Neurocomputing ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A real time clustering and SVM based price-volatility prediction for optimal trading strategy

Subhabrata Choudhury [a,1], Subhajyoti Ghosh [b,2], Arnab Bhattacharya [c], Kiran Jude Fernandes [d,3,4], Manoj Kumar Tiwari [e,*,5]

[a] Department of Metallurgical & Materials Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721302, India
[b] Department of Ocean Engineering and Naval Architecture, Indian Institute of Technology Kharagpur, Kharagpur 721302, India
[c] University of Pittsburgh, Pittsburgh, PA 15213, United States
[d] Department of Management, Durham University Business School, Mill Hill Lane, Durham University, Durham DH1 3LB, United Kingdom
[e] Department of Industrial Engineering and Management Indian Institute of Technology, Kharagpur 721302, India

ARTICLE INFO

Article history:
Received 17 October 2012
Received in revised form
22 July 2013
Accepted 10 October 2013
Communicated by Bijaya Ketan Panigrahi

Keywords:
Stock market
Clustering
Self-Organizing Maps
Trading strategy
Support vector machine

ABSTRACT

Financial return on investments and movement of market indicators are fraught with uncertainties and a highly volatile environment that exists in the global market. Equity markets are heavily affected by market unpredictability and maintaining a healthy diversified portfolio with minimum risk is undoubtedly crucial for any investment made in such assets. Effective price and volatility prediction can highly influence the course of the investment strategy with regard to such a portfolio of equity instruments. In this paper a novel SOM based hybrid clustering technique is integrated with support vector regression for portfolio selection and accurate price and volatility predictions which becomes the basis for the particular trading strategy adopted for the portfolio. The research considers the top 102 stocks of the NSE stock market (India) to identify set of best portfolios that an investor can maintain for risk reduction and high profitability. Short term stock trading strategy and performance indicators are developed to assess the validity of the predictions with regard to actual scenarios.

1. Introduction

The global financial markets are again fraught with uncertainties at every level of investment and over every possible investment vehicle. Recent developments like down grading of US credit rating by Standards and Poor's (S and P) from the embellished AAA to a prudent AA+ and ongoing Euro credit crunch involving massive government debts have forced several countries into tailspin and the contagion have heavily affected many economies all around the world, taking the investors by surprise and proving even their worst case predictions wrong. The implication of such astounding events could be seen in the massive price surge in the global gold markets whereas a complete opposite scenario evolved in the US equity, stock and commodities market which was complemented by a weakened dollar and an even frailer euro. An overwhelming majority of investors and investment institutions tend to formulate their strategies based on extrapolating simple recent trends and calculate the portfolio return-risk trade-off to formulate an optimal one.

The fallacy lies in the predictions and decisions based only upon price movements of the indices or individual stock in the market and the technical analysis for the various strategies for a range of investment vehicles. The concept of risk or volatility takes a very important meaning in this context.

Determining the standard deviation or variance of a particular asset class or its derivative becomes absolutely crucial in giving a holistic view of the market uncertainties existing. The structural exogenous changes in the market are extremely hard to predict, hence the analysis used in this piece of research focuses on giving an investor a robust tool that can accurately gauge the mood of the market and the asset under consideration which becomes crucial in formulating a trading strategy that matches the risk averseness or risk empathy of individual investors or conglomerates. Over reaction is extremely dangerous at arriving at investment decisions which becomes the cornerstone of formulating analytical or heuristic solutions towards strategy formulation that can hedge against such paranoia upon any disruptive influence. As no model is fool proof, the performance should be gauged by real time

* Corresponding author. Tel.: +91 3222 283 746.
E-mail addresses: subhabrata.iitkgp@gmail.com (S. Choudhury), g.subhajyoti90@gmail.com (S. Ghosh), cfcarnabiitkgp@gmail.com (A. Bhattacharya), mkt09@hotmail.com, mkt009@gmail.com (M.K. Tiwari).
[1] Tel.: +91 974 993 5575.
[2] Tel.: +91 99 337 954 32.
[3] Tel.: +44 191 33 45512.
[4] URL: http://www.durham.ac.uk/k.j.fernandes.
[5] Web: http://sites.google.com/site/mktiwari09.

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.10.002

portfolio return and risk analysis on a daily basis which gives an investor to choose a set of classes that will diversify the risk which aims to maximize or more correctly optimize the returns in lieu with his targeted value [1].

The work presented here focuses on devising an optimal portfolio of risky asset or asset clusters which will present an investor to assess the risk-return involved in selection. The model is applied for all the tradable stocks in the National Stock Exchange (NSE) of India, which provides alternative derivative instruments like index futures and options. The work chooses Indian stock for the analysis because of the high beta of the Indian economy and where distortion effects of financial fallouts can be studied amidst a favorable environment for long term and short term investments fueled by genuine economic growth. The work envisages a complete predictive module that can accurately forecast the prices and the inherent volatilities associated with the asset types, which are first clustered based upon their risk and return profiles. A hybrid SOM (Self-Organizing Maps) using K means clustering is used for clustering the stocks whereas Support Vector Regression (SVR) is used to predict the future price and volatility for short trading cycles for better forecasts. Based on the results, a trading strategy module is articulated which selects the best strategy for trading under the implied uncertain scenario in the market indicated by the forecasts. As stated earlier, more than the accuracy in predicting the actual prices and volatilities, what becomes important is the relative direction of the movements and a definite range of values that the model can propose based upon the different strategy that an investor chooses based upon his risk aversion. Existing work which incorporates clustering techniques [2,3] and advanced statistical and data mining techniques [4] have undergone several transformations for better predictive ability. At the moment of writing, the authors have no knowledge of an assimilated coupled system which uses both clustering and predictive techniques linked to strategy selection for stock market portfolio.

The paper is structured as follows. The next section discusses about the existing literature existing in our field of application. The following section introduces the complete clustering, prediction and strategy selection module methodology and the various algorithms involved at each stage. This methodology is then tested and analyzed for performance on the stocks of the NSE for a given trade cycle. The paper concludes with a definitive discussion on the scope of work and the existing shortcomings that can become a future area of research. The lists of all references appear thereafter.

2. Literature review

In this section, related work in clustering techniques, regression models and trading strategies are discussed which sets the stage for the problem addressed in this paper.

2.1. Clustering techniques

For the past few decades clustering techniques has been used to assort varied data sets but it is only in the later 1990s when clustering techniques were exploited on financial data. Conventional partitive and hierarchal algorithms evolved into their genetic hybrids like GA–K-means. These were extensively used for market segmentation [5]. In case of financial data, clustering algorithms has also been used to cluster time series data. Clustering of the time series facilitates a better regression model for future prediction [6] as it considerably reduces the noise of a non-stationary time series, like that of a stock price.

Though there are numerous partitive and agglomerative clustering algorithms SOM has emerged as one of the more popular choice in clustering multi-dimensional data, as analyzed by Shahapurkar and Sundareshan [7]. SOM use artificial neural networks for cluster data and helps mapping multi-variate data on a 2-D space as shown by [8]. The complexity of these clustering algorithms are proportional to amount of data fed, thereby heavily increasing the computational time for the clustering operation. In our case data is both noise prone and exhaustive.

To eliminate the noise and subsequently improve the computation time Vesanto and Alhoniemi [9] proposed a novel two level abstraction to cluster the Self Organizing Map. Their experiments hinted that clustering the SOM is a more computationally effective approach than directly clustering the data set. On the other hand SOM helps to visualize higher dimensional data sets on a 2-D space, which will be helpful if we increase the dimensions representing a stock to more than 2.There have been efforts to create a portfolio using different single layer clustering methods [10], but to the best of our knowledge, the proposed two layers is yet to be implemented for generating an optimal portfolio. The two stage subroutine – first using SOM to produce the prototypes that are then clustered in the second stage – is found to perform well when compared with direct of the data and to reduce the computation time A research by Canetta et al. [11] also concludes the same.

A few articles like [12–14] compared different clustering techniques to find the optimal number of clusters and to validate the results. The results obtained from the clustering techniques have been validated.

2.2. Regression model

Over the last few years growing number of researchers are studying the price and volatility movement of different kinds of financial instruments. Academicians and corporate researchers are trying their best to formulate methods to predict the future economic market and devise an effective trading system to maximize the profit [15].

Before the introduction of computational intelligence traditional statistical techniques such as multivariate regression, autoregressive integrated moving average (ARIMA) [16], generalized autoregressive conditional heteroskedasticity (GARCH) [17] were being used for forecasting. They are unable to produce significant result as stocks data are generally complex and noisy in nature. To correct the problem artificial intelligence techniques such as Generic Algorithms, Artificial Neural Networks (ANN) were proposed to approach this problem. Researchers are now inclining towards Support Vector Machine (SVM), first suggested by Vapnik [18] to improve the forecast [6,19]. Most comparison results of show that SVM surpasses ANN in terms of prediction performance [20]. This is due to less complex structure of SVR and due to the implementation of structural risk minimization principle, SVR attempts to minimize the upper bound of generalization error whereas in ANN empirical risk minimization principle is implemented which seeks to minimize misclassification error or deviation of the solution from the test data. Also there is lesser chance of over fitting with SVM as it is global optimum whereas ANN may generate only local optimal solutions [21,22]. A convex quadratic optimization is used to get the solution for the SVM where a Karush–Kuhn–Tucker (KKT) statement molds the necessary and sufficient conditions for a global optimum [23]. On the other hand, in case of ANN even a well optimized algorithm may not be able attain global optimum within finite process time [24].

The modeling of five futures of Chicago Mercantile Market was done by [25] using SVM. They also compared the performance with multilayer back-propagation (BP) neural network models to

Please cite this article as: S. Choudhury, et al., A real time clustering and SVM based price-volatility prediction for optimal trading strategy, Neurocomputing (2013), http://dx.doi.org/10.1016/j.neucom.2013.10.002

find that SVM outperforms the latter one. SVM was used by [21] to predict the daily change of price of Korea Composite Stock Price Index (KOSPI). Further the SVM model was compared with case-based reasoning (CBR) and back-propagation neural network (BPN) where SVM outperformed the other two. Prediction of weekly movement trend of NIKKEI 225 Index was carried out using SVM [26]. To evaluate the forecasting ability of SVM, its performance was compared with those of Linear Discriminant Analysis, Quadratic Discriminant Analysis and Elman Back propagation Neural Networks (NN). SVM outperformed other methods in the experiment. A combining model by integrating SVM with the other classification methods was also proposed in the paper. Gavrishchaka and Banerjee [27] addressed the problem of volatility forecasting from high-dimensional stock market. SVM-based volatility model was comparable often superior to the established volatility prediction models for instance GARCH and its generalizations. Forecasting of S&P CNX NIFTY Market Index of the NSE was carried out by Kumar and Thenmozhi [28] using SVM and Random Forest Regression. In the test SVM was superior to Random Forest, ANN in addition to traditional methods like ARIMA. The empirical studies in the Hong Kong securities market substantiate that the ANN and SVR approaches noticeably shrink the average forecast errors and accordingly improve the forecasting accuracy. A nonparametric approach based on ANN and SVM along with improved conventional option pricing technique was presented to forecast option prices from the Hong Kong securities market [29]. Results showed significant improvement in forecasting accuracy with ANN and SVR based models. The performance on literature case studies of SVM regression is measured against other advanced learning methods such as the Radial Basis Function, the traditional Multilayer Perceptron model, Box–Jenkins autoregressive-integrated-moving average and the Infinite Impulse Response Locally Recurrent Neural Networks [22].The comparison shows that in the analyzed cases, SVM comparable to and in most cases outperforms the other techniques. Classical methods like ARMA and GARCH, requires huge sample size for better prediction leading to a reduced original sample size for higher order models [30]. They also showed that SVM and Back propagation (BP) performs better than the ARMA model in the deviation measurement criteria. Experiments also show that SVM outperforms both finite mixture of ARMA-GARCH and BP models in deviation performance criteria [31].

More recent studies have been targeting a hybrid SVR approach to improve the forecast performance of SVR [32,33]. Signal processing techniques like wavelet transform and nonnegative matrix factorization helps in improving the forecasts.

The forecasts obtained from the Support Vector Regression model, provides closing prices and volatility values for the next 15 days. These values formed the platform for formulating the trading strategy to maximize profit.

2.3. Trading strategy

A great deal of work has been published over the past decade on Stock market trading strategies. Most of these strategies are either plagued down by lags, as it is in the case of Candle stick strategy, or is fit for long term investments, like the case of momentum trading. The strategy that we propose is based on the analysis of two trading days, which gives it an advantage over the other strategies keeping in mind the present turmoil in the financial markets. A decision matrix is also prepared to help the trader take decisions.

The short term trading strategy is ideal for unstable markets. The detailed methodology is discussed in the next section.

3. Methodology

Through the literature survey it is found that the problem of formulating optimal trading strategies in a volatile environment can be tackled more efficiently with the help of clustering and prediction methodologies. It involves three discrete steps which are incorporated in the formulation of the final trading strategy for different market conditions. It is discussed below in the following section.

3.1. Clustering

The first step of the solution involves clustering of stocks listed in NSE based on their *Logarithmic Returns* and *Daily Underlying Volatilities*. Clustering is a method of unsupervised learning to partition a data set into a set of clusters. This paper proposes a two layer abstraction to cluster the stocks using SOM followed by K-means clustering of the SOM as shown in Fig. 1. The first level, SOM, comprises of a 2-D neural network with neighborhood relations among the neurons. The Input vectors are connected to the output layers but the neurons are not inter-connected. SOM is suited for clustering and mapping of higher dimensional data on 2-D plots, proving to be ideal in the case of multi-dimensional market data. The visualization obtained from SOM fails to infer quantitative description of the data parameters, thereby promoting the need for another clustering algorithm for proper interpretation of the SOM prototypes. The prototypes obtained represent the local averages of the data belonging within a particular radius of that prototype. These local averages are further used as the data set for the second level of clustering.

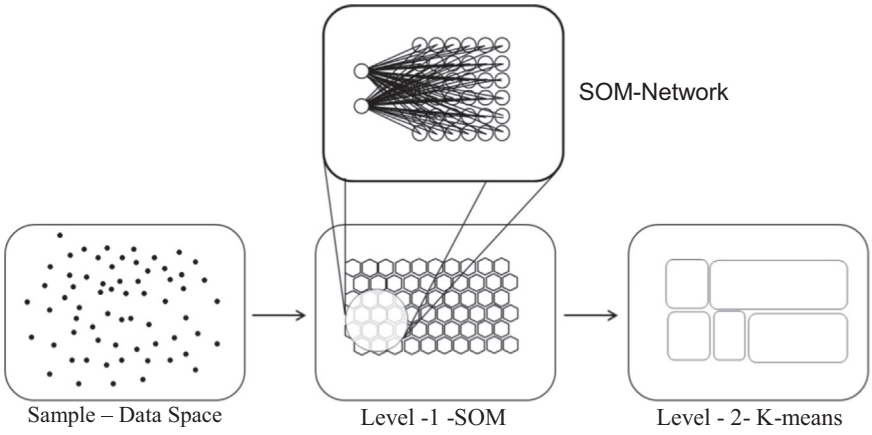


Fig. 1. Two levels of abstraction used to cluster stocks. Level 1: SOM and Level 2: K-means.

Please cite this article as: S. Choudhury, et al., A real time clustering and SVM based price-volatility prediction for optimal trading strategy, Neurocomputing (2013), http://dx.doi.org/10.1016/j.neucom.2013.10.002

The second level in the approach involves clustering using a partitive algorithm. Partitive algorithms like K-means partitions the data set into k-clusters, in which every prototype is clustered based on its closest mean. Results obtain from K-means represent clusters of prototypes, which in turn are a map of the underlying stocks. The optimal number of clusters is determined from various cluster validity indices. Indices calculated also hints an optimal clustering in this case, that is tabulated in Table 1.

This two level approach substantially reduces computation time in large data sets, because it is much lighter to cluster a set of epitomes. Consider clustering N samples using K-means. It involves making several clustering trials with different values for. The computational complexity is proportional to, $\sum_{k=2}^{C_{max}} N_k$ where C_{max} is preselected maximum number of clusters. When a set of prototypes are used as an intermediate step, the total complexity is proportional to $NM+\sum_k M_k$, where M is the number of prototypes. With $C_{max}=\sqrt{N}$ and $M=5\sqrt{N}$, the reduction of computational load is about $\sqrt{N}/15$, or about six-fold for $N=1000$. Of course, this is a very rough estimate since many practical considerations are ignored.

The cluster having the best underlying stocks is then selected for the regression function and eventually the trading strategy is contrived. From Fig. 2, it can be easily inferred that Clusters 2 and 3 will correspond to higher returns. It is to be noted that so called poor performing stocks can also be an investment option. In our case the two stage clustering yielded stocks with better returns and low volatility, thereby making them an investor's safe choice.

3.2. Support vector machine

The second step of the solution involves regressing the time series of the stocks obtained from the first step and thereby prediction future values using Support Vector Machines. SVMs are very specific class of algorithms, characterized by usage of kernels, flatness of solution and capacity control obtained by acting on the number of support vectors. SVMs are learning machines which implements the structural risk minimization inductive principle to obtain good generalization on a limited number of learning data set.

Set of points in the network: (x_1,y_1), (x_2,y_2), …, (x_m,y_m)

$$h(x)=w^T\varphi(x_i)+b \tag{1}$$

We define the above function taking into account:

Table 1
Cluster validity.

Index	Value
Silhouette	0.47904
Davies–Bouldin	0.50596
Calinski–Harabasz	66.9185
Dunn	2.4381
R-squared	0.98224

C penalty cost for deviation during training process
F a feature space
W a vector in F
w^Tw regularized term controlling the function capacity
ε an insensitive loss function as proposed by Vapnik [18]. It is the maximum error allowed during training of the dataset
$\xi_i^+\ \xi_i^-$ slack variables corresponding to the size of the excess deviations
$\sum(\xi_i^+ +\xi_i^-)$ empirical error calculated from the insensitive loss function
$\Phi(x_i)$ a mapping function which maps each x_i to a vector in F

In ε-SVR the objective is to obtain a function $h(x)$ that accounts for the most ε deviation from the actually obtained target values in the training data set. The absolute errors are neglected as long as they are less than ε. Initially we consider the function to be a linear function for the sake of simplicity

Based on SVM literature, the given optimization problem is formulated:

$$\text{Minimize } \frac{1}{2}w^Tw+C\sum(\xi_i^+ +\xi_i^-) \tag{2}$$

$$\text{Subject to } y_i-w^T\Phi(x_i)-b\le\varepsilon+\xi_i^+ \tag{3}$$

$$\text{Subject to } w^T\Phi(x_i)+b-y_i\le\varepsilon+\xi_i^- \tag{4}$$

$$\text{Subject to } \xi_i^+,\quad \xi_i^-\ge 0 \tag{5}$$

constant $C>0$ determines the offset between the flatness of f and the amount up to which deviations larger than ξ are tolerated. Thus the ξ can be deduced to another form considering the error weightage. $|\xi|E$ – insensitive loss function (Fig. 3).

The problem is solved using the KKT Conditions. SVR maps the training data set onto a feature space to enable non liner prediction. The parameters of ε-SVR were optimized using both grid and pattern search algorithms. The parameters with the minimum Mean Absolute Percentage Error (MAPE) were chosen. MAPE is a measure of the accuracy of a method for constructing fitted time series values. The accuracy is expressed in terms of percentage. Finally the forecasts were compared with Neural Network model for the same.

$$\text{MAPE } M=\frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t-F_t}{A_t}\right| \tag{6}$$

A_t – Actual
F_t – Forecast

The closing prices and Volatility predicted forms the basis of the different trading strategies, discussed in the next step.

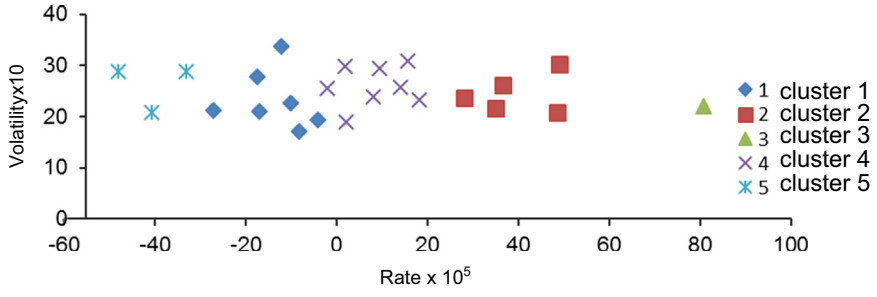


Fig. 2. Second layer of abstraction – K-means clustering.

Please cite this article as: S. Choudhury, et al., A real time clustering and SVM based price-volatility prediction for optimal trading strategy, Neurocomputing (2013), http://dx.doi.org/10.1016/j.neucom.2013.10.002

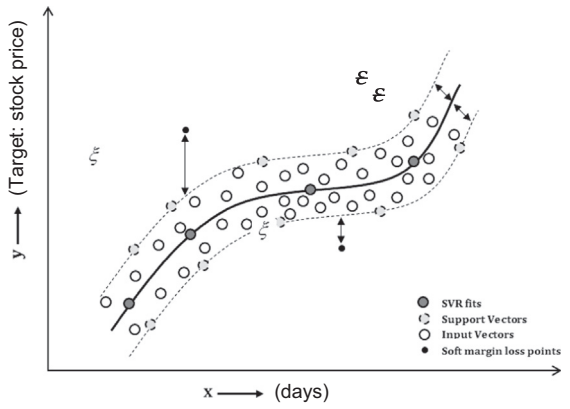


Fig. 3. Soft margin losses of linear SVM.

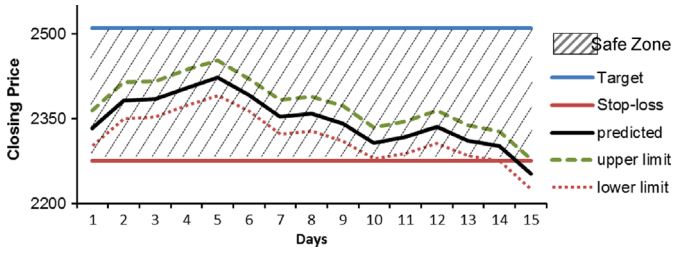


Fig. 4. General view of trading strategy.

3.3. Trading strategy

Within the realistic domain of an equity stock market and the portfolio generated from the above computational analysis, we consider a generic trading strategy for the trading space considered in the problem. We analyze the trading strategy from the view of a risk-averse trader or investor who tries to maximize his net returns focusing on minimizing the associated risk in the investment. The computational module deployed earlier gives a crystallized view of the market movements in the future, both in terms of price as well as the volatility associated. It is important to remember that throughout the entire scope of the study, only temporal effects of the time series data considered is studied. We refrain from going into long term investment strategy especially for volatile stocks like those considered simply because its erroneous on part of the researchers to not take into account dominant structural alterations that make equity markets unpredictable as they are. Instead the focus is on short term strategies (ranging from a week to a maximum of month) which take into account short term ripples and market tendencies to derive an investor strategy for the risky assets. Contrarian and momentum investment strategies are suited for longer time scales and we do not consider those strategies. Our strategy is based on a micro level strategy for the best portfolio (ranked) for each stock and its related performance based on the movement predictions. Markets are generally difficult to study owing to strong coupling between stocks, nonlinear investor response to news and imperfect dissemination of news among the investors rendering it highly inefficient, a marked deviation from traditional assumptions of market efficiency.

The investor considered has characteristic features like initial investment, target profit, stop loss price and time threshold being few of the most important ones. The investor's target margins will be reflected in the transactions (buy–sell) made based on the limits set. A general view is given below in Fig. 4 which consists of the average, stop loss limit, target limit and investment horizon (15 days for the case considered). As long as the prices remain in between these goalposts, transactions made by an investor will be less based on the preferences and the situation in the market (not considering any external stimuli of structural characteristics). However when prices exceed the limits, it is an indication of selling stocks for various reasons in accordance with the escape position existing in the market. The volume of transaction is of paramount importance because it directly affects the profit–loss scenario of the overall investment horizon. Prescient knowledge of market indicating a bullish or bearish tendency is already available for the future through the predictions made earlier.

Generally there are periods of normal and frantic trading based on the market indications and transaction volume becomes a major factor then. Normal trading involves slow oscillating price movements and low volumes where owners are reluctant to short because prices are not attractive enough and buyers prefer to wait for even lower prices. Significant trading and volume increase (rally) is followed by a crash which is indicative of settlements. Demand is truly a stochastic variable and transaction will depend upon human interpretation of the market scenario.

In the model discussed below, every investor has an initial investment to make at the starting period for the portfolio chosen. All investors intend to get rich with minimum risk Investor, with past price-volatility history, future short term predictions and incomplete news from the market. The transactions made are series of buy–sell events based on the prices existent at the particular time for each stock in the portfolio. For simplicity, volume considered is constant based on investor preference. The tendency to buy or sell is decided based upon a decision matrix (Fig. 5) based on bullish or bearish indicators of prices and the movement of the volatility giving the investor a holistic view of the strategy he should use. The scale of the investments and its profitability can be gauged by the slope between a buy–sell pair. A steep slope implies high positive or negative impact of the transaction whereas a mild slope implies less impact on overall profitability. We define a few indices to gauge the relative performance of the strategy.

Suppose we define m trading pairs of buy–sell. The absolute impact of the transaction can be valued as absolute profit index λ and relative profit index υ given by

$$\lambda = \sum_{m=1}^{n} p_{bm} v_{bm} - \sum_{m=1}^{n} p_{sm} v_{sm} \tag{7}$$

$$\upsilon = \lambda / I \tag{8}$$

where p_b, p_s, v_b, v_m are the price and volume of buying and selling shares for the transaction pair and I is the initial investment. Based on the λ, υ values of each stock we can determine the performance of the portfolio as a whole by deriving indices for the entire portfolio or compare among individual stocks to foresee the performance. We assume there are no dividends involved in the time horizon considered.

Further analysis of the performance can be gauged by the slope of buy–sell price transaction graph. It is assumed that volume involved is same, though in real cases we can derive values for volumes actually traded.

Volatility / Price	HIGH	LOW
HIGH	BUY	HOLD
LOW	SELL	HOLD

Fig. 5. Proposed trading strategy.

Please cite this article as: S. Choudhury, et al., A real time clustering and SVM based price-volatility prediction for optimal trading strategy, Neurocomputing (2013), http://dx.doi.org/10.1016/j.neucom.2013.10.002

4. Computational results and analysis

In this paper historical stock data (closing price and intra-day volatility) for the companies listed NSE was collected from NSE database and Yahoo finance. The closing prices and intra-day volatility data has been collected for 102 stocks for the horizon 2008–2011. We collected a fairly mixed data from various sectors and capital sizes. To model the functions the logarithmic returns and daily volatility were considered.

From the initial pool of 102, 31 poor performing stocks were eliminated to form the final data pool. The aforementioned two layer clustering algorithm was applied on the data pool to obtain clusters based on their Returns and Daily Underlying Volatility. The clusters with relatively higher performing stocks constituted the optimal portfolio and subsequently the underlying stocks were regressed. The underlying stocks are listed in Table 2.

The ε-SVR parameters for one of the stocks (GLAXO), along with its training data is shown in Table 3. Fig. 6 shows the plot between actual and predicted closing prices for the validation data space for GLAXO.

It has been stated earlier, that the optimal parameters for the SVR were found out by Grid and Pattern Search algorithms to minimize the Mean Absolute Percentage Error (MAPE). The behavior of the predicted value varies extensively with the change of parameters. In Table 4 MAPE and Root Mean Square Error (RMSE) for the underlying stocks are compared with the Neural Network model.

The results of the trading analysis are shown below. The stock taken is GLAXO from the best portfolio. The analysis is shown in Fig. 7

For GLAXO, three transactions were done based on the proposed trading strategy. The results are shown in Table 5.

If a total investment of 50,000 INR is assumed, the above transaction results in a return of 1.5% over the investment in a 15 day span. The strategies for other selected stocks with initial capital of 1000 stocks each are shown in Table 6.

The obtained Return on Investments (ROI) for all the five stocks using the same strategy with the predictions obtained from Neural Networks and SVM has been compared in Table 7.

5. Conclusion

In this paper a clustering and prediction based short term trading strategy has been discussed. The model for classifying best

Table 2
Optimal portfolio – highest performing stocks listed in National Stock Exchange (NSE).

Name	Volatility × 10	Rate × 10^5
TRIVENI	31.92331825	83.6908
GLAXO	12.15919784	77.8804
BHUSANSTL	29.73370912	57.0786
DR REDDY	18.64341629	53.702
HERO HONDA	19.59290779	48.54

Table 3
SVM parameters and statistical data of training data for GLAXO.

Mean target value for input data	1289.6758
Mean target value for predicted values	1285.5683
Variance in input data	324357.5
Correlation between actual and predicted	0.998536
Maximum error	293.33115
MAPE (Mean Absolute Percentage Error)	1.4246145
Proportion of variance explained by model (R^2)	0.99699

Type of SVM model: Epsilon-SVR.
SVM kernel function: Radial Basis Function (RBF).
Epsilon = 0.001, **C** = 2000, **Gamma** = 8, ***P*** = **0.**

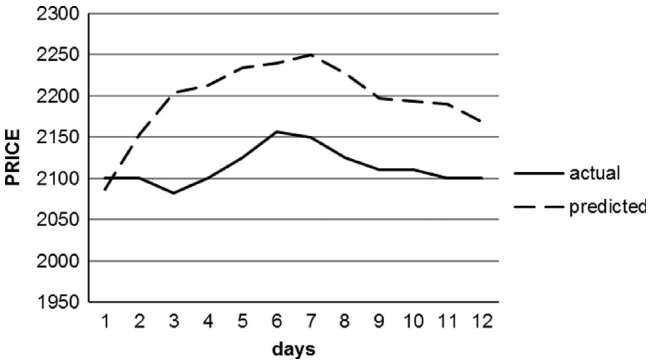


Fig. 6. Actual v/s predicted closing prices (validation sample) – GLAXO.

Table 4
Mean absolute percentage errors of the portfolio.

Name	MAPE (Closing prices) SVM	RMSE (Closing prices) SVM	RMSE (Closing prices) NN
TRIVENI	2.959	2.579	1.161
GLAXO	4.473	2.768	1.248
BHUSANSTL	3.341	2.139	5.234
DR REDDY	2.308	1.807	1.648
HERO HONDA	3.303	2.528	4.237

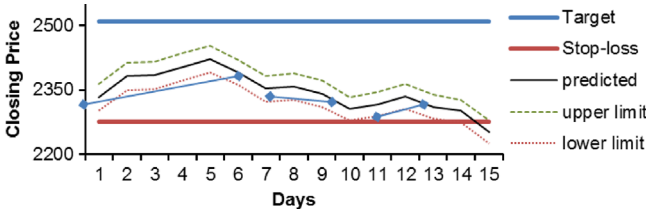


Fig. 7. Trading analysis – GLAXO.

Table 5
Trading strategy results – GLAXO.

Bought (day)	Sold (day)	Buy price	Sell price	Profit	Slope
1	6	2333	2391	58000	0.232
7	9	2353	2341	−12000	−0.12
10	12	2306	2335	29000	0.29

Table 6
Trading strategy results for 1000 stocks each.

Stock	Bought (day)	Sold (day)	Buy price	Sell price	Profit
BHUSANSTL	1	3	449	468	19,000
	6	15	447	516	69,000
HERO HONDA	1	9	1535	1740	205,000
	12	15	1680	1755	75,000
DR REDDY	1	3	1600	1630	30,000
	5	8	1603	1657	54,000
	9	12	1645	1685	40,000
	13	15	1712	1719	7000
TRIVENI	1	5	101.31	102.52	1210
	10	15	96.26	96.13	−130

Please cite this article as: S. Choudhury, et al., A real time clustering and SVM based price-volatility prediction for optimal trading strategy, Neurocomputing (2013), http://dx.doi.org/10.1016/j.neucom.2013.10.002

Table 7
Comparison between ROIs with starting capital of 1000 stocks for a 15 day horizon.

Stock	ROI SVM (%)	ROI NN (%)
BHUSANSTL	19.6	3.4
DR REDDY	8.2	7.0
HERO HONDA	18.2	8.1
TRIVENI	1.1	5.9
GLAXO	3.2	5.3

stocks at NSE and predicting the volatility of chosen stocks were formulated and analyzed. The model was developed in a generic fashion so it may be customized for most of the stock markets in the world. Two level of classification of different stocks was done by applying SOM on the pool of stocks and then applying K-Means clustering technique on the first layer of abstraction. Subsequently, SVR was applied on the clustered output for predicting price and volatility. Finally a user based decision system based on dynamic two-day price prediction was implemented.

Our work can find various applications in software development acting as investing guide to a target trader in a volatile market and Investor's technical information in financial markets. The clustering approach can also be used to tackle problems that have issues with noisy data set as in the case of multidimensional financial data.

In order to make this more suitable for front end solutions there are a few avenues of improvement that need to be addressed. The future work may include evaluating the performance of SVR and optimizing the computational time. SVMs sometimes account for poor scaling with the data size due to quadratic optimization algorithm and kernel transformation; furthermore the apt choice of the kernel parameters leads to extensive computation time through gird searches. Implementing heuristic approach in the intermediate steps to obtain the SVM parameters optimizes the solution. This will be another step towards arriving to a more real time accurate solution. We believe the next step of this problem is to investigate how to increase the accuracy and reliability by including the real world constrains before implementing it to real instances. A real life scenario of an effective trading system can be envisaged when this model can be integrated with a fundamental model which can take into account the market shocks and probability of external events which can then scale the parameters accordingly based on the probability of events. This will give more real time dynamic response to investor decisions rather based on just time series historical data as is used here.

References

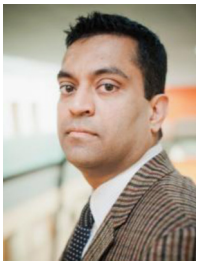
[1] H. Markowitz, Portfolio selection, J. Finance 7 (1) (1952) 77–91.
[2] A. Khan, K. Khan, B.B. Baharudin, Frequent patterns mining of stock data using hybrid clustering association algorithm, in: Proceedings of the International Conference on Information Management and Engineering, ICIME'09, 2009.
[3] C. Guo, H. Jia, N. Zhang, Time series clustering based on ICA for stock data analysis, in: Proceedings of the 4th International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM'08, 2008.
[4] P. Tino, C. Schittenkopf, G. Dorffner, Financial volatility trading using recurrent neural networks, IEEE Trans. Neural Networks 12 (4) (2001) 865–874.
[5] P. Ahmadi, F. Samsami, Pharmaceutical Market Segmentation using GA K-means, Eur. J. Econ. Finance Administrative Sci. 22 (2010).
[6] S. Hsu, J. Hsieh, T. Chih, K. Hsu, A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression, Expert Syst. Appl. 36 (4) (2009) 7947–7951.
[7] S.S. Shahapurkar, M.K. Sundareshan, Comparison of self-organizing map with K-means hierarchical clustering for bioinformatics applications, in: Proceedings of the IEEE International Joint Conference on Neural Networks, vol. 2, 2004, pp. 1221–1226.
[8] B.S. Penn, Using self-organizing maps to visualize high-dimensional data, Comput. Geosci. 31 (5) (2005) 531–544.
[9] J. Vesanto, E. Alhoniemi, Clustering of the self-organizing map, IEEE Trans. Neural Networks 11 (3) (2000) 586–600.
[10] S.R. Nanda, B. Mahanty, M.K. Tiwari, Clustering Indian stock market data for portfolio management, Expert Syst. Appl. 37 (12) (2010) 8793–8798.
[11] L. Canetta, N. Cheikhrouhou, R. Glardon, Applying two-stage SOM-based clustering approaches to industrial data analysis, Prod. Plann. Control 16 (8) (2005) 774–784.
[12] C. Budayan, I. Dikmen, M.T. Birgonul, Comparing the performance of traditional cluster analysis, self-organizing maps and fuzzy C-means method for strategic grouping, Expert Syst. Appl. 36 (9) (2009) 11772–11781.
[13] K.K. Delibasis, N. Mouravliansky, G.K. Matsopoulos, K.S. Nikita, A. Marsh, MR functional cardiac imaging: segmentation, measurement and WWW based visualisation of 4D data, Future Gener. Comput. Syst. 15 (2) (1999) 185–193.
[14] S.A. Mingoti, J.O. Lima, Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithm, Eur. J. Oper. Res. 174 (3) (2006) 1742–1759.
[15] A. Chen, M.T. Leung, H. Daouk, Application of neural networks to an emerging financial market: forecasting and trading the Taiwan Stock Index, Comput. Oper. Res. 30 (6) (2003) 901–923.
[16] G.E.P. Box, G.M. Jenkins, Time Series Analysis, Forecasting, and Control, Holden-Day, San Francisco, 1970.
[17] T. Bollerslev, Generalized autoregressive conditional heteroskedasticity, J. Econ. 31 (3) (1986) 307–327.
[18] V.N. Vapnik, The Nature of Statistical Learning Theory, 2nd Ed., Springer, New York, 1995.
[19] C. Yeh, C. Huang, S.A. Lee, Multiple-kernel support vector regression approach for stock market price forecasting, Expert Syst. Appl. 38 (3) (2011) 2177–2186.
[20] F.E.H. Tay, L. Cao, Application of support vector machines in financial time series forecasting, Omega 29 (4) (2001) 309–317.
[21] K. Kim, Financial time series forecasting using support vector machine, Neurocomputing 55 (1–2) (2003) 307–319.
[22] M.C. Moura, E. Zio, I.D. Lins, E. Droguett, Failure and reliability prediction by support vector machines regression of time series data, Reliab. Eng. Syst. Saf. 96 (11) (2011) 1527–1534.
[23] B. Schölkopf, A.J. Smola, Learning with Kernels – Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, Cambridge, 2002.
[24] P. Secchi, E. Zio, F.D. Maio, Quantifying uncertainties in the estimation of safety parameters by using bootstrapped artificial neural networks, Ann. Nucl. Energy 35 (12) (2008) 2338–2350.
[25] L.J. Cao, F.E.H. Tay, Support vector machine with adaptive parameters in financial time series forecasting, IEEE Trans. Neural Networks 14 (6) (2003) 1506–1518.
[26] W. Huang, Y. Nakamori, S. Wang, Forecasting stock market movement direction with support vector machine, Comput. Oper. Res. 32 (10) (2005) 2513–2522.
[27] V. Gavrishchaka, S. Banerjee, Support vector machine as an efficient framework for stock market volatility forecasting, Comput. Manage. Sci. 3 (2) (2006) 147–160.
[28] M. Kumar, M. Thenmozhi, Forecasting stock index movement: a comparison of support vector machines and random forest, IIMB Manage. Rev. 21 (1) (2008) 41–55.
[29] X. Liang, H. Zhang, J. Xiao, Y. Chen, Improving option price forecasts with neural networks and support vector regressions, Neurocomputing 72 (13-15) (2009) 3055–3065.
[30] A.F.Z. Hossain, M. Nasser, M. Islam, Comparison of GARCH, Neural Network and Support Vector Machine in Financial Time Series Prediction, Pattern Recognition and Machine Intelligence, Lecture Notes in Computer Science, Springer, Berlin/Heidelberg (2009) 597–602, http://dx.doi.org/10.1007/978-3-642-11164-8_97.
[31] A. Hossain, M. Nasser, Comparison of GARCH and neural network methods in financial time series prediction, in: Proceedings of the 11th International Conference on Computer and Information Technology, ICCIT 2008, X, 2008.
[32] Ling-Jing Kao, Chih-Chou Chiu, Chi-Jie Lu, Chih-Hsiang Chang, A hybrid approach by integrating wavelet-based feature extraction with MARS and SVR for stock index forecasting, Decision Support Syst. 54 (3) (2013) 1228–1244.
[33] Ju-Jie Wang, Jian-Zhou Wang, Zhe-George Zhang, Shu-Po Guo, Stock index forecasting based on a hybrid model, Omega 40 (6) (2012) 758–766.

Subhabrata Choudhury is currently pursuing his Bachelors of Technology degree in Metallurgical and Materials Engineering at Indian Institute of Technology Kharagpur, India and is in final year. His current research interest includes Data Mining, Operations Research, Machine Learning and their applications in Finance and Steel Industry.

Please cite this article as: S. Choudhury, et al., A real time clustering and SVM based price-volatility prediction for optimal trading strategy, Neurocomputing (2013), http://dx.doi.org/10.1016/j.neucom.2013.10.002

Subhajyoti Ghosh is a fourth year undergraduate student in Indian Institute of Technology Kharagpur, India enrolled in the five year dual degree program (B. Tech and M. Tech) in Ocean Engineering and Naval Architecture. His current research focuses on Operations Research, Financial Markets, and Scheduling.

Arnab Bhattacharya is currently a PhD candidate in Operations Research at University of Pittsburgh, US. He completed the five year dual degree program (B. Tech and M. Tech) in Industrial Engineering and Management at Indian Institute of Technology Kharagpur, India in 2011. His research areas include Operations Research and Data Mining.



Kiran Jude Fernandes is the Research Director and Head of the Operations Management Group at the York Management School, UK. He is also one of the Principal Investigators at the interdisciplinary York Centre for Complex Systems Analysis (YCCSA). He holds a PhD in Operations Management and Systems from the University of Warwick; a Masters (MS) from the James Worth Bagley College of Engineering at Mississippi State University (MSU) and a Bachelors of Engineering (Hons) degree in Production from Waltech. His research focuses on modeling of complex social and business domains using a complex system perspective.

Manoj Kumar Tiwari is a professor in the Department of Industrial Engineering and Management in Indian Institute of Technology Kharagpur, India. He is an associate editor of journals which include *IEEE Transactions on SMC*, *Part A: Systems and Humans, International Journal of System Science*, *Journal of Decision Support System*. He has more than 200 publications in various international journals and conferences. His research interests are Decision Support Models, Planning, Scheduling and Control Problems of Manufacturing System, Supply Chain Network.

Please cite this article as: S. Choudhury, et al., A real time clustering and SVM based price-volatility prediction for optimal trading strategy, Neurocomputing (2013), http://dx.doi.org/10.1016/j.neucom.2013.10.002