Prediction of an Economics Time Series Using Support Vector Machines

No Author Given

No Institute Given

Abstract. In the current paper the series of consumer prices index (CPI) of the Spanish economy with 1992 as a basis is analyzed, using a new approach based on the theoretical developments of V. N. Vapnik on "Theory of the statistical learning".
The well-known technique is applied as "Support Vector Machines" (SVM) in order to build a regression function, starting from the series, that allows us to carry out predictions. To do that, the embedding dimension of the times series used in the chaotic analysis of series is considered.
This study concluded by giving the predictions at the end of 1999 and exposing a theoretical problem that arises from the predictions of this example.

1 Introduction

Following Mukherjee [1], who indicates it is necessary in order to assure the reliability of the Support vector Machines, to carry out a lot of experimentation, the current practical work of implementatión of this tool is established to get predictions of this examples.

The times series we are interested in, is the consumer prices index of the Spanish economy with 1992 as a basis whose values are given in the table 1.

Table 1. *Series of numbers consumer prices index of the Spanish economy.*

Year	January	February	March	April	May	June	July	August	September	October	November	December
1992	98.576	99.233	99.592	99.485	99.745	99.726	100.050	100.962	101.795	101.856	101.921	102.227
1993	103.185	103.218	103.581	104.035	104.322	104.581	104.955	105.583	106.180	106.576	106.755	107.262
1994	108.346	108.385	108.743	109.171	109.394	109.512	109.941	110.651	110.988	111.229	111.422	111.914
1995	113.074	113.628	114.290	114.896	114.942	115.051	115.069	115.394	115.848	116.064	116.372	116.748
1996	117.462	117.782	118.200	118.871	119.281	119.181	119.340	119.678	119.970	120.134	120.141	120.497
1997	120.847	120.765	120.825	120.869	121.045	121.041	121.263	121.798	122.401	122.356	122.599	122.925
1998	123.215	122.927	122.984	123.289	123.450	123.530	123.986	124.318	124.410	124.421	124.309	124.653
1999	125.111	125.185	125.737	126.202	126.198	126.225	126.773	127.313	127.559	...	...	...

As is recommended in all studies about times series, the first step is to visualize the graph of the observations. This graph is place in the top left corner of the figure 1 and it seems to suggest a tendency close to the lineal type. From a work plan we don't follow the approach of classifying the series when these series have tendency and seasonality, as is done in the classical developments of the analysis of times series. The point of view used it to choose one model that minimizes an appropriate error indicator.

2 Theory

Given times series $\{y_1, \cdots, y_{T'}\}$, we consider a partition from it as follows,

$$\{\overbrace{y_1, , y_2, \cdots, y_T}^{training} \ , \ \overbrace{y_{T+1}, \cdots, y_{T'}}^{test}\} \tag{1}$$

The objective, given the training values, is to build a regression function. This function will be useful for us in order to carry out a prediction of each test values. Let $\widehat{y}_T(\ell)$ be the prediction of $y_{T+\ell}$ starting from the training values, where $\ell = 1, \cdots, H$ with $H = T' - T$. It is considered as an error measure:

$$EAPM = \frac{1}{H} \sum_{\ell=1}^{H} \frac{|y_{T+\ell} - \widehat{y}_T(\ell)|}{y_{T+\ell}}. \tag{2}$$

Fixed a value m, let $\mathcal{X}_m = \{x_i : i = m+1, \cdots, T\}$ be, where $x_i = (y_{i-1}, y_{i-2}, \cdots, y_{i-m}) \in R^m$ a set of m-dimensional vectors[1]. We suppose that this set is the realization of a continuous random vector, X, m-dimensional that is unknown but we do know of its existence.

Let $\mathcal{Y}_m = \{y_i : i = m+1, \cdots, T\}$ be the set generated by a random conditional variable $Y/X = x$ with distribution function $F_{Y/X}(y)$ that is also unknown but we do know of its existence.

The problem is to find the regression function of Y on X, which we do know for definition:

$$r(x) \stackrel{def}{=} E[Y/X] = \int y \, dF(y/x). \tag{3}$$

The problem of regression estimation is very hard because the conditional distribution function is unknown and all we have is information about it by the sample $Z = \{(x_i, y_i),\ i = m+1, \cdots, T\}$. The way of approaching this problem from a theoretical point of view can be found in [3].

The estimation of $r(x)$ is carried out in a class of functions $\mathcal{F}$ where its elements take the form:

$$f : \mathbb{R}^m \longrightarrow \mathbb{R}, \qquad f(x) = \omega \cdot x + b \qquad \text{with} \quad \omega \in \mathbb{R}^m,\ b \in \mathbb{R}. \tag{4}$$

that is the class of the lineal functions[2].

When a function is chosen $f \in \mathcal{F}$, it is very possible for the set Z to have errors ($f(x_i) \neq y_i$). With the aim of penalizing these errors, the ε-insensitive loss function (Vapnik function) is considered:

$$|f(x) - y|_\varepsilon = \begin{cases} |f(x) - y| - \varepsilon, & \text{if } |f(x) - y| > \varepsilon, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

that is, if the difference between $f(x_i)$ and y_i is less that ε then, it is considered that no errors have been made error[3] in estimating the true value of y_i. These errors are denoted by ξ_i if $f(x_i) > y_i$ and ξ_i^* otherwise.

[1] This value m is called embedding dimension in the chaotic study of times series.
[2] " $\cdot$ " denote the dot product in $\mathbb{R}^m$.
[3] A detailed study is in [**?**].

Fixed a value $\varepsilon > 0$ and a value $C > 0$ (the parameter C is known in regularization theory to be related to the amount of noise in the data), the optimization problem that it is considered on the Support vector Machines is:

$$\min ||\omega||^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*), \quad \text{subject to } |f(x_i) - y_i| < \varepsilon \text{ for } i = 1, \cdots, n. \quad (6)$$

It can be shown that the system (6) has a unique solution because the objective function is strictly convex and the equality constraints are linear. The solution of this problem (see[?,3]) is:

$$\omega = \sum_{j=1}^{N_{SV}} \alpha_j \, x_j \quad (7)$$

where α_j are the so called Lagrange multipliers (also called dual variables) associated with the problem (6) and N_{SV} denote the number of support vectors from $\mathcal{X}_m$. Therefore the regression estimation[4] is:

$$f(x) = \sum_{j=1}^{N_{SV}} \alpha_j \, x_j \cdot x + b. \quad (8)$$

Keeping in mind the form of the solution given in (8), it is possible to consider other classes of functions $\mathcal{F}$ without considering more than some admissible kernels functions[5] $k(x, y)$ (Polinomial, B-Splines, RBF, Splines, ...).

3 Implementation

In order to solve the optimization problem (6) Matlab program version 5.3.0 and the SVM package are used. We take $C = 1$ and the value $\varepsilon = 0.1$ is chosen because of our empirical point of view; cost is not implied if to take a error in less than a tenth of point in the prediction of the CPI.

With the aim of selecting the embedding dimension, m, this value is allowed of varying and it is observed how the errors indicator is behaved. In this way table 2 is obtained.

From the table 2 and the observation in the inferior left corner of figure 1, we conclude that the embedding dimension more appropriate is $m = 16$, because it provides the minor $EAPM$. It is also observed that starting from this value of m an effect takes place from backfitting in the predictions that motivates the $EAPM$ grows. In the superior right corner of figure 1 we observe how the predictions, that are obtained on the test values of the CPI series, behave. In that corner, we indicate the symbol $\circ$ for the real values and with the symbol $+$ the predictions.

Once the value of m is selected the overall CPI serie is considered and all the well-known values are incorporated with the objective of obtaining a predictor

[4] The parameter b is estimated from the conditions of Karush-Kuhn-Tucker.
[5] See [2].

Table 2. *Result of the execution of the SVM.*

m	1	2	3	4	5	6	7	8	9
error·100	0.1622	0.1574	0.1436	0.1450	0.1450	0.1401	0.1446	0.1419	0.1358
N_{SV}	47/71	44/70	47/69	48/68	48/67	45/66	48/65	45/64	47/63
$\%N_{SV}$	0.6620	0.6286	0.6812	0.7059	0.7164	0.6818	0.7385	0.7031	0.7460
m	10	11	12	13	14	15	**16**	17	18
error·100	0.1325	0.1282	0.1337	0.1300	0.1277	0.1270	**0.1251**	0.1553	0.1643
N_{SV}	44/62	38/61	39/60	38/59	38/58	38/57	**37/56**	39/55	39/54
$\%N_{SV}$	0.7097	0.6230	0.6500	0.6441	0.6552	0.6667	**0.6607**	0.7091	0.7222
m	19	20	21	22	23	24	25	26	27
error·100	0.1601	0.1468	0.2115	0.2221	0.2874	0.4525	0.6736	0.7230	0.7253
N_{SV}	40/53	38/52	40/51	39/50	35/49	35/48	35/47	30/46	30/45
$\%N_{SV}$	0.7547	0.7308	0.7843	0.7800	0.7143	0.7143	0.7447	0.6522	0.6667

for future values of the series. In this way we are in good conditions of carrying out the following predictions for the last months of 1999[6].

Table 3. Predictions

	October	November	December
Predictions	127,6868	127,7571	127,8915

4 Conclusion

From the current paper, it can be concluded that it is interesting applying the SVM for the study of times series in general, and the economic series in matter because this new tool can be more appropriate in some situations than in the traditional formulations.

On the other hand, if the percentage of the total support vectors is considered of training vectors as is observed in the inferior right corner of figure 1. This percentage varies according of the embedding dimension. It is a well-known result that this percentage grows when the value of ε it is approach to zero, but the following question arises: How does the percentage of support vectors affect the embedding dimension? In this practical example, we can observe how the band of fluctuation is relatively small, because the percentage varies between 0.6230 and 0.7843.

[6] To close the work it is know that in October from 1999, the CPI is 127.559. These predictions are translated in an annual inflation rate of 2.6% for 1999, a tenth higher than the last government prediction.

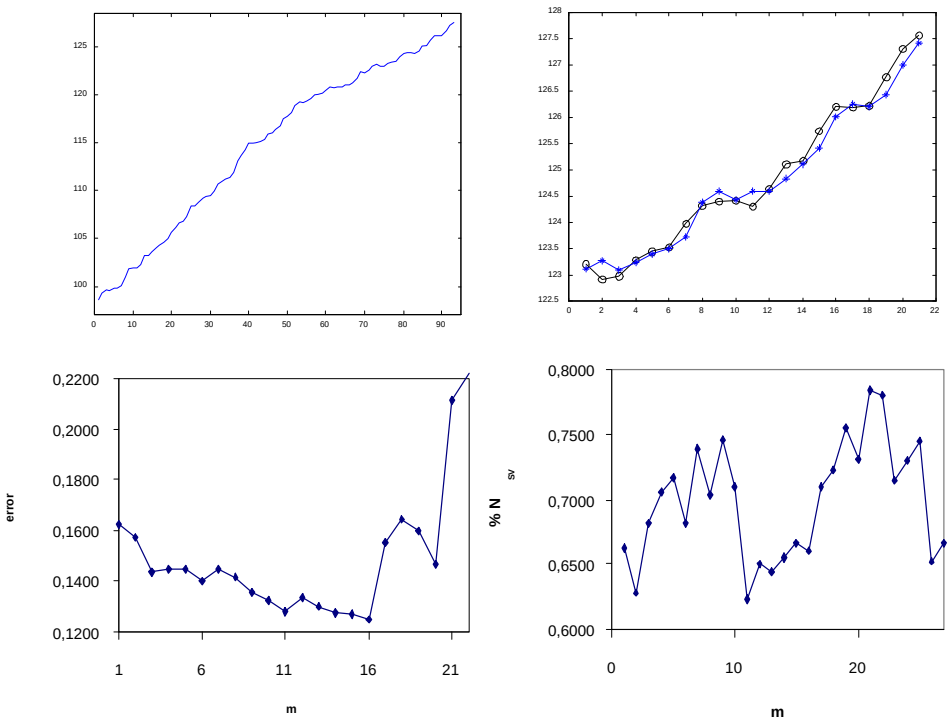

Fig. 1. *Top left corner: Graph of the CPI series. Top right corner: Representation of the test values (∘) and its predictions (+) for m= 16. Bottom left corner: Graphic embedding dimension-error. Bottom right corner: Comparative between the percentage of supports vectors and the embedding dimension.*

References

1. S. Mukherjee, E. Osuna, and F. Girosi. Nonlinear prediction of chaotic time series using support vector machines. *IEEE Transactions on Neural Networks*, 1997.
2. A. J. Smola. *Learning with Kernels*. PhD thesis, Vom Fachbereich 13- Informatik der Technischen Universitat Berlin, 1998.
3. V.N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc, 1998.