International Journal of Computer Science and Business Informatics

IJCSBI.ORG

A Predictive Stock Data Analysis with SVM-PCA Model

Divya Joseph
PG Scholar, Department of Computer Science and Engineering
Christ University Faculty of Engineering
Christ University, Kanmanike, Mysore Road, Bangalore - 560060

Vinai George Biju
Asst. Professor, Department of Computer Science and Engineering
Christ University Faculty of Engineering
Christ University, Kanmanike, Mysore Road, Bangalore – 560060

ABSTRACT
In this paper the properties of Support Vector Machines (SVM) on the financial time series data has been analyzed. The high dimensional stock data consists of many features or attributes. Most of the attributes of features are uninformative for classification. Detecting trends of stock market data is a difficult task as they have complex, nonlinear, dynamic and chaotic behaviour. To improve the forecasting of stock data performance different models can be combined to increase the capture of different data patterns. The performance of the model can be improved by using only the informative attributes for prediction. The uninformative attributes are removed to increase the efficiency of the model. The uninformative attributes from the stock data are eliminated using the dimensionality reduction technique: Principal Component Analysis (PCA). The classification accuracy of the stock data is compared when all the attributes of stock data are being considered that is, SVM without PCA and the SVM-PCA model which consists of informative attributes.

Keywords
Machine Learning, stock analysis, prediction, support vector machines, principal component analysis.

1. INTRODUCTION
Time series analysis and prediction is an important task in all fields of science for applications like forecasting the weather, forecasting the electricity demand, research in medical sciences, financial forecasting, process monitoring and process control, etc [1][2][3]. Machine learning techniques are widely used for solving pattern prediction problems. The financial time series stock prediction is considered to be a very challenging task for analysts, investigator and economists [4]. A vast number of studies in the past have used artificial neural networks (ANN) and genetic algorithms for the time series data [5]. Many real time applications are using the ANN tool for time-series modelling and forecasting [6]. Furthermore the

researchers hybridized the artificial intelligence techniques. Kohara et al. [7] incorporated prior knowledge to improve the performance of stock market prediction. Tsaih et al. [8] integrated the rule-based technique and ANN to predict the direction of the S& P 500 stock index futures on a daily basis.

Some of these studies, however, showed that ANN had some limitations in learning the patterns because stock market data has tremendous noise and complex dimensionality [9]. ANN often exhibits inconsistent and unpredictable performance on noisy data [10]. However, back-propagation (BP) neural network, the most popular neural network model, suffers from difficulty in selecting a large number of controlling parameters which include relevant input variables, hidden layer size, learning rate, and momentum term [11].

This paper proceeds as follows. In the next section, the concepts of support vector machines. Section 3 describes the principal component analysis. Section 4 describes the implementation and model used for the prediction of stock price index. Section 5 provides the results of the models. Section 6 presents the conclusion.

2. SUPPORT VECTOR MACHINES

Support vector machines (SVMs) are very popular linear discrimination methods that build on a simple yet powerful idea [12]. Samples are mapped from the original input space into a high-dimensional feature space, in which a 'best' separating hyperplane can be found. A separating hyperplane H is best if its margin is largest [13].

The margin is defined as the largest distance between two hyperplanes parallel to H on both sides that do not contain sample points between them (we will see later a refinement to this definition) [12]. It follows from the risk minimization principle (an assessment of the expected loss or error, i.e., the misclassification of samples) that the generalization error of the classifier is better if the margin is larger.

The separating hyperplane that are the closest points for different classes at maximum distance from it is preferred, as the two groups of samples are separated from each other by a largest margin, and thus least sensitive to minor errors in the hyperplane's direction [14].

2.1 Linearly Separable Data

Consider that there exist two classes and uses two labels -1 and +1 for two classes. The sample is $\chi = \{x^t, r^t\}$ where $r^t = +1$ if $x^t \in C_1$ and $r^t = -1$ if $x^t \in C_2$. To find w and w_0 such that

where, χ represents set of n points

x^t represents p dimensional real vector

r^t represents the class (i.e. +1 or -1)

$$w^T x^t + w_0 \geq +1 \text{ for } r^t = +1$$

$$w^T x^t + w_0 \leq -1 \text{ for } r^t = -1$$

Which can be rewritten as:

$$r^t (w^T x^t + w_0) \geq +1 \qquad (1)$$

Here the instances are required to be on the right of the hyperplane and what them to be a distance away for better generalization. The distance from the hyperplane to the instances closest to it on either side is called the margin, which we want to maximize for best generalization.

The optimal separating hyperplane is the one that maximizes the margin. The following equation represents the offset of hyperplane from the origin along the normal w.

$$\frac{|w^T x^t + w_0|}{||w||}$$

which, when $r^t \in \{+1,-1\}$, can be written as

$$\frac{r^t (w^T x^t + w_0)}{||w||}$$

Consider this to be some value ρ:

$$\frac{r^t (w^T x^t + w_0)}{||w||} \geq \rho, \ \forall t \qquad (2)$$

In order to maximize ρ but there are an infinite number of solutions that are obtained by scaling w, therefore consider ρ ||w|| = 1. Thus to maximize the margin ||w|| is minimized.

$$\min \frac{1}{2} \| w \|^2 \quad \text{subject to } r^t (w^T x^t + w_0) \geq +1, \ \forall t \quad (3)$$

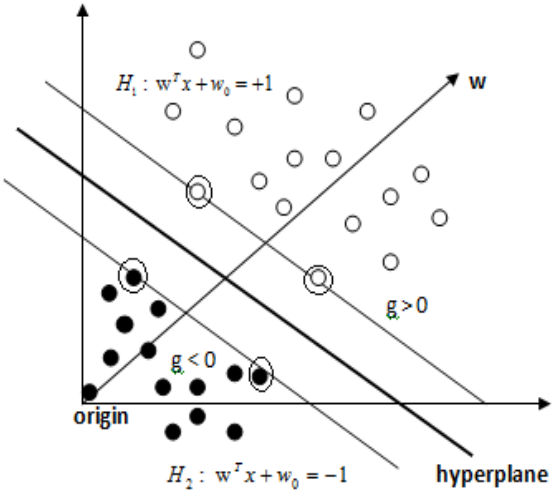

Figure 1 The geometry of the margin consists of the canonical hyperplanes H_1 and H_2.

The margin is the distance between the separating (g(x) =0) and a hyperplane through the closest points (marked by a ring around the data points). The round rings are termed as support vectors.

This is a standard optimization problem, whose complexity depends on d, and it can be solved directly to find w and w_0. Then, on both sides of the hyperplane, there will be instances that are $\frac{1}{\| w \|}$. As there will be two margins along the sides of the hyperplane we sum it up to $\frac{2}{\| w \|}$.

If the problem is not linearly separable instead of fitting a nonlinear function, one trick is to map the problem to a new space by using nonlinear basis function. Generally the new spaces has many more dimensions than the original space, and in such a case, the most interesting part is the method whose complexity does not depend on the input dimensionality. To obtain a new formulation, the Eq. (3) is written as an unconstrained problem using Lagrange multipliers α^t:

$$L_p = \frac{1}{2}\| w \|^2 - \sum_{t=1}^{N} \alpha^t [r^t (w^T x^t + w_0) - 1]$$

$$= \frac{1}{2}\| w \|^2 - \sum_{t=1} \alpha^t r^t (w^T x^t + w_0) + \sum_{t=1} \alpha^t$$

This can be minimized with respect to w, w_0 and maximized with respect to $\alpha^t \geq 0$. The saddle point gives the solution.

This is a convex quadratic optimization problem because the main term is convex and the linear constraints are also convex. Therefore, the dual problem is solved equivalently by making use of the Karush-Kuhn-Tucker conditions. The dual is to maximize L_p with respect to w and w_0 are 0 and also that $\alpha^t \geq 0$.

$$\frac{\partial L_p}{\partial w} = 0 \Rightarrow \text{w} = \sum_{i=1}^{n} \alpha^t r^t x^t \qquad (5)$$

$$\frac{\partial L_p}{\partial w_0} = 0 \Rightarrow \text{w} = \sum_{i=1}^{n} \alpha^t r^t = 0 \qquad (6)$$

Substituting Eq. (5) and Eq. (6) in Eq. (4), the following is obtained:

$$L_d = \frac{1}{2}(w^T w) - w^T \sum_t \alpha^t r^t x^t - w_0 \sum_t \alpha^t r^t + \sum_t \alpha^t$$

$$= -\frac{1}{2}\sum_t \sum_s \alpha^t \alpha^s r^t x^s (x^t)^T x^s + \sum_t \alpha^t \qquad (7)$$

which can be minimized with respect to α^t only, subject to the constraints

$$\sum_t \alpha^t r^t = 0, \text{ and } \alpha^t \geq 0, \forall t$$

This can be solved using the quadratic optimization methods. The size of the dual depends on N, sample size, and not on d, the input dimensionality.

Once α^t is solved only a small percentage have $\alpha^t > 0$ as most of them vanish with $\alpha^t = 0$.

The set of x^t whose $x^t > 0$ are the support vectors, then w is written as weighted sum of these training instances that are selected as support vectors. These are the x^t that satisfy and lie on the margin. This can be used to calculate w_0 from any support vector as

$$w_0 = r^t - w^T x^t \qquad (8)$$

For numerical stability it is advised that this be done for all support vectors and average be taken. The discriminant thus found is called support vector machine (SVM) [1].

3. PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) is a powerful tool for dimensionality reduction. The advantage of PCA is that if the data patterns are understood then the data is compressed by reducing the number of dimensions. The information loss is considerably less.

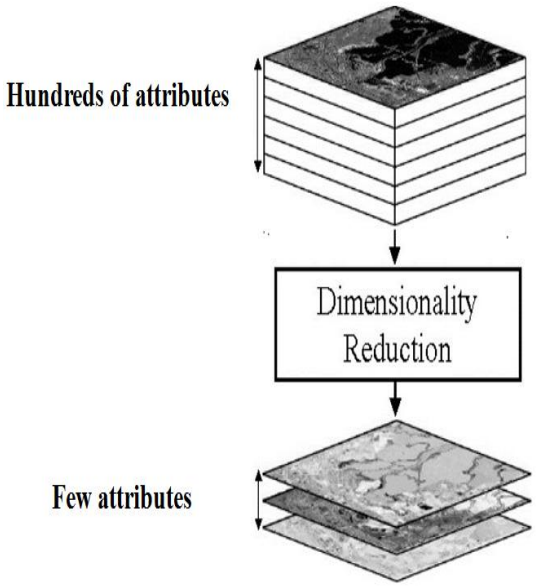

Figure 2 Diagrammatic Representation of Principal Component Analysis (PCA)

4. CASE STUDY

An investor in stocks ideally should get maximum returns on the investment made and for that should know which stocks will do well in future. So this is the basic incentive for forecasting stock prices. For this, he has to study about different stocks, their price history, performance and reputation of the stock company, etc. So this is a broad area of study. There exists considerable evidence showing that stock returns are to some extent predictable. Most of the research is conducted using data from well established stock markets such as the US, Western Europe, and Japan. It is, thus, of interest to study the extent of stock market predictability using data from less well established stock markets such as that of India.

Analysts monitor changes of these numbers to decide their trading. As long as past stock prices and trading volumes are not fully discounted by the market, technical analysis has its value on forecasting. To maximize profits from the stock market, more and more "best" forecasting techniques are used by different traders. The research data set that has been used in this study is from State Bank of India. The series spans from 10th January 2012 to 18th September 2013. The first training and testing dataset consists of 30 attributes. The second training and testing dataset consists of 5 attributes selected from the dimensionality reduction technique using Weka tool: PCA.

Table 1 Number of instances in the case study

State Bank of India Stock Index	
Total Number of Instances	400
Training Instances	300
Testing Instances	100

The purpose of this study is to predict the directions of daily change of the SBI Index. Direction is a categorical variable to indicate the movement direction of SBI Index at any time t. They are categorized as "0" or "1" in the research data. "0" means that the next day's index is lower than today's index, and "1" means that the next day's index is higher than today's index.

The stock data classification is implementation with Weka 3.7.9. The k-fold cross validation is considered for the classification. In the k-fold cross-validation, the original sample is randomly partitioned into k subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k – 1 subsamples are used as

training data [15]. The cross validation variable k is set to 10 for the stock dataset [16].The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds then can be averaged (or otherwise combined) to produce a single estimation.

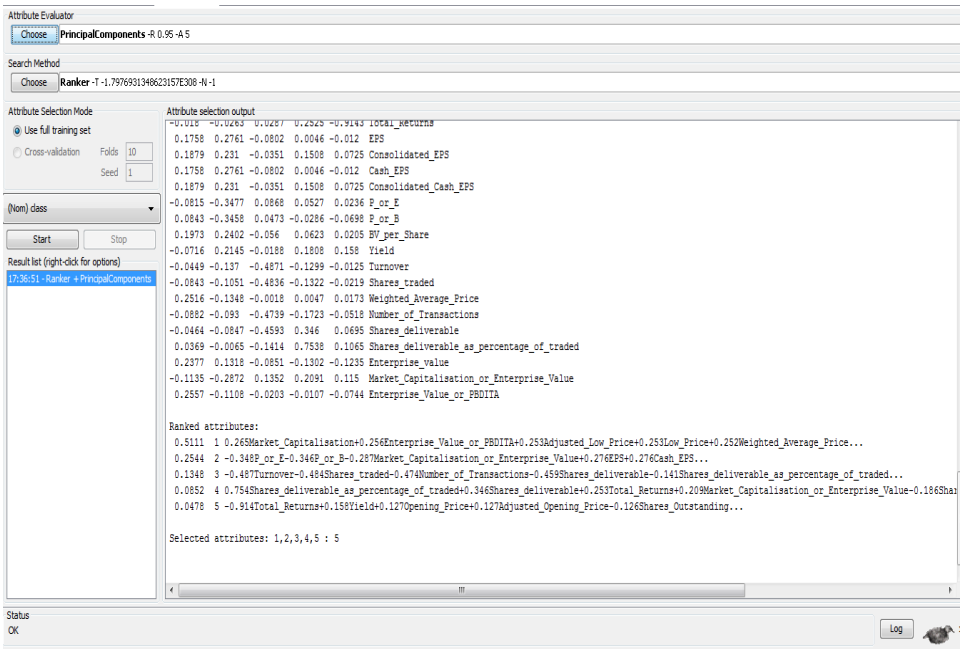

Figure 3 Weka Screenshot of PCA

At first the model is trained with SVM and the results with the test data is saved. Second, the dimensionality reduction technique such as PCA is applied to the training dataset. The PCA selects the attributes which give more information for the stock index classification. The number of attributes for classification is now reduced from 30 attributes to 5 attributes.

The most informative attributes are only being considered for classification. A new model is trained on SVM with the reduced attributes. The test data with reduces attributes is provided to the model and the result is saved. The results of both the models are compared and analysed.

5. EXPERIMENTAL RESULTS

5.1 Classification without using PCA

From the tables displayed below 300 stock index instances were considered as training data and 100 stock index instances were considered as test data. With respect to the test data 43% instances were correctly classified and 57% instances were incorrectly classified.

Table 2 Number of instances for classification without using PCA

Number of Instances and Attributes		
Number of Train Instances	Number of Test Instances	Number of Attributes
300	100	30

Table 3 Classification accuracy without using PCA

Classification Accuracy	
Correctly Classified Instances	43%
Incorrectly Classified Instances	57%

5.2 Classification with PCA

From the tables displayed below 300 stock index instances were considered as training data and 100 stock index instances were considered as test data. With respect to the test data 59% instances were correctly classified and 41% instances were incorrectly classified.

Table 4 Number of instances for classification without using PCA

Number of Instances and Attributes		
Number of Train Instances	Number of Test Instances	Number of Attributes
300	100	5

Table 5 Classification accuracy without using PCA

Classification Accuracy	
Correctly Classified Instances	59%
Incorrectly Classified Instances	41%

6. CONCLUSION

The Support Vector Machines can produce accurate and robust classification results on a sound theoretical basis, even when input stock data are non-monotone and non-linearly separable. The Support Vector Machines evaluates more relevant information in a convenient way. The principal component analysis is an efficient dimensionality reduction method which gives a better SVM classification on the stock data. The SVM-PCA model analyzes the stock data with fewer and most relevant

features. In this way a better idea about the stock data is obtained and in turn gives an efficient knowledge extraction on the stock indices. The stock data classified better with SVM-PCA model when compared to the classification with SVM alone. The SVM-PCA model also reduces the computational cost drastically. The instances are labelled with nominal values for the current case study. The future enhancement to this paper would be to use numerical values for labelling instead of nominal values.

7. ACKNOWLEDGMENTS

We express our sincere gratitude to the Computer Science and Engineering Department of Christ University Faculty of Engineering especially Prof. K Balachandran for his constant motivation and support.

REFERENCES

[1] Divya Joseph, Vinai George Biju, "A Review of Classifying High Dimensional Data to Small Subspaces", Proceedings of International Conference on Business Intelligence at IIM Bangalore, 2013.

[2] Claudio V. Ribeiro, Ronaldo R. Goldschmidt, Ricardo Choren, A Reuse-based Environment to Build Ensembles for Time Series Forecasting, Journal of Software, Vol. 7, No. 11, Pages 2450-2459, 2012.

[3] Dr. A. Chitra, S. Uma, "An Ensemble Model of Multiple Classifiers for Time Series Prediction", International Journal of Computer Theory and Engineering, Vol. 2, No. 3, pages 454-458, 2010.

[4] Sundaresh Ramnath, Steve Rock, Philip Shane, "The financial analyst forecasting literature: A taxonomy with suggestions for further research", International Journal of Forecasting 24 (2008) 34–75.

[5] Konstantinos Theofilatos, Spiros Likothanassis, Andreas Karathanasopoulos, Modeling and Trading the EUR/USD Exchange Rate Using Machine Learning Techniques, ETASR - Engineering, Technology & Applied Science Research Vol. 2, No. 5, pages 269-272, 2012.

[6] A simulation study of artificial neural networks for nonlinear time-series forecasting. G. Peter Zhang, B. Eddy Patuwo, and Michael Y. Hu. *Computers & OR 28(4):381-396* (*2001*)

[7] K. Kohara, T. Ishikawa, Y. Fukuhara, Y. Nakamura, Stock price prediction using prior knowledge and neural networks, Int. J. Intell. Syst. Accounting Finance Manage. 6 (1) (1997) 11–22.

[8] R. Tsaih, Y. Hsu, C.C. Lai, Forecasting S& P 500 stock index futures with a hybrid AI system, Decision Support Syst. 23 (2) (1998) 161–174.

[9] Mahesh Khadka, K. M. George, Nohpill Park, "Performance Analysis of Hybrid Forecasting Model In Stock Market Forecasting", International Journal of Managing Information Technology (IJMIT), Vol. 4, No. 3, August 2012.

[10] Kyoung-jae Kim, "Artificial neural networks with evolutionary instance selection for financial forecasting. *Expert System. Application* 30, 3 (April 2006), 519-526.

[11] Guoqiang Zhang, B. Eddy Patuwo, Michael Y. Hu, "Forecasting with artificial neural networks: The state of the art", International Journal of Forecasting 14 (1998) 35–62.

[12] K. Kim, I. Han, Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index, Expert Syst. Appl. 19 (2) (2000) 125–132.

[13] F. Cai and V. Cherkassky "Generalized SMO algorithm for SVM-based multitask learning", IEEE Trans. Neural Netw. Learn. Syst., Vol. 23, No. 6, pp.997 -1003, 2012.

[14] Corinna Cortes and Vladimir Vapnik, Support-Vector Networks. Mach. Learn. 20, Volume 3, 273-297, 1995.

[15] Shivanee Pandey, Rohit Miri, S. R. Tandan, "Diagnosis And Classification Of Hypothyroid Disease Using Data Mining Techniques", International Journal of Engineering Research & Technology, Volume 2 - Issue 6, June 2013.

[16] Hui Shen, William J. Welch and Jacqueline M. Hughes-Oliver, "Efficient, Adaptive Cross-Validation for Tuning and Comparing Models, with Application to Drug Discovery", The Annals of Applied Statistics 2011, Vol. 5, No. 4, 2668–2687, February 2012, Institute of Mathematical Statistics.

This paper may be cited as:

Joseph, D. and Biju, V. G., 2014. A Predictive Stock Data Analysis with SVM-PCA Model. *International Journal of Computer Science and Business Informatics, Vol. 9, No. 1, pp. 1-11*.