# An Efficient and Simplified Model for Forecasting using SRM

HAFIZ MUHAMMAD SHAHZAD ASIF*, MUHAMMAD FAISAL HAYAT*, AND TAUQIR AHMAD*

RECEIVED ON 15.04.2013 ACCEPTED ON 09.01.2014

## ABSTRACT

**Learning form continuous financial systems play a vital role in enterprise operations. One of the most sophisticated non-parametric supervised learning classifiers, SVM (Support Vector Machines), provides robust and accurate results, however it may require intense computation and other resources. The heart of SLT (Statistical Learning Theory), SRM (Structural Risk Minimization )Principle can also be used for model selection. In this paper, we focus on comparing the performance of model estimation using SRM with SVR (Support Vector Regression) for forecasting the retail sales of consumer products. The potential benefits of an accurate sales forecasting technique in businesses are immense. Retail sales forecasting is an integral part of strategic business planning in areas such as sales planning, marketing research, pricing, production planning and scheduling. Performance comparison of support vector regression with model selection using SRM shows comparable results to SVR but in a computationally efficient manner. This research targeted the real life data to conclude the results after investigating the computer generated datasets for different types of model building.**

**Key Words:** **Structural Risk Minimization, Support Vector Regression, Statistical Learning Theory, Sales Forecasting, Support Vector Machine, VC Dimension.**

## 1. INTRODUCTION

The retail sale of manufactured products is a key index of business management in consumer intensive industries such as consumer electronics, automobile, oil and gas industry [1-2] and high risk angro-chemical and pharmaceutical industries [3]. To anticipate the sales of a product, customer demand acts as an important parameter along with several other factors. From a historical perspective, exponential smoothing methods and decomposition methods were the key forecasting approaches to be developed back in the mid-1950s.

More sophisticated automated forecasting methods appeared during the 1960s, as computational resources became more available and cheaper.

Box-Jenckins methodology gave rise to the ARIMA (Auto Regressive Integrated Moving Average) models [4]. During the 1970s and 1980s, sophisticated forecasting approaches were developed including econometric methods and Bayesian methods[5]. ANN (Artificial Neural Networks) emerged as a promising forecasting approach in the 1990s for forecasting [6].

An important observation during the analysis of the existing forecasting techniques is that the increase in complexity of forecasting approaches does not always result in the increase predictive accuracy as pointed out by Makridakis and Hibon [7]. Therefore the objective of new forecasting methods is not only to improve the accuracy but also to achieve the results with minimum resources.

* Assistnat Professor, Department of Computer Science & Engineering, University of Engineering & Technology, Lahore.

The risk involved in the forecasting may result in poor performance and may cause negative implication on the business management [8]. This risk in forecasting is due to the specific nature of forecasting [9]. Another factor that may effect the forecasting adversely is environmental instability [10-11]. Almost all forecasting technique has pros and cons and the technique that gives minimum forecast error must be selected [12]. Motivation behind the accurate sales forecasting is to reduce the cost involved. Research in this area shows that 10% or more of the total profit is lost due to the forecast errors that is caused because of the overages and shortages of the product [13]. Furthermore, forecast errors adversely effects the purchasing, financing and scheduling [12] resulting is lots of hidden losses.

It is, therefore, obvious to have short term forecasting plans with more reliability and higher performance. The forecasting strategies should be checked on regular short term basis and necessary changes should be made routinely in order to achieve good results and to increase the profitability. This is only possible if the model building techniques are simple to understand  easy to implement and computational efficient.

## 1.1 Statistical Learning Theory

SLT is considered to be one of the best available formalized theories for finite-sample inductive learning [14]. The main goal of SLT is to provide a framework for studying the problem of inference that is constructing models, gaining knowledge and making predictions from a finite set of data. Predictive learning from finite data is the fundamental task of machine-learning [15]. The goal in the predictive learning is to derive an unknown input-output dependency or structure of a system using limited number of observations. The task of selecting a mathematical model from a set of potential models, that best represents the system, is called model selection. SLT that is based on the SRM principle can be used for model selection using a finite set of data.

The SRM principle stems from the idea that in order to achieve optimal generalization performance, it is necessary to find a trade-off between the quality of the approximation of the given data and the complexity of the approximating model. To achieve this, SRM principle employs the concept of nested structure with increasing complexity i.e.

$$S_1 \subset S_2 \subset \ldots \subset S_k \subset \ldots$$

Where $S_1$, $S_2$,…,$S_k$ are the models with increasing complexity.

By definition, a structure provides ordering of its elements according to their complexity (i.e. VC-Dimension) [16-17].

$$h_1 < h_2 \ldots < h_k \ldots$$

The objective is to select the structure with optimal complexity that captures the inherent trends in data. The SRM principle does not specify any information about the structure to be used. Practically, successful application of SRM principle may depend on a selected structure. In order to measure the learnability of a set of models, the SRM principle employs the concept of VC dimension. In general, the VC dimension of a function is a natural number, possibly infinite, which represents the largest number of training points that can be shattered by that function [18].

The mathematical form of SRM is described by the following set of equations. Classical regression formulation of the SRM principle is given by [15]:

$$R(\alpha) \le \frac{R_{emp}(\alpha)}{\left(1 - c\sqrt{\varepsilon}\right)_+} \tag{1}$$

Where $R(\alpha)$ is the expected risk (prediction risk) and $R_{emp}$ is the empirical risk (loss function commonly used for regression problems) is given by [16-17]:

$$R_{emp}(\alpha) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i, \alpha))^2 \tag{2}$$

And $\varepsilon$ is given by [15]

$$\varepsilon = \varepsilon\left(\frac{n}{h}, \frac{-\ln\eta}{n}\right) = a_1 \frac{h\left(\ln\frac{a_2 n}{h}+1\right)-\ln(\eta/4)}{n} \quad (3)$$

And

$$\eta = \min\left(\frac{4}{\sqrt{n}}, 1\right) \quad (4)$$

Where $h$ is the VC-dimension of the set of approximating functions, $\alpha_1$ and $\alpha_2$ are theoretical constants. The values of these theoretical constants and the confidence level $1-\eta$ must be set for practical use of the VC generalization bound in model selection [15]. Substituting values for these constants in Equation (1), we get [14]:

$$R(h) \le R_{emp}(h)\left(1-\sqrt{p - p\ln p + \frac{\ln n}{2n}}\right)_+^{-1} \quad (5)$$

Where $p=h/n$. The practical VC-bound Equation (2) contains VC penalization factor [15], also called Vapnik's measure ($vm$):

$$r(p,n) = \left(1-\sqrt{p - p\ln p + \frac{\ln n}{2n}}\right)_+^{-1} \quad (6)$$

Penalization factor has been used for VC-based complexity control linear estimators, i.e. algebraic polynomials of degree $m$, VC-dimension is $h = m+1$.

Several empirical comparisons suggest that Vapnik's measure provides superior model selection than classical analytic model selection for linear regression problems [18-20].

## 1.2 Support Vector Regression

SVMs employ the notion of implicit mapping via kernels. SVMs when applied to the regression problem incorporates a different loss function[21] than that of classification problem. The loss function is similar to standard least square method. Different loss functions are available for SVR [22]. Regression problems can be categorized into linearly separable and linearly non-separable problems. Linearly separable solution of the SVR is obtained by minimizing the following functional [22]:

$$f(\omega,\xi) = \frac{1}{2}\left\|w^2\right\| + C\sum_i (\xi^- + \xi^+) \quad (7)$$

Where $C$ is a constant, $\xi^-$ and $\xi^+$ are the parameters controlling the behavior of the system $f(\omega,\xi)$.

For non-linearly separable case, kernels are employed to overcome the problem of curse of dimensionality. The regression function for support vector estimation is given by [22]:

$$f(x) = \sum_{i=1}^{l} (\overline{\alpha_i} - \overline{\alpha_i^*}) K(x_i, x) + \bar{b} \quad (8)$$

Where the kernel $K$ computes the dot product between the input patterns [23] $k(x,x_i)=\{\phi(x),\phi(x_i)\}$ while overcoming curse of dimensionality. Kernel basically transforms the input patterns to into some feature space $\Im$ given by $\phi:\chi\rightarrow\Im$ [23]. Using kernels, we reduce the problem complexity and then apply the SVR.

## 1.3 Model Selection via Statistical Learning Theory

Model selection is the essential goal of machine learning. The problem of selection among competing models has been a fundamental issue in this regard. The techniques available today for model estimation of Polynomial based models include FPE (Finite Prediction Error) [24], AIC (Akaike's Information Criterion) [25], SCH (Schwartz's Criterion) [26] and GCV (Generalized Cross Validation) [27]. Wallace's MML (Minimum Message Length) principle [28-30] and Vapnik's SRM [16, 17, 31] based on the classical Vapnik Chervonenkis theory of VC dimensionality.

Amongst these generic techniques, SRM and MML can be applied to any family of problems [20]. Both of these generic techniques define a trade off between the complexity of a given model and its suitability to the data being analyzed [29]. AIC and SCH address the number of free parameters which is a prominent difference among models [24-25]. The increase in number of free parameters result in the model to provide a better fit to the data being observed [17]. The best fit obtained with the extra parameters is required to provide the justification for the necessity of the extra parameters in incorporating the intricacies of the underlying system. GCV, MML, and SRM are sensitive to both model's functional form i.e. selection of the set of hypothesis and the model complexity [16-17,31].

The SRM principle is embodied in SVM and is capable enough to be applied to model selection in a regression framework. SRM is built on classical learning paradigm which provides simple formulation and increased flexibility [18]. The effectiveness of the SRM is mainly due to the solid theoretical basis and the practical applications in different areas of scientific research [15].

Regression using polynomials from noisy data may result in to the phenomenon of overfitting. Different techniques have been developed in the literature to balance the complexity of model selection with their training error. An empirical evaluation [19] gives detailed comparison of the performance of different such methods in a classical polynomial regression problem that includes the MML Principle [24], Vapnik's SRM [17], FPE [24], SCH [26] and GCV [28]. The results from the above mentioned empirical evaluation builds strong basis in the favor of the MML and SRM methods over the other techniques [14-15,19].

## 1.4 Research Focus

We present a solution to the problem of model selection from real world sales data of a product (home appliance) over a period of almost three years. The total monthly sales are averaged and hence the average price of the product is calculated to use in this analysis. At this stage the model complexity is only at the base level and none of the market influencing factors have been included in the data therefore the sharp rise of sales in December because of $X_{mas}$ has been replace with the average sales values in the rest of the months of the year. In the comparative study this assumption has least effect on the research targets at this stage.

The problem at hand is to discover the trends in the sales data to forecast the future trends in the sales of the product. We are targeting a real world data set containing the price and the associated sales of a product over a time-period of three years. The model is assumed to be a univariate polynomial function with Gaussian noise of the form [20]:

$$y = f(x) + \varepsilon \tag{9}$$

Considering the behavior of the sales data, we select the univariate polynomials as a nested structure of hypothesis. This structure contains polynomials of varying degree whose coefficients are calculated by minimizing the empirical risk as shown in the previous section. The model description needs to specify the degree of the polynomial, the coefficients of the polynomial and the VC dimension of the polynomial. VC dimension of the univariate polynomial of degree $d$ can show to be equal to $(d+1)$ [20].

Initially in this work, we apply two techniques, i.e. SRM and SVR on the synthetic dataset generated to prove the applicability of SRM principle in its application to the learning from continuous financial systems. Results obtained show that both techniques provide approximately same results. Then, we consider a real world domain where we have prepared the sales dataset in the form that we have independent and dependant quantities. Based on the prepared data, we need to estimate the unknown dependency that generated the behavior of the data. Basically, we are aiming developing a simplified model for sales forecasting and yet comparable to the SVR which is computationally much more expensive. Also presented is an extensive empirical evaluation of the SRM method using sales forecasting dataset and its comparison with the SVR.

In this research, we have covered different aspects of the sales forecasting and performance comparison of different models selection techniques for sales forecasting. We have used standard implementation of support vector machines i.e. LSSVM, available in the form of MATLAB toolbox. For SRM, we have used polynomials of increasing degree as a set of hypothesis for building an optimum model that reflects the inherent details of the dataset.

Our emphasis is on the fact that SRM could be a better technique for forecasting in comparison to the SVR in application where real time system is in operation. In the literature studies so far, SVR is preferred over SRM because of better precision. In our opinion, SRM could be a better way of handling the forecasting issues in real time systems. It has been shown in this research that a simple model based on SRM can give us comparable results and is easy to incorporate with a little trade off of accuracy with the huge advantage of computation of fine tuning the model based on SVR.

## 2. EXPERIMENTAL DETAILS

We have applied the above mentioned techniques on the synthetic dataset which is prepared by adding the Gaussian noise with zero mean and one standard deviation in the randomly selected datasets. The results we are getting are very much in the favor of SRM when computational time is a critical factor. Accuracy of the results obtained using SRM and SVR is comparable as shown later but SRM is much less computationally expensive when compared to SVR. Then, we apply the two techniques on the sales dataset to analyze the performance of the two techniques. The whole process of model selection is shown in Fig. 1.

The simple implementation of SRM guarantees a good performance of the SRM even on the larger dataset. As mentioned earlier, we have used polynomials as a set of hypothesis. The simple nature of polynomials requires less computation yet accurate results. Using SRM regression for function estimation, we obtained the following results shown in Fig. 1. Result is shown in Fig. 2 that provides insight into the model selected and its comparison to the actual synthetic dataset.

Fig. 2 shows that the model obtained using SRM is not smooth and may have some error but within the bounds as specified in the implementation of the SRM.

Model selection using SVR is carried our using the standard implementation of the SVR in the form of solution to the quadratic programming problem. Algorithmically,

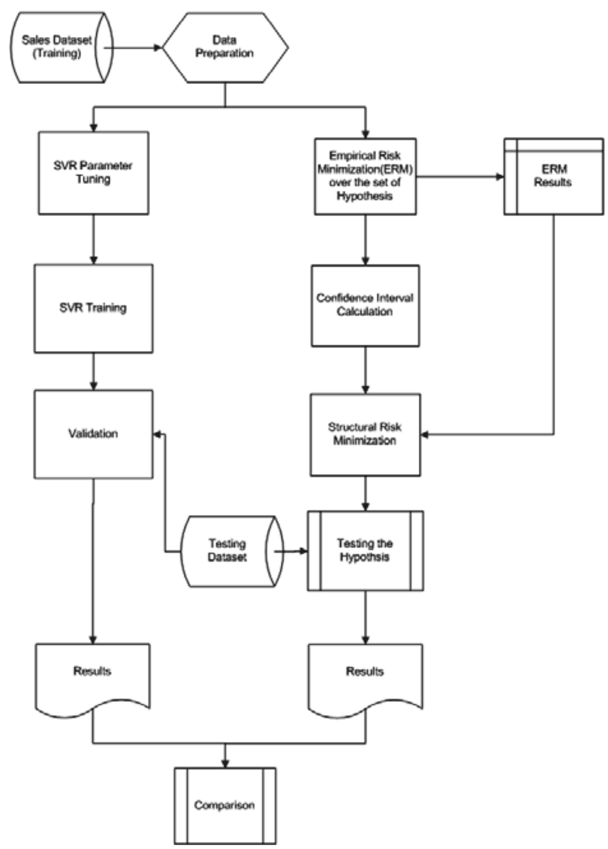


*FIG. 1. COMPARISON OF DIFFERENT TECHNIQUES IN SALES FORECASTING*

SVR is complex due to the generality of the formulation hence computationally expensive. But SVR in itself is a complete solution and can be applied to any problem domain with little or no customization.

In our problem setting, SVR is used to estimate the model using the synthetic and real datasets.

It can be seen that SVR result in Fig. 3 gives a model that is much smoother than that of SRM model selection and is a better solution to the problem of model selection.

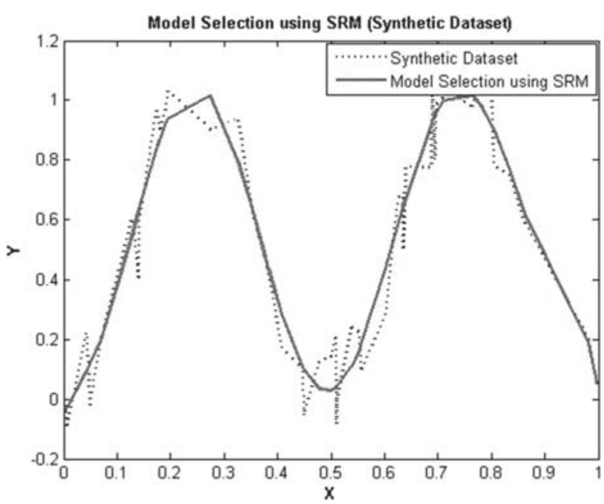


*FIG. 2. MODEL SELECTION USING SRM (SYNTHETIC DATASET)*

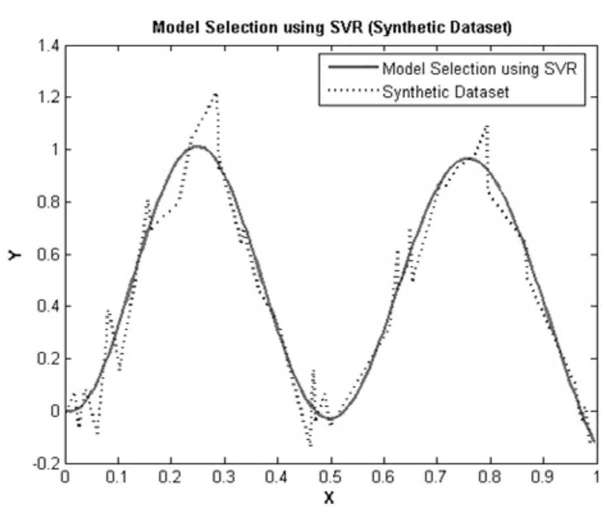


*FIG. 3. MODEL SELECTION USING SVR (SYNTHETIC DATASET)*

We have employed the following measure to calculate the forecasting performance for both the techniques, i.e. for SRM and SVR:

$$\left[ Performance = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_{target} - f\left(\mathbf{x}_i, \alpha\right)\right)^2}{\sigma\left(y_{target}\right)} \right] \quad (10)$$

The reason behind using the aforementioned technique as the measure of the accuracy is that it takes the randomness in the data as the parameter in the determination of the prediction performance [20]. The results obtained using this measure are shown in Table 1.

For our real world sales dataset, we have applied the following steps. Initially, the sales training data is used to estimate the model by minimizing the empirical risk for the functions from each element. For each element of every structure the guaranteed risk is found using the VC generalization bound. Using SRM, an optimal structure element providing minimal guaranteed risk is chosen. Secondly, we have used the Support vector regression for model selection using the RBF kernel after tuning the parameters of the kernel.

Fig. 4 gives the insight about the nature of the data that is analyzed in the upcoming text. Dataset is partitioned into two halves for extensive testing of the trained system.

**TABLE 1. PERFORMANCE OF SVR AND SVM ON SYNTHETIC AND SALES DATASET**

| | Structural Risk Minimization | Support Vector Regression |
|---|---|---|
| Synthetic Dataset | 0.245414 | 0.269527 |
| Sales Dataset | 0.182917 | 0.154506 |

SRM principle suggests the process of model selection consists of two steps. First is selecting an element of a structure with optimal complexity, and then estimating the model from this element, where the model parameters are found using empirical risk minimization [17]. Calculating the VC dimension of the polynomials of the form:

$$p(n) = \sum_{i=0}^{n} a_i x^i \qquad (11)$$

can be done using the formula $d=n+1$ where $n$ is the degree the hypothesis, in our case polynomials. SRM gives an optimal structure element from the set the nested set of structure that provides minimal guaranteed risk. In our case, SRM gives the following fit for the sales dataset.

The comparison of the original dataset and the model obtained using the SRM is shown in Figs. 5-6 to visualize the accuracy in the model selection of SRM. However, to compute the performance of the SRM, we have used Equation (10) and results are shown in Table 1. Details about this experimental setup can be found in [32].

For sales dataset, SVR gives the following results in comparison with the original dataset as shown in Fig. 7.

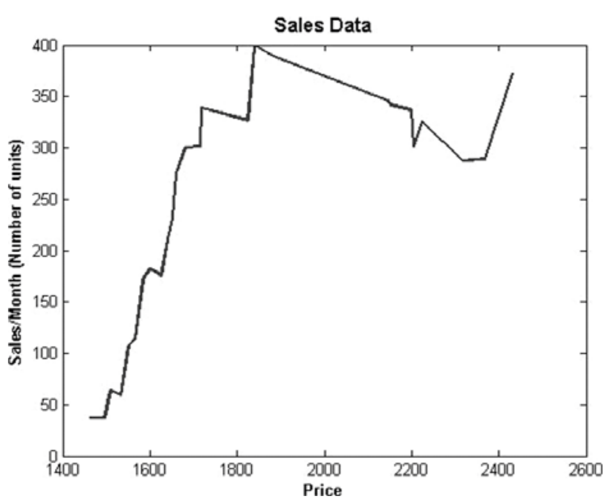


FIG. 4. SALES DATASET

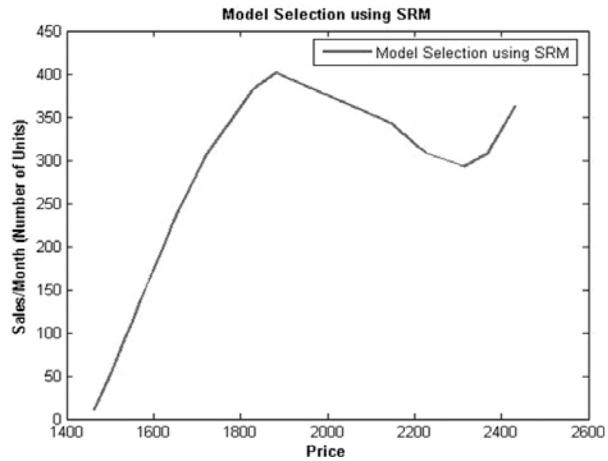


FIG. 5. MODEL SELECTION USING SRM (SALES DATASET)

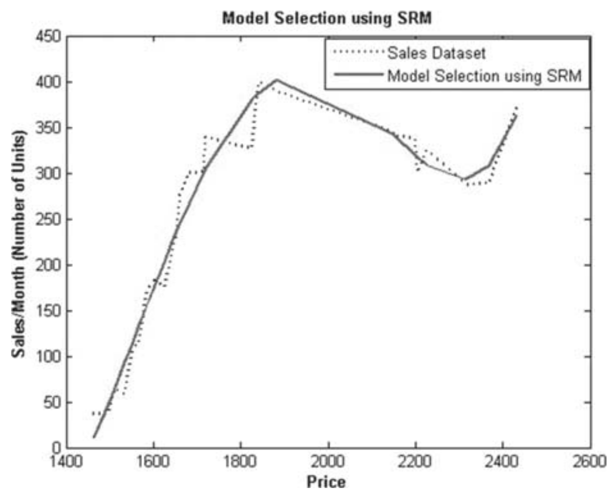


FIG. 6. COMPARISON OF MODEL SELECTION USING SRM (SALES DATASET)

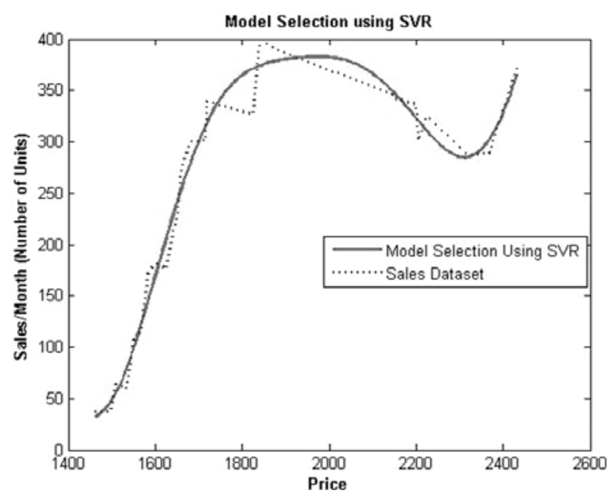


FIG. 7. MODEL SELECTION USING SVR (SALES DATASET)

## 3. COMPARISON

We have focused on two different aspects for sales forecasting namely; performance and computational time. To measure the performance, we have sued the Equation (10) as the measure. Results are shown in Table 1. It shows that SRM performance is comparable to the performance of the SVR with small error margin. An important factor is the utilization of the time resource. As we have seen that error rate of both of the techniques is comparable so the time to obtain the results is important. We can see from Table 2 that time consumption for SRM is much less as compared with SVR. This is very helpful where enterprises are interested in evaluating their models in short interval of times so that they can compare their performance and change their goals accordingly. Function estimation using SRM and SVR results into comparable results but at the different computational cost. Table 2 shows the computational time for the two different techniques using two different dataset.

Form Table 2 it can be seen that SRM is much more computationally efficient than SVR and can be used in forecasting where processing time is a critical factor. Also, performance of SVR is approximately as good as SRM as shown in Table 1.

## 4 CONCLUSIONS

We have examined the feasibility of SRM and SVM for sales forecasting to develop a generic methodology. We then compared the performance of SRM and SVR for sales forecasting in terms of prediction errors to improve the forecasting outcome. Then an extensive evaluation is done to prove that SRM is computationally efficient than SVM with a little trade off in accuracy. It has been shown in this research that a simple model based on SRM can give us comparable results and is easy to incorporate with a little trade off of accuracy with the huge advantage of computation of fine tuning the model based on SVR. We have not incorporated the different parameters of the sales domain in our estimation using structural risk minimization which can be included in the future to narrow down the error margins. Also, we can improve the performance of the SRM by using the different set of hypothesis. This research shows that SRM principle is much more efficient for linear regression models and gives comparable results with SVR.

TABLE 2. COMPUTATIONAL TIME USED BY SRM AND SVR IN SECONDS

| | Structural Risk Minimization | Support Vector Regression |
|---|---|---|
| Synthetic Dataset | 1.8252 | 246.9184 |
| Sales Dataset | 1.1676 | 162.7246 |


## ACKNOWLEDGEMENT

This research work has been completed with the support of the University of Engineering & Technology, Lahore, Pakistan.


## REFERENCES


[1] Lasschuit, W., and Thijssen, N., "Supporting Supply Chain Planning and Scheduling Decisions in the Oil and Chemical Industry", Computers & Chemical Engineering, Volume 28, No. 6, pp. 863-870, 2004.

[2] Maravelias, C.T., and Grossmann, I.E., "Simultaneous Planning for New Product Development and Batch Manufacturing Facilities", Industrial & Engineering Chemistry Research, Volume 40, No. 26, pp. 6147-6164, 2001.

[3] Levis, A.A., and Papageorgiou, L.G., "A Hierarchical Solution Approach for Multi-Site Capacity Planning under Uncertainty in the Pharmaceutical Industry", Computers & Chemical Engineering, Volume 28, No. 5, pp. 707-725, 2004.

[4] Box, G.E., and Jenkins, G.M., and Reinsel, G.C., "Time Series Analysis: Forecasting and Control", Wiley, Volume 734, 2011.

[5] Remus, M.S., "The Handbook of Forecasting: A Manager's Guide", John Wiley and Sons, New York, USA, 1982.

[6] Foster, W.R., Collopy, F., and Ungar, L.H., "Neural Network Forecasting of Short, Noisy Time Series", Computers & Chemical Engineering, Volume 16, No. 4, pp. 293-297, 1992.

[7] Makridakis, S., Hibon, M., and Moser, C., "Accuracy of Forecasting: An Empirical Investigation", Journal of the Royal Statistical Society, Series-A (General), pp. 97-145, 1979.

[8] Gardner, E.S., "Evaluating Forecast Performance in an Inventory Control System", Management Science, Volume 36, No. 4, pp. 490-499, 1990.

[9] Schultz, R.L., "The Implementation of Forecasting Models", Journal of Forecasting, Volume 3, No. 1, pp. 43-55, 1984.

[10] Dalrymple, D.J., "Sales Forecasting Methods and Accuracy", Business Horizons, Volume 18, No. 6, pp. 69-73, 1975.

[11] White, H.R., "Sales Forecasting: Time Saving and Profit-Making Strategies that Work", Scott, Foresman, 1984.

[12] Wacker, J.G., and Lummus, R.R., "Sales Forecasting for Strategic Resource Planning", International Journal of Operations & Production Management, Volume 22, No. 9, pp. 1014-1031, 2002.

[13] Ritzman, L.P., and King, B.E., "The Relative Significance of Forecast Errors in Multistage Manufacturing", Journal of Operations Management, Volume 11, No. 1, pp. 51-65, 1993.

[14] Cherkassky, V., Shao, X., Mulier, F.M., and Vapnik, V.N., "Model Complexity Control for Regression using VC Generalization Bounds", IEEE Transactions on Neural Networks, Volume 10, No. 5, pp. 1075-1089, 1999.

[15] Cherkassky, V., and Mulier, F.M., "Learning from Data: Concepts, Theory, and Methods", Wiley, IEEE Press, 2007.

[16] Vapnik, V., "The Nature of Statistical Learning Theory", Springer, 1999.

[17] Vapnik, V., "Statistical Learning Theory", 1998.

[18] Burges, C.J., "A Tutorial on Support Vector Machines for Pattern Recognition", Data Mining and Knowledge Discovery, Volume 2, No. 2, pp. 121-167, 1998.

[19] Cherkassky, V., and Shao, X., "Model Selection for Wavelet-Based Signal Estimation", IEEE World Congress on Computational Intelligence, Volume 2, pp. 843-848, 1998.

[20] Viswanathan, M., Yang, Y.K., and Whangbo, T.K., "Minimum Message Length and Classical Methods for Model Selection in Univariate Polynomial Regression", ETRI Journal, Volume 27, No. 6, pp. 747-758, 2005.

[21] Smola, A.J., "Regression Estimation with Support Vector Learning Machines", Master's Thesis, Technische Universit, Munchen, 1996.

[22] Gunn, S.R., "Support Vector Machines for Classification and Regression", ISIS Technical Report 14, 1998.

[23] Smola, A.J., and Schökopf, B., "A Tutorial on Support Vector Regression", Statistics and Computing, Volume 14, No. 3, pp. 199-222, 2004.

[24] Akaike, H., "Fitting Autoregressive Models for Prediction", Annals of the Institute of Statistical Mathematics, Volume 21, No. 1, pp. 243-247, 1969.

[25] Akaike, H., "Statistical Predictor Identification", Annals of the Institute of Statistical Mathematics, Volume 22, No. 1, pp. 203-217, 1970.

[26] Schwarz, G., "Estimating the Dimension of a Model", The Annals of Statistics, Volume 6, No. 2, pp. 461-464, 1978.

[27] Craven, P., and Wahba, G., "Smoothing Noisy Data with Spline Functions", Numerische Mathematik, Volume 31, No. 4, pp. 377-403, 1978.

[28] Wallace, C.S., and Boulton, D.M., "An Information Measure for Classification", The Computer Journal, Volume 11, No. 2, pp. 185-194, 1968.

[29] Wallace, C.S., and Freeman, P.R., "Estimation and Inference by Compact Coding", Journal of the Royal Statistical Society, Series-B (Methodological), pp. 240-265, 1987.

[30] Wallace, C.S., "Statistical and Inductive Inference by Minimum Message Length" Springer Science + Business Media, 2005.

[31] Vapnik, V., "Estimation of Dependencies Based on Empirical Data", Springer-Verlag. 1982.

[32] Asif, H.M.S., Mohyuddin, M.A., and Shahbaz, M., "Sales Forecasting Using Structural Risk Minimization Principle of Statistical Learning Theory", International Conference on Data Mining, CSREA, pp. 418-423, USA, 2007.