4OR-Q J Oper Res (2016) 14:309–325
DOI 10.1007/s10288-016-0316-0

INDUSTRY

An application of support vector machines to sales forecasting under promotions

G. Di Pillo[1] · V. Latorre[1] · S. Lucidi[1] · E. Procacci[2]

Received: 23 December 2015 / Published online: 29 April 2016

Abstract This paper deals with sales forecasting of a given commodity in a retail store of large distribution. For many years statistical methods such as ARIMA and Exponential Smoothing have been used to this aim. However the statistical methods could fail if high irregularity of sales are present, as happens for instance in case of promotions, because they are not well suited to model the nonlinear behaviors of the sales process. In recent years new methods based on machine learning are being employed for forecasting applications. A preliminary investigation indicates that methods based on the support vector machine (SVM) are more promising than other machine learning methods for the case considered. The paper assesses the application of SVM to sales forecasting under promotion impacts, compares SVM with other statistical methods, and tackles two real case studies.

Keywords Machine learning · Support vector machines · Sales forecasting · Promotion policies · Nonlinear optimization

Mathematics Subject Classification 62M20 · 68T05 · 90B05 · 90B60 · 90C30

This work was partially supported by ACT-OperationsResearch under contract 602/2010 on “Forecasting by Neural Networks and SVM”.

✉ G. Di Pillo
dipillo@dis.uniroma1.it

[1] Department of Computer Control and Management Engineering, Sapienza University of Rome, Via Ariosto 25, 00185 Rome, Italy

[2] ACT-OperationsResearch SRL, Via Nizza 45, 00198 Rome, Italy

1 Introduction

This paper is concerned with sales forecasting of a given commodity in a retail store of large distribution. In past times managers of these stores used their experience to predict the daily sales and to decide the resupply quantities. In more recent years, with the development of computer aided decision making, the use of mathematical methods has became more widespread. In years 70s and 80s the principal methods used were statistical methods based on time series autoregressive models, like the ARIMA method, the Box-Jenkins' method and the Winter's exponential smoothing method, see e.g. Makridakis et al. (1998). These methods perform the forecast by processing as input data samples of the same time series that one wants to forecast. If we consider the forecast as the output of the process, we can say that for these methods the input and the output pertain to the same time series.

Since the 90s the new mathematical model of Artificial Neural Network (ANN) was developed and employed also for forecasting applications, see e.g. Zhang et al. (1997). A neural network bases its prediction not only on samples of the time series that one wants to forecast, but also on samples of other input time series on which the output forecast may depend. These input series are called attributes of the output. The basic structure of an ANN is a multi-layer network of neurons, characterized by an activation function that depends on some parameters. Neurons are connected by weighted arcs. By properly tuning the parameters and weights, the ANN may become able to perform a forecast. An alternative characterization of the artificial neuron of an ANN is obtained, rather than in terms of activation functions, in terms of radial basis functions (RBF).

By the end of 90s, a mathematical model different than ANN was also developed for classification and forecasting, named Support Vector Machine (SVM), see e.g. Cristianini and Shawe-Taylor (2000). The analytical roots of SVM are in the Statistical Learning Theory, the algorithmic roots for its training are in the duality theory of Mathematical Programming. Also the SVM performs the forecast using samples of the time series that one wants to forecast, as well as samples of other input attributes, and since its introduction the SVM has been considered a valid competitor of the ANN in the same fields of application.

Multi-layer ANN, RBF ANN and SVM are tools for machine learning methods, based on a training process that, using given sets of input and output data, enables to forecast outputs corresponding to sets of input data not used for training. In all cases the training process is performed by solving mathematical optimization problems. In this way the machine learning method provides a surrogate model of a complex unknown phenomenon.

In this paper the complex phenomenon of concern is how the amount of sales of a given commodity depends on different suitable input attributes, and in particular on an abnormal input attribute, that is occurrence of promotions on sales.

There are several works in literature that deal with these issues. One of the first works dating to the 90s, Ansuj et al. (1996) showed the superiority of ANN on the ARIMA method in sales forecasting. In Alon et al. (2001) several comparisons are made between machine learning methods and statistical methods, showing from empirical results that machine learning methods have an edge on statistical methods especially in periods of volatile economic conditions. An extended evaluation on the use of neural networks in sales forecasting is reported in Crone and Preßmar (2006). Sales forecasts on a weekly basis using different inputs are obtained in Thiesing et al. (1995) and Thiesing and Vornberger (1997), proving again the efficacy of ANN. An application of ANN to retail sales of footwear is described in Das and Chaundhury (2007). As concerns SVM, their potential application in sales forecasting is dealt with in Levis and Papageorgiou (2005). Other works focus on the flexibility of machine learning methods. For example in Kuo and Xue (1998) fuzzy neural networks, and in Chang et al. (2009) both fuzzy neural networks and clustering methods, are used, to improve neural networks results. In Kuo et al. (2009) and Wu et al. (2008) particular optimization procedures are used, like genetic algorithms or swarm optimization, to improve the forecast and to obtain better results than the statistical methods. In a more general framework, see Cui and Curry (2005) and West et al. (1997), the authors use learning methods in the economical context of marketing for predicting consumer's future choices.

While there exists a wide literature on sales forecasting, the effect of promotions seems to be neglected, unless for some consumer behavioural aspects. On the other hand, it is of main importance in the marketing practice. As far as we know the effect of promotions in sales forecasting has been addressed explicitly only in Gür Ali et al. (2009), where the forecast is performed by aggregating at different levels stores and commodities, and in Trapero et al. (2015), in the context of the time series autoregressive models.

The lack of attention on the effects of promotions motivated a research effort dedicated in particular to approaches for sales forecasting based on machine learning, funded by ACT-OperationsResearch, an OR-oriented private company with branches in Italy, USA and UK, specialized, among others, in forecasting platforms.

In Di Pillo et al. (2013) we have done a preliminary assessment of machine learning methods in sales forecasting under promotions, by comparing the relative performances of Multi-layer ANN, RBF ANN and SVM. For the sake of brevity, in the present paper we focus on SVM, because the preliminary assessment indicates that SVM are most promising among them. With respect to Di Pillo et al. (2013) we also perform a wider experimentation. We point out that we do not intend to develop behavioral models on how consumers react to promotion, we confine ourselves to verify if the SVM is able to perform a reliable forecast of sales based on the considered input attributes.

The paper is organized as follows. In Sect. 2 we shortly describe the SVM employed for forecasting and the optimization problem to be solved in its training. In Sect. 3 we consider the implementation issues to be taken into account in the practical applications of the SVM. In Sect. 4 we describe the experimental environment of our application, making use of real sales data from two retail stores of large distribution. In Sect. 5 we report and analyze the results obtained in sales forecasting under promotion policies, making a comparison between the SVM and the statistical methods. Section 6 summarizes some concluding remarks.

As already mentioned this work has been funded by ACT-OperationsResearch, vendor of the multi-paradigm forecasting platform *Before!* and willing to improve the performance of the platform under difficult conditions, as it is the case in the presence of promotions. As a result of this work, the SVM is now fruitfully an integral part of the platform, in particular of the *Before! Promo Forecasting* module used to predict the demand for future promotion events associated with temporary price reductions and/or media and advertising pressure, see the URL http://www.act-operationsresearch.com for more details. Users of the platform in Italy are the companies in the large scale retail distribution COIN, CONAD and OVS; user of the platform in several European countries is the car rental company EUROPCAR.

2 Support vector machine

The SVM have been developed in the context of the Statistical Learning Theory, originally for solving classification problems [see e.g. Burges (1998), Cristianini and Shawe-Taylor (2000)]. The results obtained within this framework show that SVM has very good extrapolating properties, thus motivating the use of SVM for forecasting. In this section we describe shortly the mathematical model of the SVM and the related optimization problems solved in its training.

The optimization problem makes use of a training set

$$\mathcal{P} = \left\{ \left(x^p, y^p \right), x^p \in \Re^n, y^p \in \Re, \quad p = 1, \dots, P \right\},$$

where P is the cardinality of the set, and (x^p, y^p) is an input-output pair, a sample of the relation that we want to reproduce.

A *linear* SVM aims to realize a linear regression function

$$y(x) = w^T x + b$$

with the property that for each sample the regression error is bounded by a value $\epsilon \geq 0$ so that:

$$|y^p - w^T x^p - b| \leq \epsilon, \quad p = 1, \dots, P,$$

and with the property of being as much flat as possible, where flatness is measured by the squared norm of w. Therefore we are lead to the problem:

$$\begin{array}{l} \min_{w,b} \ \frac{1}{2}\|w\|^2 \\ |y^p - w^T x^p - b| \leq \epsilon, \, p = 1, \dots, P \end{array} \tag{1}$$

However, problem (1) could be infeasible. To tackle this possible failure, slack variables $\xi^p, \hat{\xi}^p, p = 1, \dots, P$ are introduced, and Problem (1) is modified as:

$$\begin{array}{l} \min_{w,b,\xi,\hat{\xi}} \ \frac{1}{2}\|w\|^2 + C \sum_{p=1}^{P} \left(\xi^p + \hat{\xi}^p \right) \\ w^T x^p + b - y^p \leq \epsilon + \xi^p \\ y^p - w^T x^p - b \leq \epsilon + \hat{\xi}^p \quad p = 1, \dots, P, \\ \xi^p, \quad \hat{\xi}^p \geq 0. \end{array} \tag{2}$$

where the second term in the objective function, with $C > 0$, provides a measure on how much the regression errors exceed the value ϵ.

Problem (2) is a quadratic convex problem in the variables $w, b, \xi, \hat{\xi}$, and therefore the solution can be found by solving its Wolfe dual problem, which is easier to be solved. Denoting by $\lambda^p, \hat{\lambda}^p, p = 1, \dots, P$ the dual variables corresponding to the Lagrange multipliers associated with the constraints on the regression errors, and by $\lambda, \hat{\lambda}$ the vectors with components $\lambda^p, \hat{\lambda}^p,\ p = 1, \dots, P$, the dual problem is obtained as:

$$\begin{aligned} \min \Gamma(\lambda, \hat{\lambda}) = \tfrac{1}{2} \sum_{p=1}^{P} \sum_{q=1}^{P} & \left(\hat{\lambda}^p - \lambda^p\right)\left(\hat{\lambda}^q - \lambda^q\right)(x^p)^T x^q \\ & - \sum_{p=1}^{P} \left(\hat{\lambda}^p - \lambda^p\right) y^p + \epsilon \sum_{p=1}^{P} \left(\hat{\lambda}^p + \lambda^p\right) \\ & \sum_{p=1}^{P} \left(\hat{\lambda}^p - \lambda^p\right) = 0 \\ & 0 \le \lambda^p \le C \quad p = 1, \dots, P \\ & 0 \le \hat{\lambda}^p \le C \quad p = 1, \dots, P. \end{aligned} \tag{3}$$

The structure of Problem (3) is of main interest, because it can be exploited for generalizing the linear SVM to the *nonlinear* SVM. To this aim it is sufficient to substitute the inner product $(x^p)^T x^q$ with the value $k(x^p, x^q)$ given by a suitable kernel function $k : \Re^n \times \Re^n \rightarrow \Re$. Since we will make use of nonlinear SVM, the problem of concern becomes the following:

$$\begin{aligned} \min \Gamma(\lambda, \hat{\lambda}) = \tfrac{1}{2} \sum_{p=1}^{P} \sum_{q=1}^{P} & \left(\hat{\lambda}^p - \lambda^p\right)\left(\hat{\lambda}^q - \lambda^q\right) k(x^p, x^q) \\ & - \sum_{p=1}^{P} \left(\hat{\lambda}^p - \lambda^p\right) y^p + \epsilon \sum_{p=1}^{P} \left(\hat{\lambda}^p + \lambda^p\right) \\ & \sum_{p=1}^{P} \left(\hat{\lambda}^p - \lambda^p\right) = 0 \\ & 0 \le \lambda^p \le C \qquad p = 1, \dots, P \\ & 0 \le \hat{\lambda}^p \le C \qquad p = 1, \dots, P. \end{aligned} \tag{4}$$

Problem (4) is a quadratic convex problem in the unknowns $\lambda, \hat{\lambda}$. Once solved, with solution $\lambda^*, \hat{\lambda}^*$, the nonlinear regression function $y(x)$ for the set of input-output samples $\mathcal{P}$ is given by

$$y(x) = \sum_{p=1}^{P} \left(\left(\hat{\lambda}^p\right)^* - \left(\lambda^p\right)^*\right) k\left(x, x^p\right) + b^*, \tag{5}$$

where b^* can be determined making use of the complementarity condition.

Here we adopt, as kernel function, the commonly used Gaussian Radial Basis Function (RBF) kernel:

$$k(x^p, x^q) = \exp\left(-\sigma \|x^p - x^q\|^2\right),$$

where $\sigma > 0$ is the kernel parameter. This choice is due to the fact that the Gaussian RBF kernel performed better than other kernel functions, in some preliminary testing.

The parameters ϵ, C that appears in Problem (4) and the kernel parameter σ affect the regression error, and are to be tuned in the validation phase of the learning process.

3 Implementation issues

The development of a SVM for forecasting requires the availability of:

- a data set,
- an optimization procedure,
- a validation procedure.

In this section we will shortly describe the three items.

3.1 Data set

The data set is the set of available input-output samples $\{(x^p, y^p),\ x^p \in \Re^n, y^p \in \Re,\ p = 1, \dots, R\}$, where R is usually very large. It must be partitioned into two subsets:

- the training and validation set $\mathcal{P} = \{(x^p, y^p),\ p = 1, \dots, P\}$ used by the learning procedure in the optimization and validation phases,
- a test set $\mathcal{S} = \{(x^p, y^p),\ p = P + 1, \dots, R\}$ used for measuring the quality of forecast produced by the resulting SVM within the data set.

In turn, the set $\mathcal{P}$ is partitioned into a subset $\mathcal{P}_T$ used by the optimization procedure for training and a subset $\mathcal{P}_\mathcal{V}$ used by the validation procedure. Different partitions of $\mathcal{P}$ are used in the validation procedure as described in the following.

Let x^p an input value, y^p the corresponding output and $y(x^p)$ the value predicted by the machine after the learning and validation procedures. Then the test set $\mathcal{S}$ is used to compute the mean square error $MSE(\mathcal{S})$ value:

$$MSE(\mathcal{S}) = \frac{1}{(R-P)} \sum_{p=P+1}^{R} \left(y(x^p) - y^p\right)^2,$$

which provides an overall measure of the quality of the forecast.

3.2 Optimization procedure

As shown before, the training of a SVM turns out to be mainly an optimization problem, with two significant features: the first one is that the problem is usually of very large scale, the second one is that it is a convex problem of particular structure, so that the second feature may in some way balance the first one. The fact that the optimization problem is a structured convex problem is one of the reasons why we focus on SVM, rather than on others machine learning methods, which require the solution of non-convex optimization problems. Indeed very specialized, and therefore efficient, algorithms have been proposed for the constrained optimization problem (4) of SVM training. In our application we have used the well known algorithm LIBSVM available through Chang and Lin (2014), making reference to the subset of input-output pairs $\mathcal{P}_T \subset \mathcal{P}$ not used for validation.

3.3 Validation

In the optimization problem of the SVM to be solved in the training procedure, we have to give values to the parameter ϵ that bounds the regression errors, to the parameter C which weights the fact that the regression errors exceed the value ϵ and to the parameter σ which appears in the kernel function. In particular, given ϵ, increasing C has the effect of making more regular the output function realized by the SVM; however, if exaggerated, it produces the trouble of over-fitting, that is the machine interpolates very well the training samples, but becomes inefficient on the samples of the test set, since it looses its generalization proprieties with respect to samples not in the training set.

The validation procedure aims to determine the values of the parameters that appears in the optimization model so as to obtain the best performances in forecasting. The validation procedure is performed by using the $MSE(\mathcal{P}_V)$ value, defined in a way similar to the $MSE(\mathcal{S})$, with reference to the validation set rather than to the test set:

$$MSE\,(\mathcal{P}_V) = \frac{1}{|\mathcal{P}_V|} \sum_{(x^p, y^p) \in \mathcal{P}_V} \left(y(x^p) - y^p \right)^2 .$$

The SVM is trained with different values of the parameters, and the corresponding values of the $MSE(\mathcal{P}_V)$ are evaluated. Then the values of the parameters which determine the smallest value of the $MSE(\mathcal{P}_V)$ are selected for the resulting SVM. However, in order to make the prediction less dependent on the training subset given in input to the optimization procedure, we adopt a *k-fold strategy* for validation. We divide the set of samples $\mathcal{P}$ into k different disjointed subsets of equal size and use $k-1$ of these subsets as $\mathcal{P}_{\mathcal{T}}$ for training and the remaining subset as $\mathcal{P}_{\mathcal{V}}$ to compute the *MSE* for validation. This procedure is repeated k times, every time using a different subset of $\mathcal{P}$ as validation set. Once all the partitions are used for validation and the values of the *MSE* for all the k-partitions are computed, the goodness of a particular choice of the values of the model parameters is determined by computing the average of its *MSE* on the k different partitions. Then the particular choice of values that determines the smallest value of the average *MSE* is considered to be optimal. In our application we set $k = 10$, a choice often used in literature.

After selecting the best parameter values the SVM is retrained for a last time, with all the k partitions in the training set. Then the test set $\mathcal{S}$, is used to evaluate the performance of the SVM using samples never used before.

In our application the particular choices of the SVM parameter values are taken as suggested in the user guide of LIBSVM, that is: as concerns ϵ, the fixed value $\epsilon = 0.1$ is used, as concerns C and σ a grid search is performed, with $C = 2^j$, $j = -3, \dots, 10$ and $\sigma = 2^j$, $j = -15, \dots, 3$.

4 Experimental environment

In this section we describe how the SVM has been used for sales forecasting. In our application we used two input-output time series provided by ACT-OperationsResearch, taken from the sales receipts of two different retail stores of the same chain of large distribution; the two stores are characterized by different volumes of sales.

As concerns the output y we are interested in the daily sales of a particular kind of pasta of a popular brand, measured by the number of items sold; as concerns the input vector x, we will describe below which attributes have been considered. In particular we are interested in capturing the effects of promotion policies on the sales. The input-output samples used for training and testing cover the five years 2007–2011. In particular, the years 2007–2010 have been used only for training and validation, the year 2011 for training, validation and testing in a sliding window approach, as described in the following. The first time series is taken from retail store #1, the one with the smaller volume of sales. The second series is taken from store #2, the one with the larger volume of sales. To have an idea of the sales trend, we can look at Figs. 7, 8, 9, 10, 11 and 12 where the sales of store #1 and store #2 in the year 2011 are drawn, together with their forecasts. It appears clearly that sales increase during promotion periods, that have been 5 during the year.

In our forecasting we will use as input attributes subsets of the following set of 13 attributes:

- 9 calendar attributes, linked to the specific day in which the output is given: month, day of the month and day of the week. The day of the week is represented by 7 mutually exclusive boolean attributes. These attributes bring into consideration typical human behaviors and customs. For example in Saturday it is expected to sell more than in the other days of the week;
- 4 problem specific attributes: one boolean attribute whose value is one if there is promotion of the product in that day, zero otherwise, number of hours the store is open that day and the daily price of the product; moreover the overall number of sale receipts released that day in the store for purchase of any kind of goods, which accounts for how many consumers entered the store in the same day.

As concerns the last attribute listed before, that is the number of receipts released in the same day for which the sales forecast is done, we point out that of course its value is not known. Therefore we implemented a SVM for forecasting the number of receipts per day. This SVM uses the 2007–2010 series for training and validation. Then we used this SVM to produce a forecast of the number of receipts in 2011, and for testing the quality of the forecast by comparing with the actual data. This SVM uses in input 11 attributes: the 9 calendar attributes also used in forecasting sales, the number of hours the store is open and a last attribute that indicates if in that day the expected number of sales is high or low. This attribute is 0 in normal days, 1 in days before festivities, -1 when the store is open on Sundays and 2 on the day of Christmas eve and new year eve. Actually the SVM for the receipts forecast in 2011 is the result of training and validating 12 SVM, each one for one month of the year: that is the samples of January 2007–2010 are used for building a SVM able to forecast the number of receipts in January 2011, and so on.

We are aware that a forecast attribute can be considered a risky choice for the robustness of the final predictive model. However as we will see, this attribute turns out to be effective since it can be used in place of the calendar attributes in order to smooth the curse of dimensionality which affects the training of SVM; moreover it is of interest in itself for the management of the store.

In sales forecasting we realized two experiments changing the attributes in input:

- in the first experiment we use 4 inputs: promotion, number of opening hours, price of the product and number of daily receipts (forecast);
- in the second experiment we use 12 inputs: promotion, number of opening hours, price of the product and the nine calendar attributes.

We are interested in a time horizon for the forecast of the next four weeks, suitable for resupply policies. Therefore, since we deal with a short term forecasting we adopt the *sliding window* approach often used in this kind of applications, see e.g. Bo et al. (2006).

We make reference to the time series of store #1 in order to explain how we proceeded. We use the samples of the year 2011 for forecasting and testing; by eliminating one day in which the store was closed, the year has been divided into 13 time intervals, each one containing four weeks worth of data, 28 input-output daily samples. The 13 sets of input-output samples used for testing are denoted by $\mathcal{S}_i, i = 1, \dots, 13$. The samples of the years 2007–2010 are used first for training and cross-validation, and a first SVM is built; then we use the SVM to perform the forecast of the output samples in the first time interval, and we measure the quality of the forecast by the $MSE(\mathcal{S}_1)$ value. Then we add the set $\mathcal{S}_1$ to the training and validation set, we train and validate a new SVM using also the new set of samples, we perform the forecast of the output samples in the second interval $\mathcal{S}_2$, and we measure the $MSE(\mathcal{S}_2)$ value. The procedure is repeated, until we reach the last interval of the year 2011. In this way we train 13 different and independent SVMs, enlarging each time the training and validation set. The time series of store #2 has been treated in the same way.

By adopting the sliding window method the forecast in each time interval makes use of the updated input-output data of all previous time intervals.

5 Computational results

In this section we report the results obtained in forecasting the sales during 2011, and we make a comparison with the forecasts provided by traditional statistical methods. In particular, for each store we run five computations, two for the SVM by using the two different configuration of input attributes, and three for statistical methods, the first method ARIMA being the auto-regressive moving average, the second one ES being the exponential smoothing, and the third one HWES being the Holt-Winter triple exponential smoothing.

The results of the statistical methods are obtained by using the corresponding tools of the R Statistical Computing Environment, available in the *forecast* package of the CRAN (http://cran.r-project.org/) repository, in the default setting. More in particular:

- ARIMA is the forecast obtained using the shell *forecast.arima*, using the function *auto.arima*;
- ES is obtained using the shell *forecast.ets*, using the function *ets*;
- HWES is obtained using the shell *forecast.HoltWinters*, using the function *HoltWinters*.

We point out that in the default setting all time series models and parameters are automatically optimized by the corresponding tools.

5.1 Forecast of daily sales receipts

Preliminarily we show the results obtained using a SVM for forecasting the number of daily receipts in 2011, used as input attribute for the sales forecast of concern here. As already said, we used the samples of 2007–2010 for training and cross-validation, with the 11 input attributes listed in Sect. 3.

In Figs. 1, 2, 3, 4, 5 and 6 we draw the actual number of sales receipts and the number predicted by the SVM daily during 2011, for store #1 and store #2 respectively. From the figures we can observe that the forecast of daily receipts is quite satisfactory. The $MSE(2011)$ on the number of receipts was 8.329×10^3 and 38.982×10^3 respectively for store #1 and store #2. The large values of the MSE are due to large numbers of actual and forecast receipts. Indeed, if we evaluate the Mean Absolute Percentage Error ($MAPE$) defined as

$$MAPE(\mathcal{S}) = \frac{100}{(R-P)} \sum_{p=P+1}^{R} \frac{|y(x^p) - y^p)|}{y^p},$$

we find values 4.5 and 5.4 % respectively for store #1 and store #2.

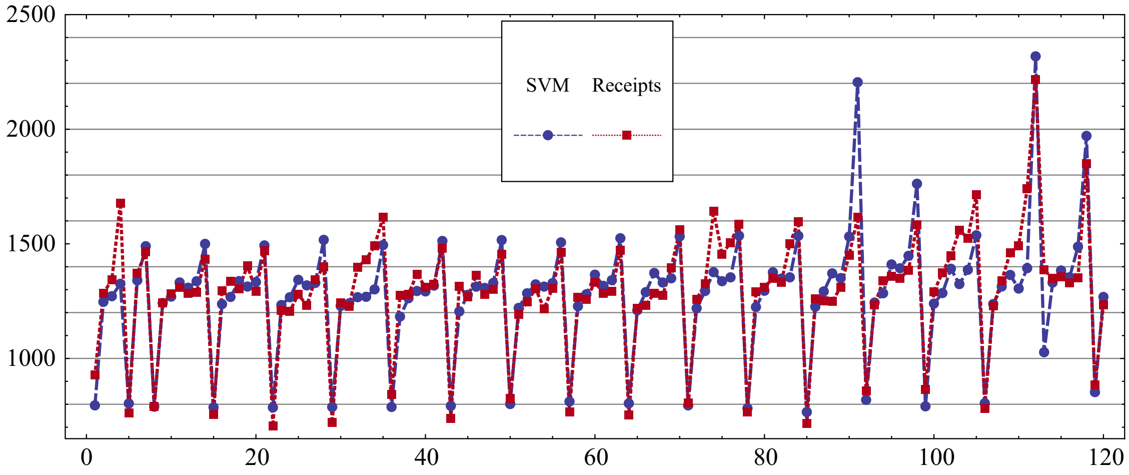

Fig. 1 Daily receipts forecast for store #1 from January to April 2011

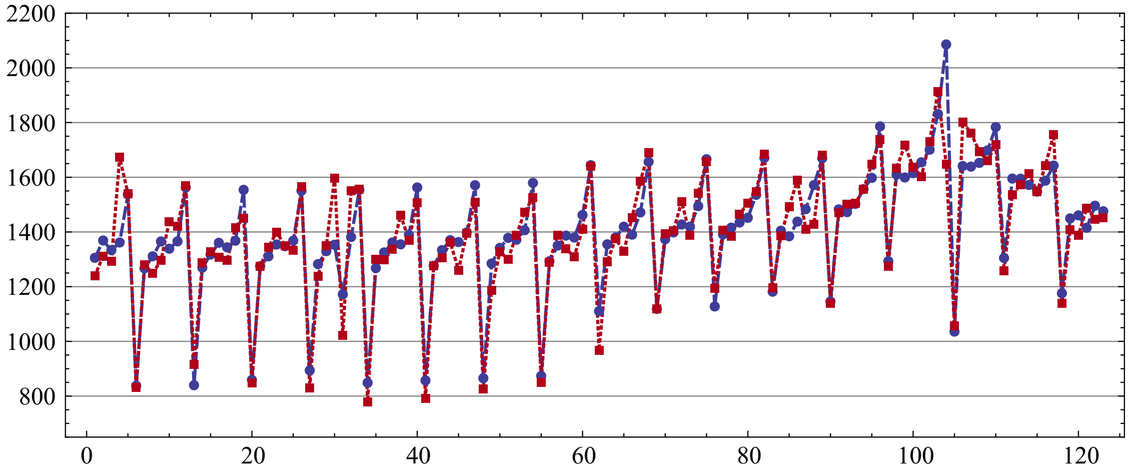

Fig. 2 Daily receipts forecast for store #1 from May to August 2011

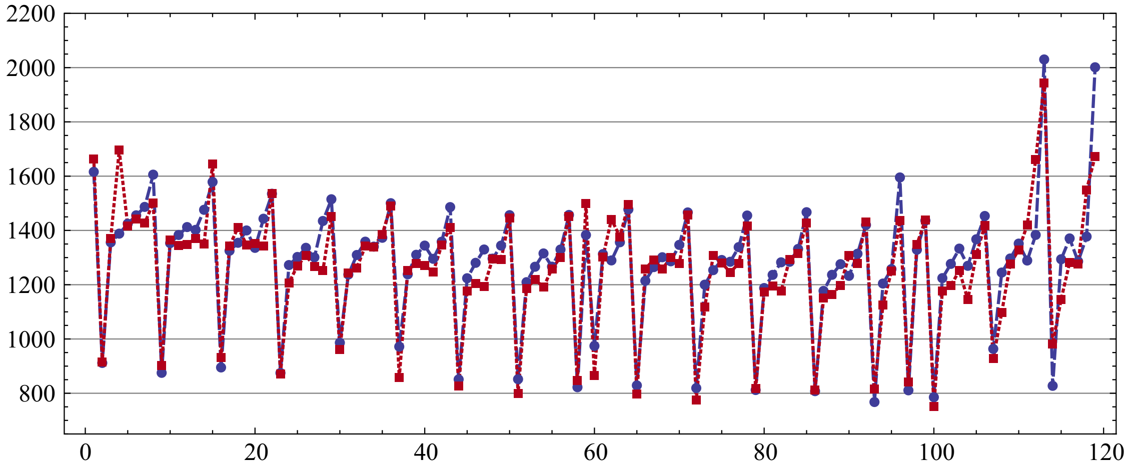

Fig. 3 Daily receipts forecast for store #1 from September to December 2011

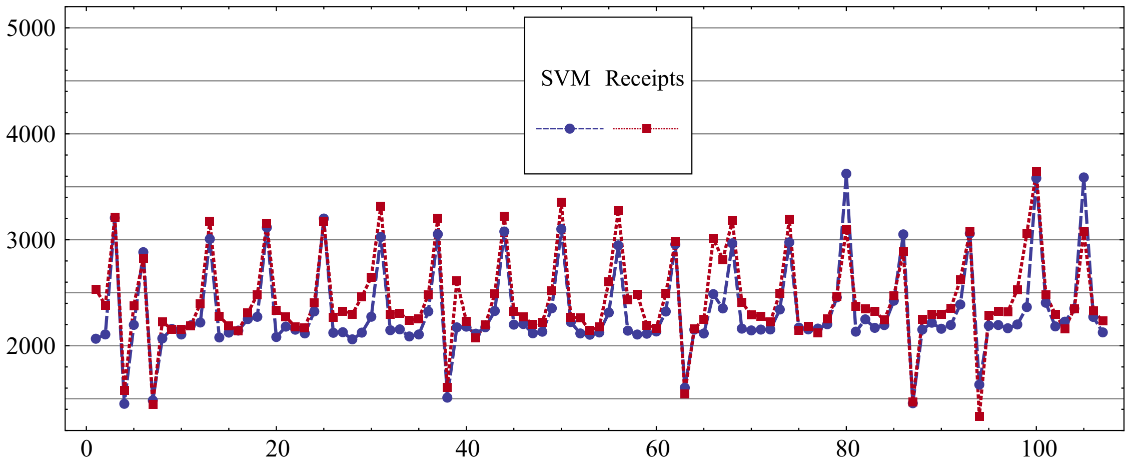

Fig. 4 Daily receipts forecast for store #2 from January to April 2011

5.2 Sales forecast in store #1

The results obtained after training and cross-validating the SVM, with two different configuration of input attributes, denoted by $4i$, $12i$, are given in terms of $MSE(\mathcal{S}_i)$, $i =$

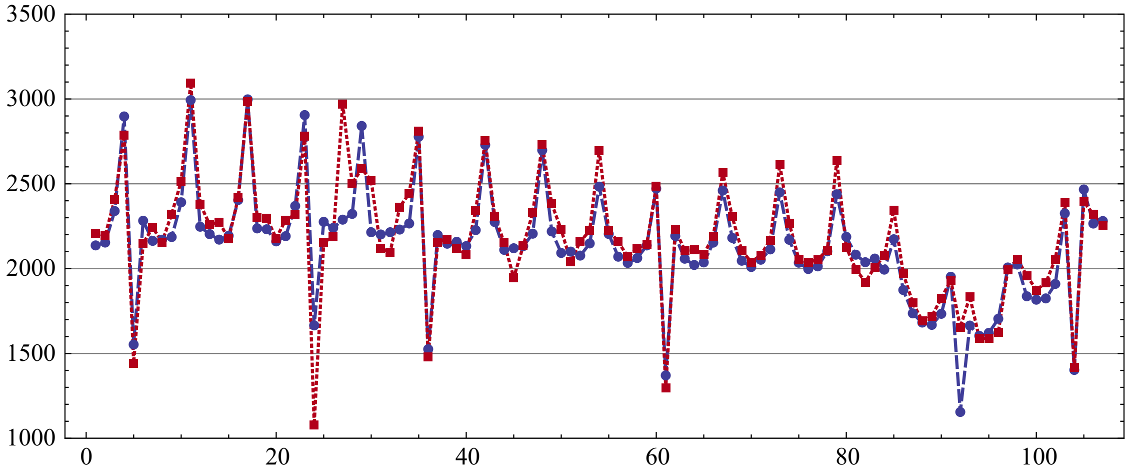

Fig. 5 Daily receipts forecast for store #2 from May to August 2011

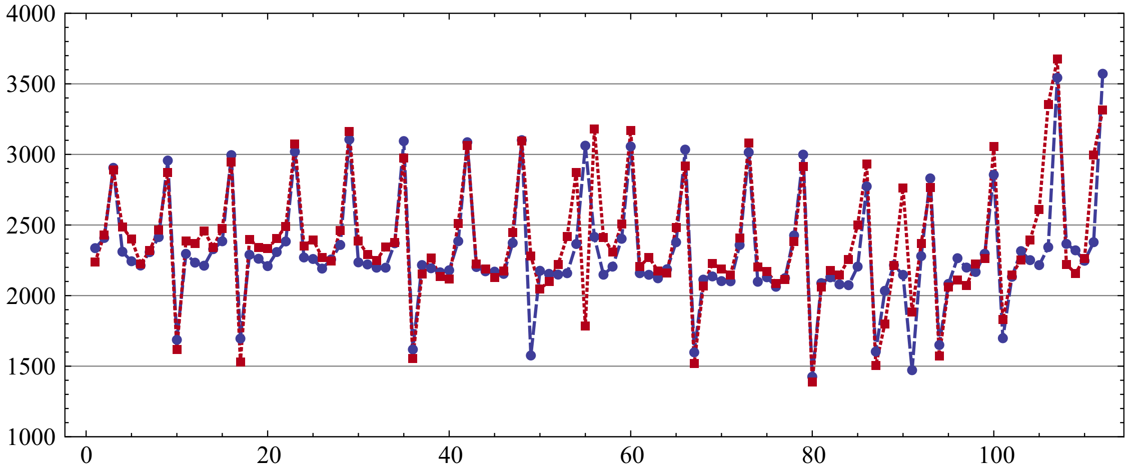

Fig. 6 Daily receipts forecast for store #2 from September to December 2011

$1, \ldots, 13$ in Table 1. In the same table are given the $MSE(\mathcal{S}_i)$ values obtained by using the three statistical methods. All results refer to the test set $\mathcal{S}$ of input-output data of 2011, not used for training and validation.

From the table we notice that the smallest mean value of the $MSE(\mathcal{S}_i)$ is given by the SVM 4*i*, and that the SVM 12*i* gives a mean value smaller than those of ES and HWES. Among the statistical methods, the smallest mean value of the $MSE(\mathcal{S}_i)$ is given by ARIMA. If we take a look to the results interval by interval, the SVM 12*i* is able to get the best forecast for five intervals on 13, while the SVM 4*i* and ARIMA give the best forecast for three intervals on 13. It is evident how the error increases in intervals corresponding to promotion periods, like for instance intervals 5, 7, 10.

In Figs. 7, 8 and 9 we draw the actual sales in store #1 during 2011, and the forecasts produced by SVM 4*i* and ARIMA. Promotion periods, that have been five, are displayed by vertical dashed lines. It appears that mainly in these periods SVM 4*i* outperforms ARIMA. Indeed we notice, especially for the periods of promotion going from April 26 to May 9, from July 7 to July 20, and from September 22 to October 10, that the weekly trend is totally disrupted by the promotion effect. The output assumes values difficult to predict and with a high variability. The statistical method, even in

Table 1 $MSE(\mathcal{S}_i)$ for store #1

Interval	SVM4i	SVM 12i	ES	HWES	ARIMA
1	44.689	37.223	74.045	70.494	73.624
2	49.287	47.2	38.459	38.453	55.592
3	40.579	37.932	39.275	38.955	36.969
4	37.823	34.913	46.911	48.804	35.938
5	141.23	172.032	174.839	167.022	167.428
6	31.656	25.636	26.498	27.32	27.772
7	80.293	88.65	107.967	100.863	116.649
8	77.818	44.933	73.705	73.611	57.188
9	52.446	67.243	47.94	54.611	54.614
10	101.804	171.564	86.232	89.004	72.997
11	30.401	26.771	28.187	28.478	31.291
12	69.175	82.413	63.411	64.316	53.365
13	44.057	36.842	81.798	84.176	50.25
Average	61.751	67.355	68.406	68.134	64.31

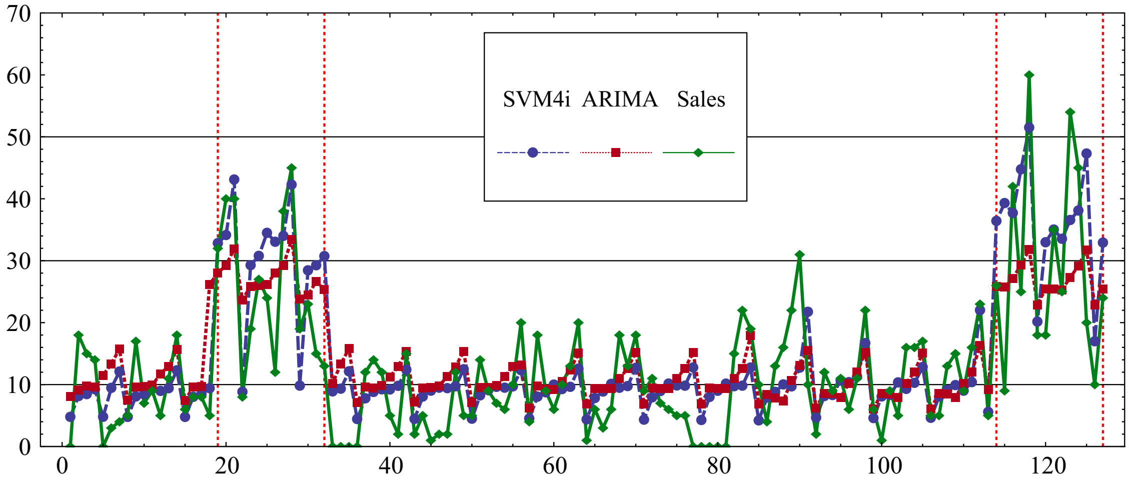

Fig. 7 Actual and forecast sales in store #1 from January to April 2011

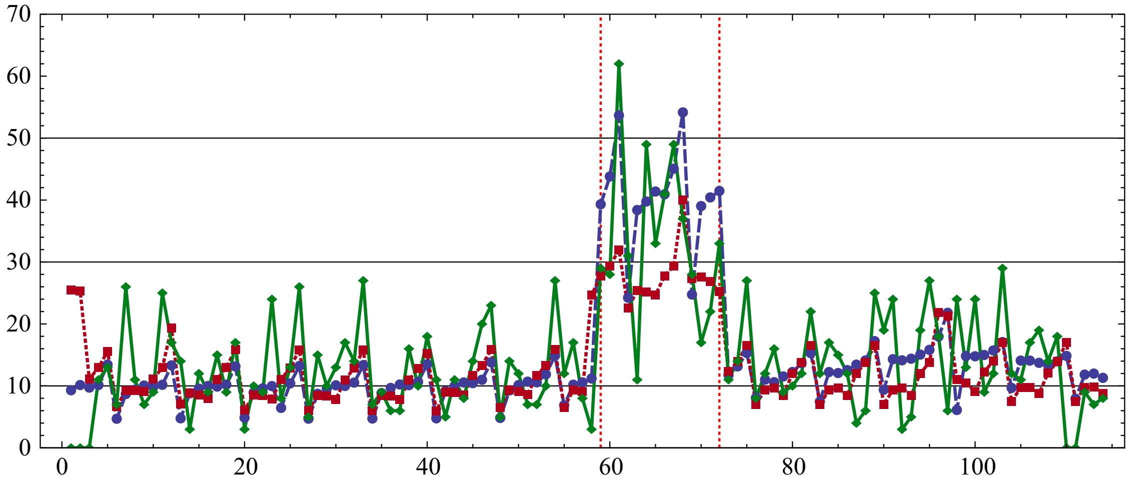

Fig. 8 Actual and forecast sales in store #1 from May to August 2011

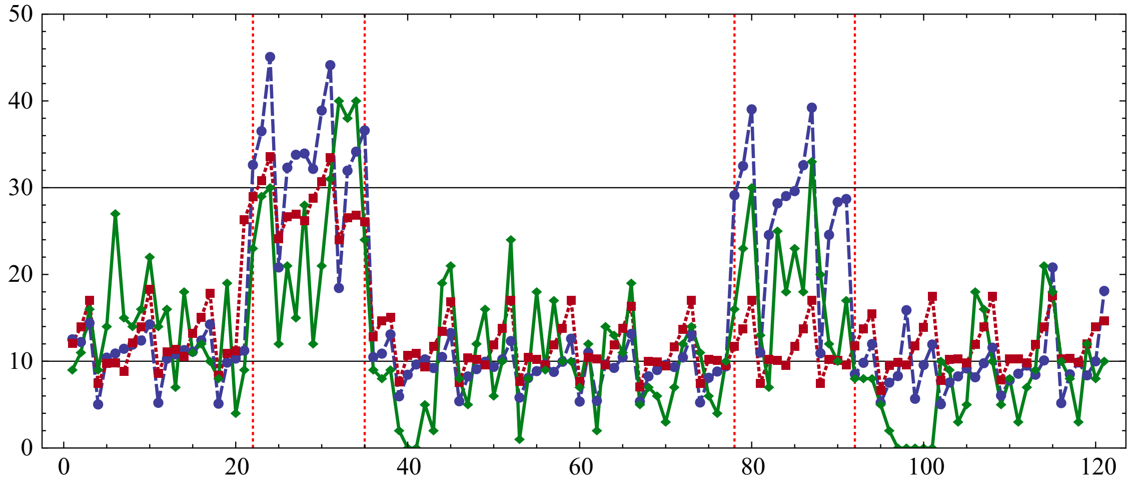

Fig. 9 Actual and forecast sales in store #1 from September to December 2011

Table 2 $MSE(\mathcal{S}_i)$ for store #2

Interval	SVM 4i	SVM 12i	ES	HWES	ARIMA
1	883.66	1232.07	1354.73	1390.98	1383.15
2	295.33	296.26	271.18	302.91	270.31
3	213.98	202.82	180.51	149.12	167.61
4	70.49	54.72	56.58	68.92	55.53
5	364.84	405.7	617.09	614.62	574.86
6	58.11	72.57	49.07	48.73	42.81
7	506.85	586.05	862.46	828.62	853.5
8	133.93	121.03	130.26	139.43	134.73
9	31.07	60.66	50.86	57.77	45.72
10	687.71	496.86	1262.27	1220.03	1177.41
11	100.98	129.08	79.54	107.4	65.85
12	185.67	264.01	805.75	843.36	817.82
13	139.92	109.5	123.48	164.07	120.1
Average	281.21	283.56	454.74	462.85	444.64

this periods of promotion, follows its weekly trend, while the SVM is able to catch a part of this significant variability.

5.3 Sales forecast in store #2

In Table 2 we report the results obtained for store #2, organized in the same way as in Table 1. From the table it appears that the best results are given again by SVM 4i, followed by SVM 12i; among the statistical methods the best results are given again by ARIMA. Looking at the intervals, we see that SVM 4i gives the best results in five on 13 intervals, and that SVM 12i gives the best results on four on 13 intervals.

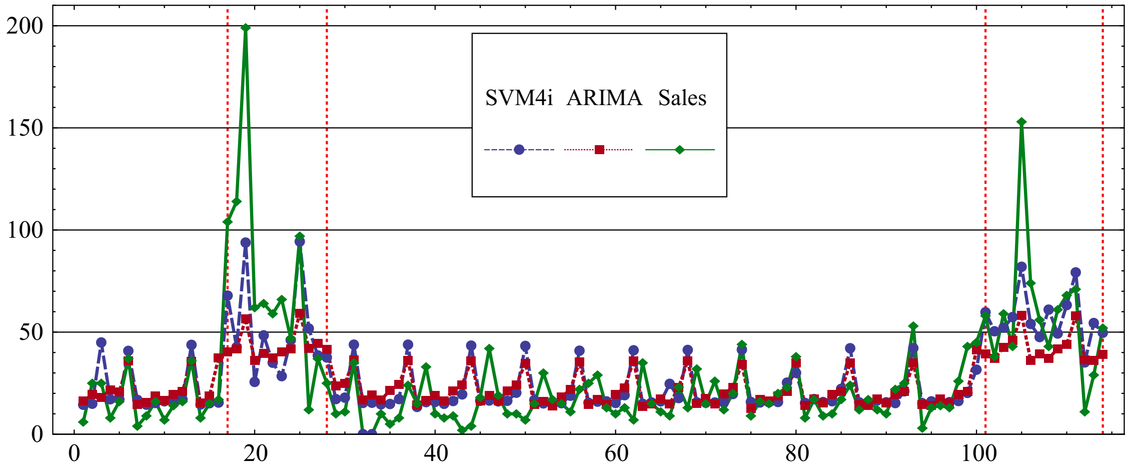

Fig. 10 Actual and forecast sales in store #2 from January to April 2011

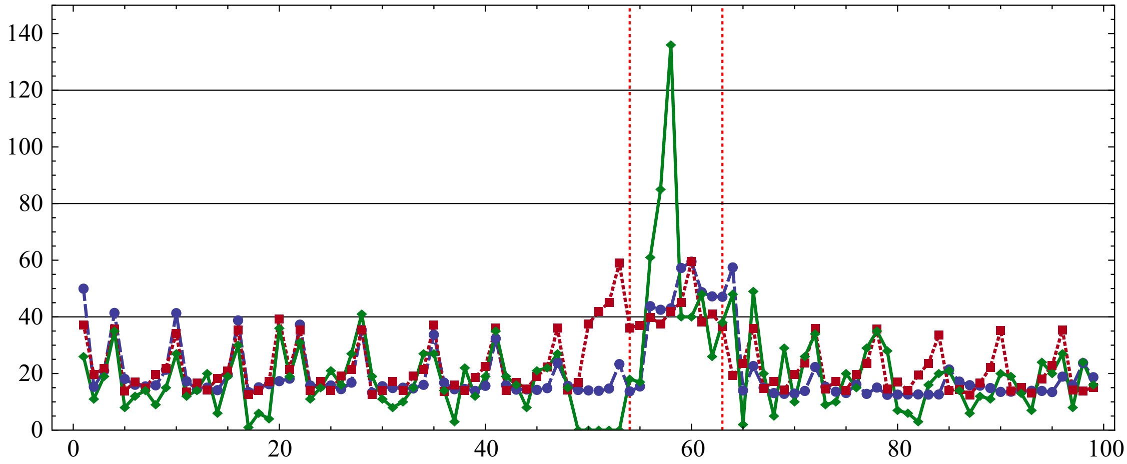

Fig. 11 Actual and forecast sales in store #2 from May to August 2011

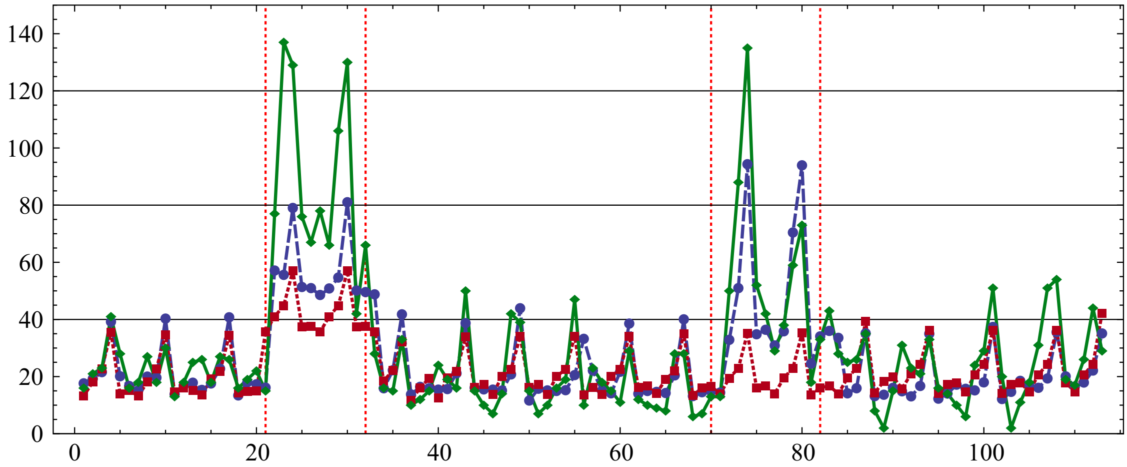

Fig. 12 Actual and forecast sales in store #2 from September to December 2011

In Figs. 10, 11 and 12 we draw the actual sales in store #2 during 2011, and the forecasts produced by SVM 4i and ARIMA. Promotion periods are again displayed by vertical dashed lines, and again it appears that in these periods SVM4i outperforms

ARIMA. Indeed, we can make remarks similar to those for store #1, as concerns the different way in which SVM and statistical methods react to promotion.

6 Conclusions

We have described how the SVM can be applied to sales forecasting. The application has concerned the daily sales of a kind of pasta in two retail stores, also in the presence of promotion. We have pointed out the importance of a suitable selection of input attributes for the machine, and the possibility of smoothing the curse of dimensionality by resorting to the forecast of an attribute which accounts for the calendar attributes. From the computational results we have shown that the SVM provides a valuable tool for sales forecasting, being the *MSE* values of the forecast smaller than the one produced by statistical methods. As a conclusion, we claim that any sales manager could take advantage by enlarging the class of methods employed for sales forecasting, so as to include, with the more traditional statistical methods, also the Support Vector Machine considered in this paper.

Compliance with ethical standards

We confirm that this submission complies with all ethical standards of 4OR. In particular we confirm that this work has been partially funded by ACT-OperatiosResearch SRL, contract 602/2010 on “Forecasting by Neural Networks and SVM”.

Author contribution All authors of this paper have directly participated in the planning, execution, or analysis of this study. All authors of this paper have read and approved the final version submitted. The contents of this manuscript have not been copyrighted or published previously. The contents of this manuscript are not now under consideration for publication else- where; The contents of this manuscript will not be copyrighted, submitted, or published else- where, while acceptance by the Journal is under consideration; Department representative is fully aware of this submission.

References

Alon I, Qi M, Sadowski RJ (2001) Forecasting aggregate retail sales: a comparison of artificial neural networks and traditional methods. J Retail Consum Serv 8:147–156

Ansuj AP, Camargo ME, Radharamanan R, Petry DG (1996) Sales forecasting using time series and neural networks. Comput Ind Eng 31:421–424

Bo C, Lu A, Wang Z, Zhang S (2006) Study and application on dynamic modeling method based on SVM and sliding time window techniques. In: Proceedings of the 6th world congress on intelligent control and automation. IEEE, pp 4714–4718

Burges CJC (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 2:121–167

Chang CC, Lin CJ (2014) LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm

Chang PC, Liu CH, Fan CY (2009) Data clustering and fuzzy neural network for sales forecasting: a case study in printed circuit board industry. Knowl Based Syst 22:344–355

Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge

Crone SF, Preßmar DB (2006) An extended evaluation framework for neural network publications in sales forecasting. In: Proceedings of the international conference on applied artificial intelligence AIA’06. ACTA Press, pp 179–186

Cui D, Curry D (2005) Prediction in marketing using the support vector machine. Mark Sci 24:595–615

Das P, Chaundhury S (2007) Prediction of retail sales of footwear using feedforward and recurrent neural networks. Neural Comput Appl 16:491–502

Di Pillo G, Latorre V, Lucidi S, Procacci E (2013) An application of learning machines to sales forecasting under promotions, Tech. Rep. Department of Computer Control and Management Engineering, Sapienza University of Rome, n. 4. http://www.dis.uniroma1.it/~bibdis/index2.php?option=com_docman&task=doc_view&gid=25&Itemid=34

Gür Ali Ö, Sayin S, van Woensel T, Fransoo J (2009) SKU demand forecasting in the presence of promotions. Expert Syst Appl 36:12340–12348

Kuo RJ, Hu TL, Chen ZY (2009) Application of radial basis function neural network for sales forecasting. In: Proceedings international Asia conference on informatics in control, automation and robotics. IEEE, pp 325–328

Kuo RJ, Xue KC (1998) A decision support system for sales forecasting through fuzzy neural networks with assymetric fuzzy weights. Decis Support Syst 24:105–126

Levis AA, Papageorgiou LG (2005) Customer demand forecasting via support vector regression analysis. Chem Eng Res Des 83:1009–1018

Makridakis S, Wheelwright SC, Hyndman RJ (1998) Forecasting: methods and applications. Wiley, New York

Thiesing FM, Middelberg U, Vornberger O (1995) Short term prediction of sales in supermarkets. In: Proceedings of the international conference on neural networks. IEEE, pp 1028–1031

Thiesing FM, Vornberger O (1997) Sales forecasting using neural networks. In: Proceedings of the international conference on neural networks. IEEE, pp 2125–2128

Trapero JR, Kourentzes N, Fildes R (2015) On the identification of sales forecasting models in the presence of promotions. J Oper Res Soc 66:299–307

West PM, Brockett PL, Golden LL (1997) A comparative analysis of neural networks and statistical methods for predicting consumer choice. Mark Sci 16:370–391

Wu Q, Yan HS, Yang HB (2008) A forecasting model based on support vector machine and particle swarm optimization. In: Proceedings of the workshop on power electronics and intelligent transportation systems. IEEE, pp 218–222

Zhang G, Patuwo BE, Hu MY (1997) Forecasting with artificial neural networks: the state of the art. Int J Forecast 14:35–62