Stock Market Forecasting using Time Series Analytics with SVR

Supreet Kaur, Mr. Sachin Majithia

Abstract—**The stock market prediction is the strategic approach to estimate the variations in the volatile stock market. The stock market prediction enables the investors to plan their investment ahead of time, which is usually practiced to reduce the investment risk and yield the maximum profit. The stock market prediction is enabled using the curve fitting using the regression mechanism or classification mechanisms based upon the versatile classification models. The existing Multi classifier ensemble (MCS) uses bagging and boosting records the accuracy of 88.2%. The proposed model has been designed by using the SVM based regression model, which utilizes the support vector regression (SVR) module on specific paradigm. The proposed model has attained the results with higher accuracy than 90% on all of the intervals. The regression models including the polynomial regression, linear regression and support vector regression (SVR) models have been tested under this experimental model. In this paper, it is found that SVR has been discovered as the most efficient on most of the intervals, whereas on the contrast the maximum number of anomalies has been discovered under SVR. The performance fixation includes the incorporation of the linear regression, SVR and polynomial regression on the final stage of assessment.**

Index Terms*— *Backtseting, Stock market prediction, Machine learning, Value prediction.

1. Introduction

The proposed model has been designed for the stock market price evaluation of the response on the popular stocks. The proposed model is designed for the automatic stock price assessment and automatic price prediction. The results have been obtained from the proposed model. The proposed model have been given thread from the stock market historical data archive source, from where the data has been obtained in the excel format and further converted to the MATLAB supported format. The stock market price prediction and curve analysis has been performed on the given data file containing previous prices of the stocks in the last 6 years starting from 1st January, 2010 to 31st December, 2016, which makes the data of five years in the continuous series. The proposed algorithm returns the positive, negative and neutral trends after analyzing the historical stock data. The stock prices are calculated by performing the overlapping segmental analysis over the given historical data using the Self Organizing Maps (SOM) and support vector machine's (SVM) derivative support vector regression (SVR). The continuous weightage assessment has been performed in order to assess the prices in the small segments, which creates the base data for the next step in the hierarchical predictive analysis. The dataset has been obtained from the online archive source of S&P500 and deeply analyzed with the predicted price, previous price, change percentage, negative, positive or neutral trends and price curve for 6 years (2010-2016). The proposed model has been designed using the three predictive models of support vector regression (SVR), polynomial regression and linear regression.

Manuscript received September, 2017.

Supreet Kaur*, Department of Information Technology, Chandigarh Group of Colleges, Landran, Mohali.*

Mr. Sachin Majithia*, Assistant Professor, Department of Information Technology, Chandigarh Group of Colleges, Landran.*

2. LITERATURE REVIEW

Chen, Chen et al. [2] has worked on exploiting social media for stock market prediction with Factorization Machine. Later although financial news is proposed to access market information, there are some disadvantages for news to predict the stock market. Recently when the micro-blogging service has grown to a popular social media and provides a number of real-time messages for a lot of users, social media is proposed for the stock market prediction.

Xu Feifei and Vlado Keelj [13] have worked on Collective Sentiment Mining of Microblogs in 24-Hour Stock Price Movement Prediction. The authors have proposed a method for collective sentiment analysis for stock market prediction and analyze its ability to predict the change of a stock price for the next day. The proposed method is a two-stage process, based on the latest natural language processing and machine learning algorithms. Their evaluation shows best performance with the SVM approach in sentiment detection, with accuracy rates of 71.84/74.3% for positive and negative sentiment, respectively. The results of sentiment analysis are used in predicting stock price movement (up or down), and we found that users' activity on Stock Twits overnight positively correlates with stock trading on the next business day.

Soujanya et al. [10] has proposed the use of Sentic patterns for the purpose of the sentiment analysis from the social data. The authors proposed the use of dependency-based rules for concept-level sentiment analysis. In this work, the authors have introduced a novel paradigm to concept-level sentiment analysis that merges linguistics, common-sense computing, and machine learning for improving the accuracy of tasks such as polarity detection.

Yassine, Mohamed et al. [14] has worked on the development of a framework for emotion mining from text in online social networks. This paper presents a new perspective for studying friendship relations and emotions' expression in online social networks where it deals with the nature of these sites and the nature of the language used.

Vivek Narayanan [11] has worked on a fast and accurate sentiment classification using an enhanced Naive Bayes model. The authors have explored different methods of improving the accuracy of a Naive Bayes classifier for sentiment analysis. They have also observed that a combination of methods like effective negation handling, word n-grams and feature selection by mutual information results in a significant improvement in accuracy.

Sasan Barak [7] designs a fusion model for returns and risk prediction of stocks in financial market by applying various diversity methods in order to achieve more precise predictions for considering the simultaneous risk and return prediction of stocks for developing a base classifier selection procedure from candidate procedures by dataset clustering and considering the accuracy of combined classifiers. Developing a wrapper-GA scheme for feature selection and pre- diction and comparing it with the fusion method

Jigar Patel et al. [4] has worked towards the prediction of the stock market prices using the fusion of machine learning models. The paper focuses on the task of predicting future values of stock market index. Two indices namely CNX Nifty and S&P Bombay Stock Exchange (BSE) Sensex from Indian stock markets are selected for experimental evaluation.

The following conclusions have been drawn from the literature survey:-

3. The existing prediction models require the well define features for the prediction of the stock prices using historical stock market data.
4. The data preparation for the stock market prediction is very important after extracting the descriptive features by observing the moving average in the different block length which yields the balanced results between 20 and 100 days which is not done in the existing system.

5. PROBLEM DEFINITION

From the literature study, we have observed maximum overall accuracy of the existing fusion model of multi classifier ensemble system (MCS) [8] at 88.2%, which can be further improved. The existing Fusion of MCS model does not utilize any of the textual information, which primarily involves the news data, social media data, etc. The existing prediction models require the well define features for the prediction of the stock prices using the historical stock market data. The length of the historical data is very important for the prediction of the stock data. The data preparation for the stock market prediction is very important, which must be prepared after extracting the descriptive features.

6. PROBLEM FORMULATION

The existing model is based upon the multi-classifier ensemble, where the classification algorithms of Bagging, Boosting and Adaboost are utilized, for the purpose of stock price prediction. The hybrid classification model in the existing model relies upon the ensemble data without the descriptive analysis of the input data using any of the feature descriptor, which does not let the existing model to neutralize the independent variables for the high accuracy models. The use of robust and flexible feature descriptor, such as Self Organized Maps (SOM), for description of the features from the input training and testing data can further improve the overall performance of the proposed model. The amalgamation of the SOM and Support vector machine (SVM) will be used in the proposed multi-classifier ensemble. For the purpose of result fusion by the multi-classifier ensemble, the SOM-SVM model can be used for the preparation of final classification results. The SVM has been utilized to predict the features over the given time line from the time series data. Finally, the application of SOM would be involved to determine the realistic patterns to determine the stock trend, which can indicate whether stock will rise or fall.

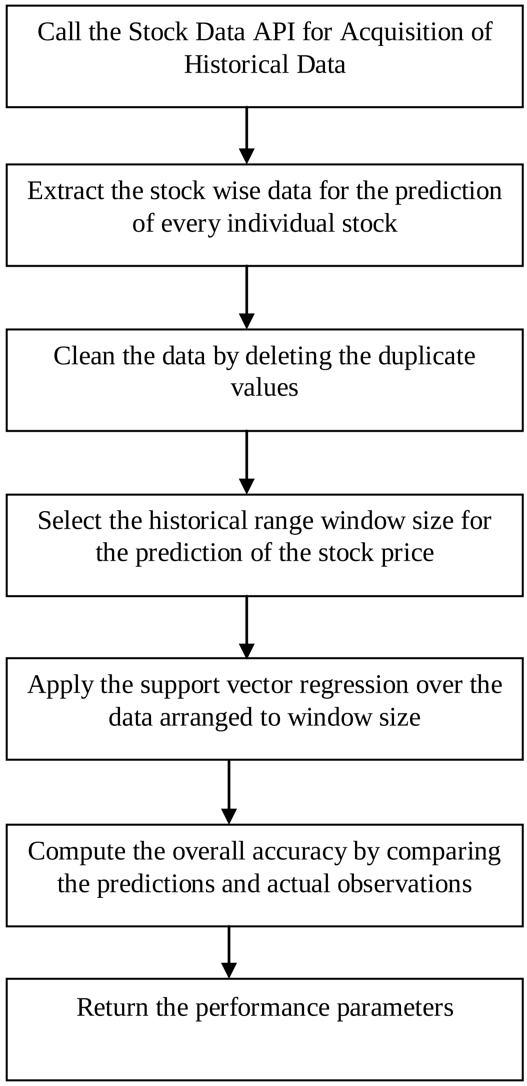

Figure 1: Flow Chart of Proposed work

7. EXPERIMENTAL DESIGN

The proposed model is designed for the prediction on the time-series data obtained from the live stocks API (application programming interface) from GOOGLE, which can be targeted for the selective stocks in one API call. The regression model has been applied over the time series data in the specific hierarchy, which mandates the preprocessing, data cleaning, missing value fixation and application of the support vector regression model for stock value prediction. The following algorithm describes the overall working of the proposed model using SVR:

Algorithm 1: Support Vector Model based Predictive Analysis on Selective Stocks

8. Initialize the tickers containing the stock names, {'AAPL', MSFT', 'SPY']
9. Initialize the data source ['GOOGLE']
10. Initialize the start and finish date to acquire the time series data ['START DATE' 'FINISH DATE']
11. Acquire the data from the API, API_CALL [tickers, data source, start date, finish date]
12. Obtain the adjacent close values for each stock from the acquired data for all weekdays
13. Extract the specific stock data from the stock data matrix, ['MSFT']
14. Fix the "Not a number" values in the extracted stock market vector
    a. If iteration index is lower than 3
        i. If current value is "NaN"
            1. Fill zero in place of NaN
    b. Otherwise
        i. Extract the three value starting from current value, initial counter at (current-3) to initial
        ii. Run the iteration over the mini vector acquired on step 7(b)(ii)
            1. If current value is NaN
                a. Save the value id in the mini vector
            2. Otherwise
                a. Add the current value total sum of vector, sumVec = sumVec + Current Value
        iii. Compute the average value of mini vector, mvAvg = sumVec / No of Values
        iv. Fill all mini vector values marked on 7 (b) (ii)(1) with average value (mvAvg) on step 7(b)(iii)
        v. Restore the mini vector to the output vector
15. Apply the support vector regression (SVR) over the MSFT vector after fixation
16. Return the prediction from the regression model
17. Save the regression result

VI. RESULT ANALYSIS

The proposed model has been analyzed for the accuracy of the predictions among the proposed prediction model. The proposed model has been designed by combining the SVM with feature descriptor extraction, missing value fixing and data cleaning in order to acquire the balanced and best value data as the training matrix. The results of the proposed model has been analyzed in the form of predicted values of the stocks in the next day of trading, which has undergone the proposed model's SVR model as well as polynomial and linear regression models. The results are shown in detail in the following:

Table 1: Table of Predictions and Actual Values

Index	Upper Bound Limit (Data Size: 100 days)	Original Value	Support vector regression	Polynomial Regression	Linear Regression
0	100	26.84	26.55	27.21	29.51
1	101	26.27	27.26	26.70	29.42
2	102	26.07	25.17	26.13	29.30
3	103	25.01	26.17	25.60	29.18
4	104	26.00	23.77	24.91	29.03
5	105	25.80	29.31	24.56	28.92
6	106	17.27	23.28	24.22	28.80
7	107	25.89	**2.70**	21.96	28.37
8	108	26.46	**38.23**	22.02	28.27
9	109	26.86	24.21	22.22	28.20
10	110	25.79	27.36	22.52	28.14
11	111	25.29	24.93	22.58	28.04
12	112	25.11	25.35	22.54	27.94
13	113	24.79	25.62	22.49	27.82
14	114	25.00	24.72	22.40	27.71
15	115	25.66	25.44	22.40	27.60
16	116	25.50	26.68	22.56	27.51
17	117	26.58	24.96	22.70	27.43
18	118	26.32	29.41	23.07	27.38
19	119	26.37	24.23	23.36	27.33
20	120	26.44	27.16	23.65	27.28
21	121	25.95	26.95	23.94	27.23
22	122	25.77	25.00	24.12	27.17
23	123	25.31	26.12	24.25	27.11
24	124	25.00	**25.05**	24.30	27.03
25	125	24.53	**24.63**	24.29	26.94
26	126	24.31	24.64	24.20	26.85

27	127	23.31	25.00	24.09	26.74
28	128	23.01	22.37	23.81	26.61
29	129	23.16	23.67	23.52	26.48
30	130	23.27	23.89	23.30	26.35
31	131	15.48	23.03	23.14	26.23
32	132	23.82	2.82	21.54	25.88
33	133	24.30	39.45	21.68	25.78
34	134	24.41	22.73	21.92	25.71
35	135	24.27	24.60	22.17	25.63
36	136	24.83	25.34	22.39	25.56
37	137	25.13	25.70	22.70	25.51
38	138	25.44	**25.24**	23.05	25.46
39	139	25.51	**25.47**	23.43	25.43
40	140	24.89	25.89	23.80	25.39
41	141	25.23	24.30	24.04	25.35
42	142	25.48	26.42	24.32	25.31
43	143	25.12	**25.07**	24.62	25.28
44	144	25.84	23.68	24.84	25.24
45	145	25.81	28.03	25.17	25.22
46	146	26.10	23.93	25.46	25.21
47	147	26.16	26.88	25.79	25.20
48	148	25.95	26.61	26.09	25.19
49	149	26.03	25.38	26.34	25.18

In the table 1, the result analysis shows the predicted values by the different testing models. The squared difference of the predicted and original values has been recorded in the table 2 between the original with SVR, Linear regression and Polynomial regression models. The original values in table 1 are explained with the help of table 2 in the following discussion. The testing models in the experimental results include the support vector regression (SVR), polynomial regression and linear regression models. In the case of SVR model, the indices on the positions 7, 8, 31, 32 and 33 shows the anomalies with higher differences between the original observations and predicted values by the proposed SVR model over the data of 100 days arranged in the linear formation with slider window function. The predicted values on the indices of 24, 25, 38, 39 and 43 are predicted with the minimum difference, which does not exceed the maximum difference of 0.04 as squared difference. The polynomial linear prediction model has been recorded with five anomalies on the indices (6, 7, 8, 9 and 31), which is equal to the anomalies counted for SVR model. The nearest prediction by linear model are recorded on the indices 2, 25, 26, 30, and 48, out of which the maximum difference has been recorded at 0.06 as squared difference, which is significantly higher than the SVR model. The linear prediction model has been found most reliable in our case, which has been recorded in two values on 28th and 31st indices in the case of anomalies. But the nearest matching values are obtained in three indices (38, 39 and 41), where maximum difference has been found at 0.01 as squared difference The linear regression model can be declared as the most significant model, if the assessment is based upon the count of anomalies, whereas the highest number of matching models elaborates the support vector regression (SVR) model, which can be declared as most significant predictive model.

Table 2: Squared Sum of Prediction v/s Actual value of next day stock value

Index	SVR	Polynomial Regression	Linear Regression
1	0.09	0.14	7.14
2	0.97	0.19	9.9
3	0.8	0	10.44
4	1.35	0.35	17.42
5	4.98	1.2	9.17
6	12.32	1.55	9.71
7	36.14	48.41	133
8	537.86	15.41	6.14
9	138.62	19.75	3.28
10	7.04	21.51	1.78
11	2.46	10.67	5.52
12	0.13	7.37	7.59
13	0.06	6.62	7.99
14	0.69	5.3	9.21
15	0.08	6.75	7.32
16	0.05	10.64	3.75
17	1.39	8.64	4.06
18	2.62	15.09	0.72
19	9.57	10.59	1.13
20	4.57	9.06	0.92
21	0.53	7.77	0.71
22	1	4.03	1.65
23	0.6	2.73	1.97
24	0.66	1.12	3.23
25	0	0.5	4.12
26	0.01	0.06	5.83
27	0.11	0.01	6.43
28	2.85	0.61	11.8
29	0.41	0.64	12.99
30	0.26	0.13	11.01
31	0.38	0	9.49
32	57.11	58.73	115.62
33	440.92	5.22	4.22
34	229.63	6.85	2.2
35	2.84	6.19	1.68
36	0.11	4.39	1.86
37	0.26	5.95	0.53
38	0.32	5.89	0.14
39	0.04	5.7	0
40	0	4.31	0.01
41	0.99	1.18	0.25
42	0.86	1.42	0.01
43	0.89	1.36	0.03
44	0	0.25	0.03
45	4.67	1	0.36
46	4.92	0.41	0.34
47	4.72	0.4	0.8
48	0.52	0.14	0.93
49	0.43	0.02	0.58
50	0.43	0.1	0.73

The table 2 shows the count of anomalies as well as the count of most matching events in the predictive model with the squared distances. The squared difference values have been obtained by computing the squared distances between the original values on the next day of the stock window (100 days), and the predicted values by support vector regression (SVR), polynomial regression and linear regression models.

VII. CONCLUSION

In this paper, the regression models are implemented in order to build the predictive model to predict the stock prices in the given stock windows, which has been sustained at the windows of 100 days. The window of 100 days has been undergone the slider window function, which selected the recent 100 days to predict the value of next trading day on the stock market. The stocks of Microsoft, Google and Apple has undergone the analysis, and the final results are predicted over the historical stock data of Microsoft, which has been obtained between "01-01-2010" and "31-12-2016", which includes the total data of six years. The proposed model has been obtained with the different number of anomalies and nearest predictions. In the case of SVR and polynomial regression, five anomalies and five nearest predictions has been found as top 10% and bottom 10% indicators, whereas in the case of linear regression only two anomalies and three nearest values in the case of top 10% and bottom 10% indicators. The maximum squared difference of 537.86 has been recorded in the case of SVR against the 58.73 for polynomial regression and 115.62 for linear regression.

ACKNOWLEDGMENT

With profound gratitude and due regards, I whole heartedly and sincerely acknowledge the efforts, encouragement and proper guidance by Mr. Sachin Majithia, Assistant Professor, Department of Information Technology, Chandigarh Group of colleges, Landran, Mohali, Punjab.

REFERENCES

[1] Blei, D.M.A.M, "Supervised topic models," In Proceedings of NIPS, 2007, 121- 128.

[2] Chen, Chen, Wu Dong Xing, Hou Chunyan, and Yuan Xiaojie, "Exploiting Social Media for Stock Market Prediction with Factorization Machine," In Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on *Web Intelligence (WI) and Intelligent Agent Technologies* (IAT)-Volume 02, IEEE Computer Society, 2014, pp. 142-149.

[3] E. Bothos, D. Apostolou, G. Mentzas, "Using Social Media to Predict Future Events with Agent-Based Markets," *Intelligent Systems*, IEEE, 2010, 50-58.

[4] Jigar Patel, Sahil Shah, Priyank Thakker, K. Kotecha, "Predicting Stock Market index using fusion of machine learning techniques," *Expert System With Application*, V.N- 42, 2014.

[5] P. Chang, D. Wang, C. Zhou, "A novel model by evolving partially connected neural network for stock price trend forecasting," *Expert Systems with Applications*, 39 (1), 2012, 611-620.

[6] Ronen Feldman, "Techniques and Applications for Sentiment Analysis," Communications of the ACM, Vol. 56 No. 4, 2013, pp. 82-89.

[7] S. Barak, J. H. Dahooie and T. Tichy, "Wrapper ANFIS-ICA method to do stock market timing and feature selection on the basis of japanese candlestick," *Expert System Application,* vol. 42, 2015, pp. 9221-9235.

[8] Sasan Barak, Azadeh Arjmand, and Sergio Ortobelli, "Fusion of multiple diverse predictors in stock market," *Springer*, Information Fusion 36 (2017): 90-102.

[9] S. Asur, Huberman, B.A, "Predicting the future with social media," In Proceeding of IEEE/WIC/ACM International Conference on *Web Intelligence and Intelligent Agent Technology* (WI-IAT), 2010, 492-499.

[10] Soujanya Poria, Erik Cambria, Grégoire Winterstein, and Guang-Bin Huang, "Sentic patterns: Dependency-based rules for concept-level sentiment analysis," *Knowledge-Based Systems* 69 (2014): 45-63.

[11] Vivek Narayanan, Ishan Arora, and Arjun Bhatia, "Fast and accurate sentiment classification using an enhanced Naive Bayes model," In *Intelligent Data Engineering and Automated Learning–IDEAL 2013*, Springer Berlin Heidelberg, 2013, pp. 194-201.

[12] W. Dai, J. Wu, C. Lu, "Combining nonlinear independent component analysis and neural network for the prediction of Asian stock market indexes," *Expert Systems with Applications*, 39 (4), 2012, 4444-4452.

[13] Xu Feifei, and VladoKeelj, "Collective Sentiment Mining of Microblogs in 24-Hour Stock Price Movement Prediction," *In Business Informatics (CBI)*, 2014 IEEE 16th Conference, vol. 2, IEEE, 2014, pp. 60-67.

[14] Yassine, Mohamed and Hazem Hajj, "A framework for emotion mining from text in online social networks," In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, IEEE, 2010, pp. 1136-1142.

Supreet Kaur is a student of M.Tech (IT) in Department of Information technology, Chandigarh Group of Colleges, Landran carrying out her research work under the guidance of Mr. Sachin Majithia. Her main research interest is in data mining.

Mr. Sachin Majithia is presently serving as Assistant Professor in Department of Information Technology, Chandigarh Group of Colleges, Landran. His key research areas are Database Management and Storage. He has published more than 50 papers in various journals and conferences of international and national repute