2018 3rd International Conference for Convergence in Technology (I2CT)
The Gateway Hotel, XION Complex, Wakad Road, Pune, India. Apr 06-08, 2018

Mutual Fund NAV Prediction Using Cascaded SVM Models

Akhila Vangara*
B.E.
Department of Electronics
and Communication
PES University, South Campus
Bangalore, Karnataka 560100
akhila.vangara.2895@gmail.com
*corresponding author

Syed Thouseef*
B.E.
Department of Electronics
and Communication
PES University, South Campus
Bangalore, Karnataka 560100
thouseef6170@gmail.com
*corresponding author

Shwetha S Bhat
Assistant Professor
Department of Electronics
and Communication
PES University, South Campus
Bangalore, Karnataka 560100

V V Rao
Former Principal Scientist
IFCPAR
New Delhi

***Abstract*—The world of finance is often viewed as a gamble because of its unpredictability. Although NAVs are not as volatile as stock prices, the mutual fund market too has its share of uncertainties. The current models for the prediction of NAVs are based on historical NAV data using neural network models or regression models. The paper aims to predict NAV by expanding the focus onto various other parameters as input, including macroeconomic factors. In addition, results have been optimized using SVM in a cascaded form.**

The experiments were conducted for the Indian market, in specific, the HDFC mid cap opportunities (Growth) mutual fund scheme.

keywords- NAV, SVM, stock prediction, mutual funds, Indicators, Overlay, Lag

I. Introduction

Mutual funds are pooled investment vehicles that are created by the capital accumulated to invest in a range of securities such as stocks, bonds or other assets. The capital is generated by investments from a large number of individuals and/or organizations. The price of a Mutual Fund unit is called its Net Asset Value and it varies on a daily basis. The returns on a mutual fund scheme are maintained by the fund managers so as to ensure results; the portfolio of the scheme is balanced. Therefore, it is less treacherous than the stock market. Hence, the comparatively slow moving mutual fund market is an investment attraction across the Indian society.

The period of investment in mutual funds may vary between 1 to 10 years. Therefore, the knowledge of the behavior of Net Asset Value of a Mutual Fund Scheme is useful for both the investors as well as the fund managers. For the common man this information can aid in determining when to withdraw from the scheme so as to reap maximum benefits. The fund managers on the other hand can conclude when to book profits on a particular stock.

Prior work in this area has used models such as ARIMA and FLANN. These models revolve around the historical NAV data only. However, we propose a system which also considers the variation in stock market among other parameters affecting the NAV.

The paper focuses on prediction of the Net Asset Value of the HDFC mid cap opportunities fund (Growth) for the period between two to four months. Weka (3.8) and correspondingly a machine learning algorithm has been used for delivering the highest accuracy for this application is the Sequential Minimal Optimization (SMO) technique of Support Vector Machine(SVM) regression.

II. Materials And Methods

A. Technical Indicators

Technical indicators of a stock are derived from the book value of a stock, actual price of the stock and variations of the price thereof. The variations are related to many macro and micro economic factors. The macro-economic factors are government policies, international prices of oil and gold and general sentiment of the markets across the world. The micro-economic factors encompass the confidence of investors in a particular stock based on its past performance, present working and future deliverables commonly called as market sentiment. Some of the technical indicators used are - **Simple Moving Average**, **Exponential Moving Average**, **Moving Average Convergence Divergence**, **Bollinger Bands**.

B. Fundamental Indicators

Fundamental analysis comes from the point of view of evaluating a stocks intrinsic value. Intrinsic value is the actual worth of the stock as opposed to the rate it is traded in the market. It involves an economic and financial assessment of a company and their assets based on their records, past values and estimated futures. A keen glance at revenues, earnings, returns and profit margins can determine the value of a company and thereby their stocks.

Some of the fundamental indicators used are- **Price to Earnings Ratio(P/E)** and **Price to Book Value (P/B)**.

C. Macroeconomic Factors

The Bombay Stock Exchange(BSE) and any stock market in general is affected by a multitude of factors that do not depend on or reflect in the fundamental and technical indicators. These are called Macroeconomic factors and are used to account for the ”human sentiment”.

978-1-5386-4273-3/18/$31.00 ©2018 IEEE

D. Weka 3.8

The tool weka consists of a package called time series forecasting which was used for prediction of stock prices and therefore the NAV. In this package SVM algorithm and SMO technique was used to train and test the data. Using two thirds of the data for training and one third for testing, evaluation of both these phases was done through Mean Absolute Error (MAE), Mean Absolute Percentage Error(MAPE), Root Mean Square Error(RMSE) and Directional Accuracy(DA).

E. Evaluation Metrics

Proposed SVM model has 3 phases of regression.
The first phase involves training the model based on a segment of the data provided. The second phase involves testing the trained model based on another segment of the data. The last phase is forecasting of the target variable based on the historic data only.
In the case of stocks, training and testing was done with two-thirds of data for training and one-third for testing. However, for NAVs, training, testing as well as forecasting was performed. This experiment was carried out with varying distributions of training and testing. The evaluation of the data is done based on four parameters:

1) **Mean Absolute Error**
2) **Directional Accuracy**
3) **Mean Absolute Percentage Error**
4) **Root Mean Square Error**

F. Importance of Overlay

The Indian stock market(BSE), especially during the period chosen for the experiment was volatile. It can be seen further that the predictions based on solely historical values of indicators and stock prices are fairly poor. This indicates that there are certain other macro-economic factors such as human sentiment and government policies which may not be reflected in the historical values per se.
Hence the Weka tool provides the facility to use certain indicators as the guidance parameters. These parameters are meant to assist the predictions and not themselves be a part of the predictions. Any of the parameters except the target variable, which in this case the closing price can be used as the overlay.

G. Lag Variables

Lag is defined as the ”window” of periodicity that the software takes into consideration for learning the behavior of the data. For instance for monthly data the lag can be 12, for weekly 7, for daily 24 or 12.
For the purpose of the experiment we customized the lag for both stock prediction and NAV prediction.

III. SVM and Mutual Funds

The backbone of this research is SVM implemented through SMO. It can be described as an algorithm that approximates the target variable y_i as a linear function of the training data.

A. SVM Basics

Assuming that the input space is $\chi \in \mathbb{R}$ and the training data is $\{(x_1, y_1)$,(x_2, y_2)....$(x_l, y_l)\}$ we define each of the training data x_i as a vector which can be of any dimension. The function $f(x)$ can thus be defined as [2]:

$$f(x) = \langle w, x_i \rangle + b \tag{1}$$

with $w \in \chi$ and $b \in \mathbb{R}$.
w represents the weight vector corresponding to vector x_i whereas b represents the bias. The essence of the algorithm is to predict with an error as less as possible; an error margin of $\pm\epsilon$. Due to this, the convex optimization problem is [2]:

$$\frac{1}{2}\|w\|^2 \tag{2}$$

subject to:
$y_i - \langle w, x_i \rangle - b \leq \epsilon$
$\langle w, x_i \rangle + b - y_i \leq \epsilon$

However, by relaxing the constraints, introduce slack variables and tunable parameter. Minimize[2]:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi + \xi^*) \tag{3}$$

subject to:
$y_i - \langle w, x_i \rangle - b \leq \epsilon + \xi_i$
$\langle w, x_i \rangle + b - y_i \leq \epsilon + \xi_i^*$
$\xi, \xi^* \geq 0$
C is the tunable parameter such that $C > 0$. It provides the trade off between the flatness of $f(x)$ and the amount to which deviation greater than ϵ can be tolerated. Referred to as the primal, this objective problem can be solved by introducing a set of variables and evaluating their saddle condition.
By solving the **Lagrangian** and using the **KKT** conditions [2] it can be easily shown that:

$$w = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)x_i. \tag{4}$$

Thus, the weight vector is described as a linear combination of training data.

B. SMO in Mutual funds

In SVM regression, the input pattern consisting of all parameter values are mapped onto a feature space ϕ. The input vector, x, is related to z by : $z_i = \phi(x_i)$. Inner products are thus calculated as $\phi(x) \cdot \phi(\hat{x})$ [1]. Each of these products are tabulated, weighted and thereafter added.
The dot product in the feature space can in turn be replaced by utilizing a user entered kernel, which is a similarity function such that the above inner product is equivalent to $k(x, \hat{x})$. The idea is to extrapolate a function $f(x)$ such that it is as close to the target variable as possible.
For the purpose of our experiment, on the first level of stock prediction, we had several parameters that were used to train the SVM(SMO) model such as SMA, EMA, VWAP etc. Therefore, it is safe to assume that each of these parameters

represented a dimension of our input training data and hence every support vector, x_i. The algorithm then calculates a weight vector w such that each parameter was weighted depending on its influence on the target variable, the closing price of the stock, y_i. This brings us to the previously established function: $f(x) = \langle w, z_i \rangle + b$, Where z_i is the transform of x_i in the feature space and b represents the bias. Using the feature space z equation (3) now translates to:
Minimize

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{5}$$

Subject to:
$y_i - w \cdot z_i - b \leq \epsilon + \xi_i$
$w \cdot z_i + b - y_i \leq \epsilon + \xi_i^*$
$\xi_i, \xi_i^* \geq 0$
Thus, it can be concluded that for optimality, $b_{low} \leq b_{up}$. [2]
The Lagrangian multiplier β and the bias threshold parameter b are closely related as can be seen from the above equations. It is important to note that at optimality, they are equal.
For more on SMO refer [2].
For improvements in SMO refer [1]

IV. Experiments

Simulations were performed on Weka 3.8 for the stock prediction. Thereafter the predicted stock prices were used to predict the values of NAV of a particular Mutual Fund Scheme.

A. *Stock prediction*

The top 30 stocks of HDFC Mid Cap Opportunities fund - Growth scheme were selected out of the investments in a total of over 80 stocks. Each stock of the 30 was used for the experiment as shown in the form of the examples *Voltas* and *MRF*. Each stock had a varying influence on NAV in terms of its percentage weight-age to the same. Consequently, the number of units invested in each stock was corresponding to its weight-age to the NAV.
Some of the top players include *Tube Investment of India, Voltas, Aurobindo Pharma, IndusInd Bank* and *Yes Bank*. The *Open, High, Low, Close* and *Volume* values were extracted. The fundamental indicator values were obtained and technical indicator values were calculated for the same.

1) Voltas: The model was trained with 2 years data and tested with 1 year data. The stock for voltas was predicted. Figures 1-4 highlight the results in the form of graphs with and without Overlay and Lag.
Figures 5-8 represent the evaluation metrics for the same.

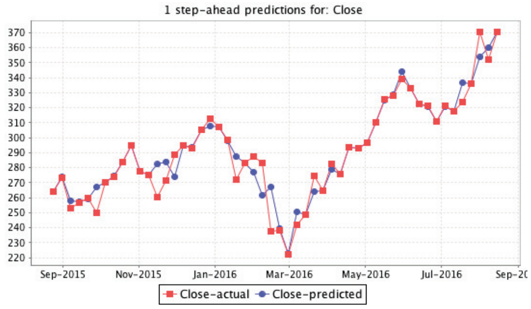

Fig. 1. Voltas Training without any lag or overlay

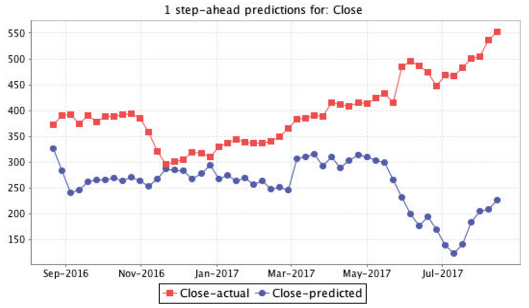

Fig. 2. Voltas Testing prediction without any lag or overlay

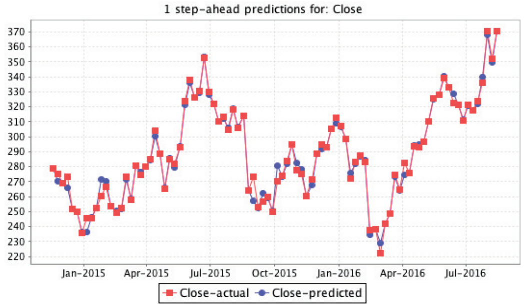

Fig. 3. Voltas Training with lag and overlay

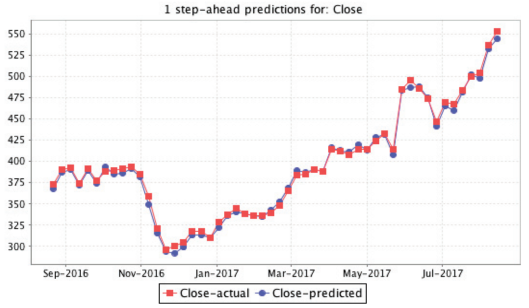

Fig. 4. Voltas Testing prediction with lag and overlay

Target : Close	1-step-ahead	2-steps-ahead	3-steps-ahead	4-steps-ahead	5-steps-ahead
N	52	51	50	49	48
MAE	4.3363	4.4307	4.6016	4.6886	4.9013
DA	76.4706	76	77.551	79.1667	78.7234
MAPE	**1.5611**	**1.5957**	**1.6618**	**1.6883**	**1.7654**
RMSE	8.3002	8.3699	8.1553	8.2254	8.3055

Fig. 5. Voltas Training data evaluation without Lag or Overlay

Target : Close	1-step-ahead	2-steps-ahead	3-steps-ahead	4-steps-ahead	5-steps-ahead
N	52	51	50	49	48
MAE	140.6973	140.3302	161.0526	177.09	175.4129
DA	50.9804	52	57.1429	58.3333	57.4468
MAPE	**32.8771**	**32.7966**	**37.6311**	**41.3004**	**40.7527**
RMSE	171.7783	170.0169	194.873	215.3129	214.4876

Fig. 6. Voltas Testing data evaluation without Lag or Overlay

Target : Close	1-step-ahead	2-steps-ahead	3-steps-ahead	4-steps-ahead	5-steps-ahead
N	92	91	90	89	88
MAE	1.8221	1.8425	1.8002	1.8529	1.8312
DA	91.2088	91.1111	91.0112	90.9091	90.8046
MAPE	**0.655**	**0.662**	**0.6462**	**0.6654**	**0.6576**
RMSE	3.3241	3.3425	3.3174	3.3611	3.2936

Fig. 7. Voltas Training data evaluation with Lag and Overlay

Target : Close	1-step-ahead	2-steps-ahead	3-steps-ahead	4-steps-ahead	5-steps-ahead
N	52	51	50	49	48
MAE	3.6128	3.4906	3.5937	3.5844	3.5178
DA	88.2353	88	87.7551	87.5	89.3617
MAPE	**0.9263**	**0.893**	**0.9217**	**0.9175**	**0.8966**
RMSE	4.382	4.2683	4.4056	4.3721	4.2988

Fig. 8. Voltas Testing data evaluation with Lag and Overlay

2) MRF: This is an example of high volatility in terms of closing prices of stocks. The MRF stock started in August 2015 at a rate of around Rs 35,000 and went on to increase to over 65,000 by the end of three years. The prices increased and decreased rather intermittently which called for the use of overlay of certain parameters for accurate predictions.
Figures 9-12 represent graphically the MRF stock predictions. Figures 13-16 show the evaluation metrics of the same.

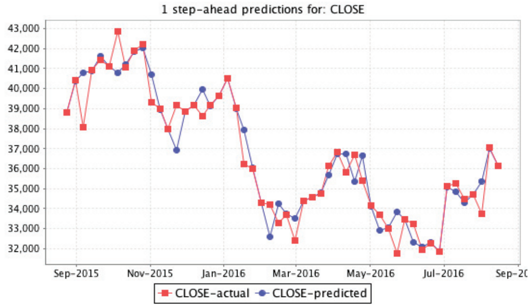

Fig. 9. MRF Training without Lag or Overlay

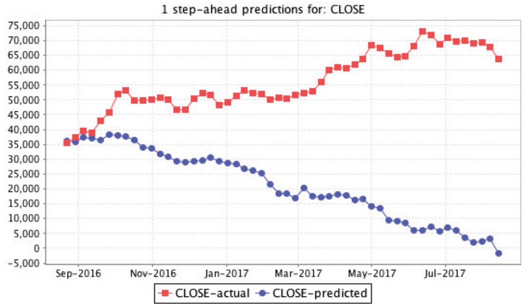

Fig. 10. MRF Testing or prediction without Lag or Overlay

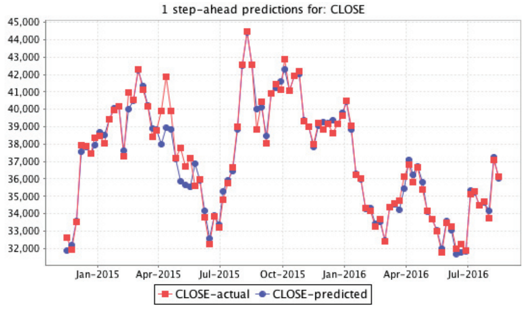

Fig. 11. MRF Training with Lag and Overlay

B. Stock Prediction - Analysis

1) Voltas: Figures 1 and 3 represent the training data of actual values and predicted values. However it should be noted that while the prediction of the training data without lag or overlay seems to be somewhat close to the actual value, it is obvious that the presence of lagged variables and overlay provide a much better training of the SVM model. From figure 5 the training evaluation is at MAPE above 1.7% as compared

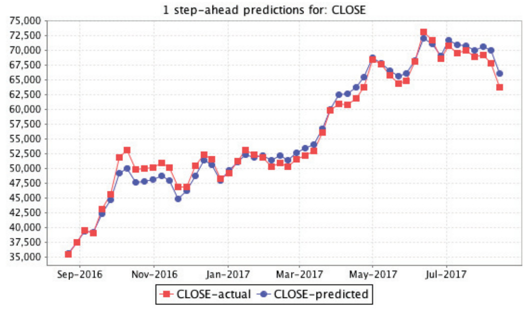

Fig. 12. MRF Testing with Lag and Overlay

Target : Close	1-step-ahead	2-steps-ahead	3-steps-ahead	4-steps-ahead	5-steps-ahead
N	52	51	50	49	48
MAE	510.9901	602.4103	662.4357	664.2533	674.4204
DA	68.6275	70	65.3061	70.8333	68.0851
MAPE	**1.412**	**1.6573**	**1.8251**	**1.8395**	**1.866**
RMSE	890.2942	904.6706	935.3875	912.2696	930.4407

Fig. 13. MRF Training Evaluation without Lag or Overlay

Target : Close	1-step-ahead	2-steps-ahead	3-steps-ahead	4-steps-ahead	5-steps-ahead
N	52	51	50	49	48
MAE	34993.8172	46980.7269	41545.2501	34931.2479	33772.929
DA	52.9412	50	44.898	41.6667	42.5532
MAPE	**57.8142**	**77.4942**	**68.6221**	**57.8177**	**55.8681**
RMSE	40817.2046	54498.0634	47490.4522	39248.7686	37581.2205

Fig. 14. MRF Testing Evaluation without Lag or Overlay

Target : Close	1-step-ahead	2-steps-ahead	3-steps-ahead	4-steps-ahead	5-steps-ahead
N	92	91	90	89	88
MAE	319.6083	317.4939	316.4265	321.2438	322.2105
DA	80.2198	81.1111	80.8989	80.6818	81.6092
MAPE	**0.8503**	**0.8407**	**0.837**	**0.8493**	**0.8518**
RMSE	579.3446	583.6484	581.9741	582.9592	582.1676

Fig. 15. MRF Training Evaluation with Lag and Overlay

Target : Close	1-step-ahead	2-steps-ahead	3-steps-ahead	4-steps-ahead	5-steps-ahead
N	52	51	50	49	48
MAE	1125.5657	1093.6647	1081.0049	1087.2904	1098.9804
DA	92.1569	92	91.8367	91.6667	91.4894
MAPE	**2.0285**	**1.9739**	**1.9486**	**1.9552**	**1.9737**
RMSE	1353.078	1308.1326	1281.8951	1281.8068	1286.1451

Fig. 16. MRF Testing Evaluation with Lag and Overlay

Target : Close	1-step-ahead	2-steps-ahead	3-steps-ahead	4-steps-ahead	5-steps-ahead
N	28	27	26	25	24
MAE	0.0224	0.0226	0.0131	0.0134	0.0131
DA	96.2963	96.1538	96	95.8333	95.6522
MAPE	**0.048**	**0.0482**	**0.0266**	**0.027**	**0.0262**
RMSE	0.0529	0.0537	0.015	0.0153	0.015

Fig. 17. NAV prediction - Training data

Target : Close	1-step-ahead	2-steps-ahead	3-steps-ahead	4-steps-ahead	5-steps-ahead
N	8	7	6	5	4
MAE	0.1642	0.1788	0.1984	0.1996	0.121
DA	85.7143	83.3333	80	75	100
MAPE	**0.2972**	**0.3228**	**0.3577**	**0.3595**	**0.2172**
RMSE	0.2244	0.2388	0.2567	0.2677	0.1533

Fig. 18. NAV prediction - Testing data

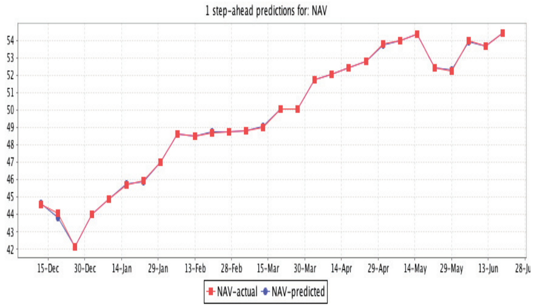

Fig. 19. Graph for Training model

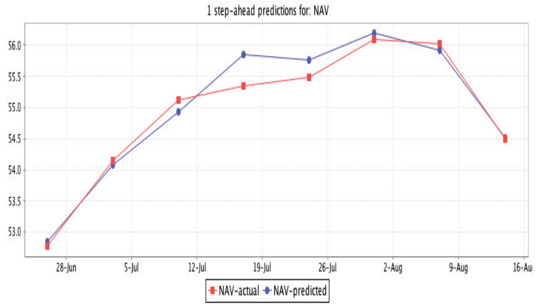

Fig. 20. Graph for Testing model

to a rather low MAPE in figure 7 representing the use of Lag and Overlay, at around 0.66%.

Analyzing the test data performance, the SVM model without any Lag or Overlay performed rather poorly. As figures 2 and 6 show that there is a high error percentage of 41%, which makes the predictions nearly meaningless. On the other hand, figures 4 and 8 describe a highly accurate scenario with an MAPE of around 0.9%.

2) MRF: Figures 9 and 11 represent the training data predictions as compared to actual values. The first observation is that the graph describing the training model without any Lag or Overlay misses most of the local minimums and maximums. The model trained with Lag and overlay performs closer predictions for training data. From figure 13 the MAPE is calculated as over 1.8% for without lag and overlay whereas in figure 15 the evaluation of the SVM model with Lag and Overlay yields improved results for training data with MAPE at around 0.85%

Similarly, figures 10 and 14 show an extremely poor test data analysis at peaking values of MAPE. The model fails to follow the trend and shows divergence. However, figures 12 and 16 show major improvement in prediction and follow convergence in actual and predicted values to an extent. The MAPE is only a mere 2%.

C. NAV prediction

Individual stocks and further on NAV were predicted using SVM predictions in cascade.

The stock prediction was performed on the top 30 stocks invested in by the HDFC mid cap opportunities - growth scheme. These predictions were used in the next stage for the prediction of NAV.

Different parameters for overlay and macroeconomic factors were used for each stage.

As can be seen the MAPE is as low as 0.3%.

Figures 17-20 indicate the results of NAV prediction.

V. Conclusion

The paper implements the cascaded SVM model (SMO regression) based ML system for the prediction of mutual fund NAV of the scheme HDFC mid cap opportunities - Growth. The accuracy of the test result in terms of MAPE is between 0.29% - 0.36%.

The algorithm proves higher efficiency, provides flexibility and with minor modifications, fund managers can maximize returns.

References

[1] S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, and K. R. K. Murthy, ”Improvements to the SMO Algorithm for SVM Regression,” *IEEE Transactions on Neural Networks*, vol 11, no. 5, Sept 2000.

[2] Alex J. Smola and Bernhard Schlkopf, ”A tutorial on Support Vector Regression,” *RSISE*, Australian National University, Canberra 0200, Australia, Max-Planck-Institut fr biologische Kybernetik,72076 Tbingen, Germany, Statistics and Computing 14: 199-222, 2004.

[3] S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, and K. R. K. Murthy, *Improvements to Platt's SMO Algorithm for SVM Classifier Design*, Control Division, Department of Mechanical and Production Engineering, National University of Singapore, Singapore

[4] Nikola Milosevic, ”Equity Forecast: Predicting long term stock price movement using machine learning,” School of Computer Science, University of Manchester, UK.

[5] C. M. Anish and Babita Majhi, ”Net Asset Value Prediction using FLANN model,”, *International Journal of Science and Research*, ISSN(Onine):2319-7064 Dept. of Computer Science and IT, Guru Ghasidas Vishwavidyalaya, Central University, Bilaspur, Chattisgarh, India.

[6] E. Priyadarshini and Dr. A. Chandra Babu, ”Prediction of the Net Asset Values of Indian Mutual Funds using Auto-Regressive Integrated Moving Average(ARIMA),” *Journal of Computer Applications Research and Development(JCARD)*, ISSN 2248-9304(Print), ISSN 2248-9312(Online), vol 1, no. 1, April-June, 2011.

[7] Ian H. Witten and Eibe Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Edition, Elsevier, 2008.