem, since if a very high value is taken for this, a lot of error is being allowed because our prediction would be farther from our objective y_i, instead if we take a small value of $\in$ we would be allowing very little error and the complexity would increase, therefore the most suitable value for this problem is $\in= 10^{-4}$. On the other hand the parameter C when it takes very small values, increases the mean square error for the estimation of the data, this is because when the value of $C = 0$ there is no regularization, while for large values of C as shown in Figure 3, the mean square error tends to become zero [5].

Figure 4 shows the estimation of the data set using a polynomial Kernel with which an estimate with a very low error is obtained, however when the data is predicted 20% of the data that was left to test, it is noted that the error starts to grow which generates a disadvantage compared to the Gaussian kernel [6].

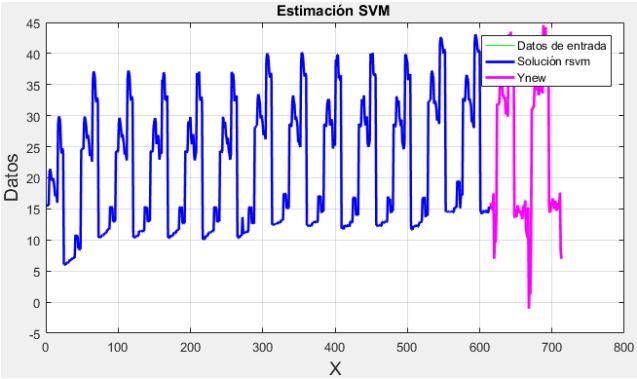

Figure 4: VSM regression with polynomial kernel of order 3.

The results obtained for the output y_2 are observed in the following figures.

4 Conclusion

The SVM algorithm has to be compared with conventional methods, providing better results, in addition to having a mathematically more elaborated description compared to other meta-cultural techniques. On the other hand, the optimization problem is a quadratic convex optimization problem, which guarantees a unique minimum. It should also be noted that the models depend on few parameters making the development methodology much easier. Finally,

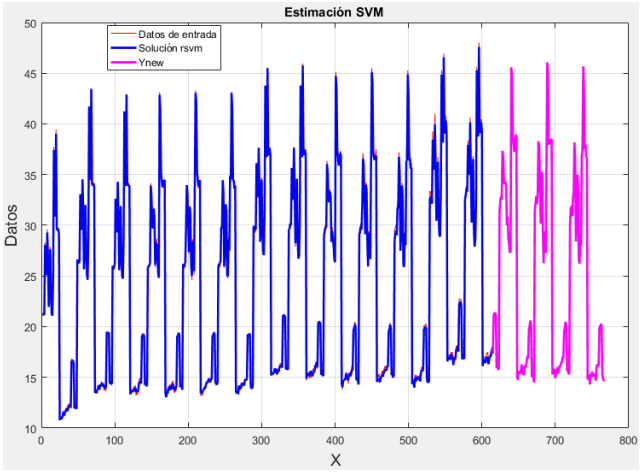

Figure 5: Regression with VSM for y_2.

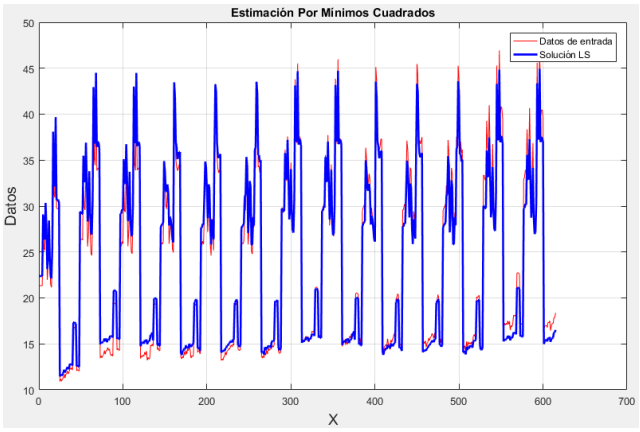

Figure 6: Regression with least squares for y_2.

the results obtained are usually simply a combination of few support vectors, which makes it easy.

Faced with the results of the treated data set, it can be shown that the prediction by VSM for the output variables has a very high performance with an average square error for y_1 of 0.63 and for y_2 of 1.12, which indicates that we can Easily deduct the heating load and the cooling load without requiring the design of a new building by simulation in *AUTODESK ECOTEC.*. The results obtained in this document are consistent with the theory of literature where it is stated that the SVM have a greater efficiency compared to the classical methods such as least squares, since these can not consider the multiple

collinearity of the input variables.

Acknowledgements. We would like to thank the referee for his valuable suggestions that improved the presentation of this paper and our gratitude to the Department of Mathematics of the Universidad Tecnológica de Pereira (Colombia), the GEDNOL Research Group and GIMAE (Grupo de Investigación en Matemática Aplicada y Educación).

References

[1] J. Montserrat et al., *Control Optimo*, Universidad Politécnica de Valencia, 2007.

[2] A. Tsanas, A. Xifara, Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools, *Energy and Buildings*, **49** (2012), 560-567. https://doi.org/10.1016/j.enbuild.2012.03.003

[3] S. Mukherjee, E. Osuna, F. Girosi, Nonlinear prediction of chaotic time series using support vector machines, *Neural Networks for Signal Processing VII. Proceedings of the 1997 IEEE Signal Processing Society Workshop*, (1997), 511-520. https://doi.org/10.1109/nnsp.1997.622433

[4] F. Takens, Detecting strange attractors in fluid turbulence, *Dynamical Systems and Turbulence*, Springer-Verlag, Berlin, 1981, 366-381. https://doi.org/10.1007/bfb0091924

[5] R. Kalman, A new approach to linear filtering and prediction problems, *Journal of Basic Engineering*, **82** (1960), no. 1, 35. https://doi.org/10.1115/1.3662552

[6] S. Schiavon, K. Lee, F. Bauman, T. Webster, Influence of raised floor on zone design cooling load in commercial buildings, *Energy and Buildings*, **42** (2010), no. 8, 1182-1191. https://doi.org/10.1016/j.enbuild.2010.02.009

[7] Y. Gala, *Algoritmos SVM para Problemas Sobre Big Data*, Universidad Autónoma de Madrid, 2013.

Received: October 7, 2018; Published: November 12, 2018