Stock Price Forecasting Framework based on the Support Vector Regression and Monte Carlo Method

M V Kamal, D Vasumathi

Abstract: ***Stock market and its prices prediction are considered as one of the challenging task in financial forecasting. In my research, the framework is created based on the support vector regression (SVR) and Monte Carlo method to predict the stock price. The radial basis function (RBF) has high capacity, simpler design, and adopted for kernel function in SVR. The stock price of four companies Microsoft, Facebook, Amazon and Google is used to analyze the efficiency of the proposed method. The different parameters like mean square error (MSE), mean absolute error (MAE) measured to estimate the outcome of the proposed method. The experimental result showed the efficiency of the SVR-Monte Carlo in terms of error value. The MSE for the SVR-Monte Carlo in Google stock obtained as 0.2162 and MAE for the predicted value is 0.0164.***

Keywords - Mean Square Error, Monte Carlo, Radial Basis Function, Stock market price prediction, and Support Vector Regression.

I. INTRODUCTION

The process of Web Mining is to extract the pattern from the web and it is classified as three types, namely: web structure mining, web usage mining and web content mining [1]. Web usage mining is defined as the extraction of the meaningful user patterns from web server access logs by using data mining techniques [2]. Data mining is a major part of the knowledge discovery in economics, finance, telecommunication, medicine, and various scientific fields. Data mining can discover the hidden information from a large amount of data that are important for identification of crucial relationships, patterns, facts and trends. Data belongs to the various fields are generated in various devices, transactional applications, networks, sensors, and social media[3].Usually, the data present in the large size and in heterogeneous form. There are different methods available to store and analyze these data, but it doesn't handle large-volume of data [4].The most common methods in data mining are unable to manage large amount data by the conventional ways. There are various hybrid methods used for handling large amount of data. The goal is to build the data mining method more effective and simpler to handle by the professionals in the business intelligences and data scientists. This allows to develop the convenient model for data mining process by incorporating many methods [5].

Business intelligence (BI) is software that helps in the tactical planning processes of the corporation. Most companies collect the data from their business functions and the BI web mining tool collect that information in time series manner. BI software provides the growth potentials based on the search criteria [6-9]. Business growth of the enterprises majorly based on the customer centric and need to understand the customer's requirement to succeed. The business community needs the BI to have the expert decision beyond the focusing customer relationship management. The stock market analysis is one of the important process in the BI and there is the need of effective methods for stock market analysis and prediction[10].In this research, the Support Vector Regression (SVR) and Monte Carlo is used to forecast the stock market value. The SVR-Monte Carlo evaluated one-year data of four companies such as Amazon, Google, FB and Microsoft. The risk factor is measured for the four companies and the performance is investigated. A number of different parameters measured from the output of the SVR-Monte Carlo and this showed the performance of the SVR-Monte Carlo.

The organization of the paper as follows, literature Review is presented in the Section 2, the proposed method is explained in Section 3, and the experimental result is given in Section 4. The conclusion of this paper is made in section 5.

II. LITERATURE REVIEW

Stock Market prediction (SMP) is one of the crucial tasks for financial managers, which in need of the robust prediction. The current research related to the SMP is surveyed in this section.

Xi Zhang, *et al*. [11] investigated the relationship between various data sources, and developed a multi-source multiple instance mode that combines event, sentiments, as well as the quantitative data into a comprehensive framework. The event extraction and representation model used to capture the news events. This method is capable of analyzing the importance of the data source and considers the crucial data as input and predict interpretable. The optimization technique can be used to increase the effectiveness of the stock prediction.

Nuno Oliveira, *et al*.[12] presented a model to predict the stock market variables and information extracted from the micro-blog of twitter. The twitter dataset collected from the year of December 2012 to October 2015, contains about 31 million messages related to 3,800 stocks traded in the US market. This method also processed a prediction procedure of four regression methods and rolling window evaluation with a statistical test of predictive accuracy. The traditional sentimental analysis assesses their data based on the micro-blogging sentiment.

Revised Version Manuscript Received on 16 September, 2019.
M V Kamal, Dept of CSE,(Research Scholar), JNTU-Hyderabad, India.
D. Vasumathi, Dept of CSE, JNTU-Hyderabad, India.

The Kalman Filter provided a unique daily sentiment indicator from a twitter and four other sentiment indicator for the analysis. This method outperformed two state-of-art method in sentiment indicator using micro-blogging data. The efficiency of the proposed method is high compared to the state-of-art method in the stock prediction. The stock data along with microblog data can be used to increase the efficiency of the prediction.

Bin Weng, *et al*. [13] developed a method combining data which is collected from online, with traditional time series and technical analysis for stack that can provide more efficient and intelligent daily trading expect system. The three machine learning methods such as decision tree, Support vector machine and neural networks used for "inference engine". The case study of AAPL (Apple NASDAQ) stock used to measure the performance of this expert system. This method achieved 85% accuracy in forecasting the next day stock value. The feature selection technique helped to select the relevant feature for the machine learning and minimized the large number of data without loss of information. The different types of data need to incorporate in this method for effective analysis.

Xi Zhang, et al., [14] created a coupled matrix and tensor factorization architecture based on the event extracted from the online news and the users' sentiments from social media, then used the information for the SMP. This model predicted the multiple correlate stocks simultaneously based on the commonalities and factorized the low rank matrices between tensor and matrices. Various methods utilized to measure the correlation between the stock to deal with the data sparsity issues. A number of features considered for the coupling effects to measure the stock correlation. This method attained the accuracy of 62.5 % and 61.7 %. The efficiency of the proposed methods need to be improved.

Ahmad Kazem, et al., [15] established a forecast model which is based on the chaotic mapping, firefly algorithm and SVR to forecast the stock market price. The chaotic firefly algorithm was used to optimize the hyper parameter of SVR. Then, optimized SVR was used to forecast the SMP. Mostly, the genetic algorithm was applied to optimize the parameter for machine learning in SMP and this method follows the chaos theory and the firefly algorithm to optimize the hyper parameters of SVR. The delay in coordinate embedding method was used to reconstruct the phase space dynamics. The high predictive accuracy was achieved by the structural risk minimization (SRM). The stock market data from the NASDAQ historical quotes namely National Bank shares, Intel and Microsoft daily closed stock price, collected and applied to the optimized SVR data. A chaotic mapping operator helped to increase the search quality of the firefly algorithm in the search space by increasing the quality of the generated population in the initial stage and prevented it from local optima. The error value needs to minimize in the method and parameter optimization is to be improved.

The efficiency of the existing method in semantic analysis is low due to the dynamic nature of the information on the web. Therefore, the proposed system analyses the stock prediction using Support Vector Regression and the risk factor is predicted by the Monte Carlo method.

III. PROPOSED METHOD

The major objective of the BI is to transform the knowledge into the better decision-making process. The stock market forecasting is the important part in the BI and a lot of methods proposed for this analysis. The main objective of this research is to effectively forecast the stock market for better financial planning. The SVR analysis the time series stock data and predict the stock price of the several companies. The Monte Carlo technique is used to forecast the risk factor of the stock and the risk factor is given as input to the SVR. The state-of art method is compared with the SVR-Monte Carlo method to investigate the efficiency of the proposed method. The real time stock data mined from the web for the specific companies such as Facebook, Google, Microsoft and Amazon to predict the stock. The block diagram of the SVR-Monte Carlo method is shown in the Fig. 1.

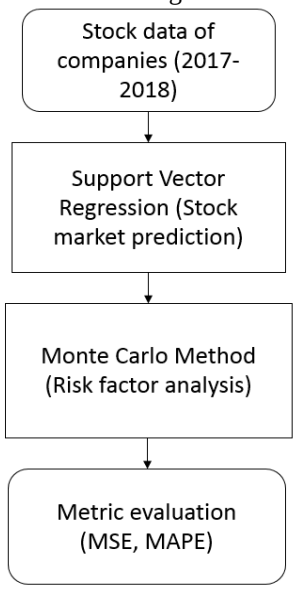

Fig 1. The block diagram of the SVR-Monte Carlo method

A. Support Vector Regression

Assume the given input a $(x_1, y_1), \ldots, (x, y)$, where $x_i \in R_n, i = 1,2, \ldots$ and $y_i \in R$ is the target value each input vector x_i.These patterns are used to train the regression model and predict the feature value. SVR is the non-linear regression method that is applied to identify the best regression hyperplane value in the high dimensional feature space [15].

The popular technique in SVRs is ϵ-SVR that finds the hyperplane with a ϵ-insensitive loss function [15]. The SVR is represented in the Eq. (1).

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

In the equation (1), $\varphi(x)$ is a nonlinear mapping from the input data to the feature space. Where w is a vector weight coefficients and bias constant is denoted as b. The w and b are calculated by reducing the optimization problem in the Eq. (2).

$$\text{Minimize}\ \frac{1}{2}\left||w|\right|^2$$

$$\text{Subjected to}\ \begin{cases} y_i - (< w, \varphi(x_i) > + b) \le \epsilon \\ (< w, \varphi(x_i) + b) - y_i \le \epsilon \end{cases} \quad (2)$$

The point from the ϵ-insensitive band is not removed in this method to solve the feasibility issues and to increase the efficiency. The points penalize by using the slack variables ξ_i, ξ_i^*, in the Eq. (3).

$$\text{Minimize}\ \frac{1}{2}\left||w|\right|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*)$$

$$\text{Subjected to}\begin{cases} y_i - (< w, \varphi(x_i) > + b) \le \epsilon + \xi_i \\ (< w, \varphi(x_i) > + b) - y_i \le \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \quad (3)$$

In the Eq. (3), the cost constant $C > 0$ determine the trade-off between model training error and complexity, l is the number of training patterns [15]. After considering the Lagrangian and optimality conditions, then the model solution in dual representation can be found using the Eq. (4).

$$f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)K(x_i, x) + b \quad (4)$$

The α_i, α_i^* in the Eq. (4) are non-zero Lagrangian multiplies and the solution to the dual problem. The kernel function $K(x_i, x)$ denotes the inner product $< \varphi(x_i), \varphi(x) >$. The RBF used in this study as the kernel function due to its simplicity and capabilities.

$$K(x_i, x_j) = \exp\left(-\gamma \left||x_i - x_j|\right|^2\right) \quad (5)$$

The γ is the width parameter of RBF kernel in the Eq. (5) and selected based on its heuristics.

B. Monte Carlo method

After the analysis of stock data by the SVR, the risk factor is identified by the Monte Carlo (MC) method. The Monte Carlo is used to numerically approximate an integral using a random sample from the domain of the integrand. Monte Carlo method is especially useful when the integrand is having multiplicative of a known probability density function (PDF) with a finite second moment.

$$F(q) = \int I_A(\hat{b}) f\left(\hat{b} \middle| b, cov(\hat{b})\right) d\hat{b}, \quad (6)$$

$I_A(\hat{b})$ in the Eq. (6) is the indicator function that is defined in the Eq. (7).

$$I_A(\hat{b}) = \begin{cases} 1 & if\ \hat{b} \in A, \\ 0 & if\ \hat{b} \in A \end{cases}$$

$$\text{And}\ A = \{\hat{b} \in \mathbb{R}^m : g(\hat{b}) \le q\}. \quad (7)$$

The random draws of B is represented as $\hat{b}^1, \ldots, \hat{b}^B$ and then estimate the probability value p for the corresponding quantile q for the sample distribution of $g(\hat{b})$ is denoted in the Eq. (8).

$$\hat{p} = \frac{\sum_{j=1}^{B} I_A(\hat{b}^j)}{B} \quad (8)$$

The above result is guaranteed to converge to a true value according to the law of large numbers. In addition, the Monte Carlo error value is calculated from the Cumulative Density Function (CDF) in the Eq. (9).

$$SE(\hat{p}) = \sqrt{\sum_{j=1}^{B} \frac{(I_A(\hat{b}^j) - \hat{p})^2}{B(B-1)}} = \frac{\sqrt{\hat{p}(1-\hat{p})}}{B-1} \quad (9)$$

The accuracy of estimate $\hat{p}$ based on the number of simulation extract B. The sampling distribution of $g(\hat{b})$ is measured from the Monte Carlo, that finds the $(1-\alpha)\%$ CI from the Monte Carlo samples. The Monte Carlo sampling distribution from B measures $\hat{b}^1, \ldots, \hat{b}^B$. Then process the random draws u^i from the sampling distribution of $g(\hat{b})$ as follows: $u^i = g(\hat{b}^i)$. The quantiles identifies the q_L and q_U of $u^i s$ related to the lower and upper confidence limit probabilities of $\alpha/2$ and $1-\alpha/2$ respectively.

IV. EXPERIMENTAL RESULT

Stock market forecasting attained much attention in the recent times and many research have been made in stock market forecasting. The efficiency of the SMP needs to be increased for business planning. The objective of the research is to increase the efficiency of the SMP. One year of stock data for four companies such as Amazon, Facebook (FB), Google and Microsoft collected and used to train the SVM-Monte Carlo. The stock data are given as input to the SVR and the Monte Carlo method used to forecast the future stock data. The risk is analyzed for the four companies and error value calculated. The SVR-Monte Carlo was implemented in the tools of python 3.7 and JypitorLab. This method was executed in the operating system of Windows 10 with 8 GB RAM in the Intel i7 processor and 500 GB hard disk.

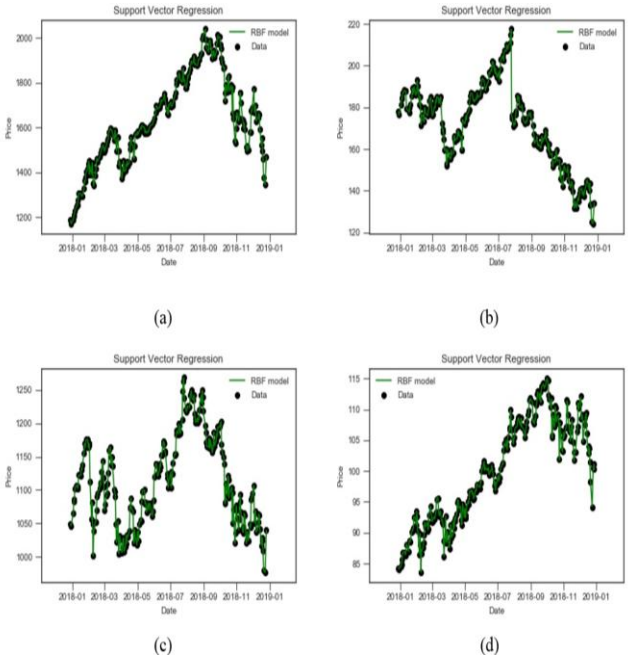

Fig 2. Stock Prediction (a) Amazon, (b) FB, (c) Google, and (d) Microsoft

The stock prediction of the four companies such as Amazon, FB, Google and Microsoft is shown in the Fig. 2 (a-d) respectively. The SVR process the data of stock and Monte Carlo predict the stock range of the four companies. The rise and fall of the predicted stock values of four company's predictions are shown in the Fig. 2(a-d).

The RBF model is considered as the kernel function in the SVR to predict the stock. The risk factor is analyzed from the predicted value for the stock values. The predicted value shows the increase in stock of three companies such as Amazon, FB and Google in the time of July. The Microsoft stock value predicted to be rising in high value in the time of September.

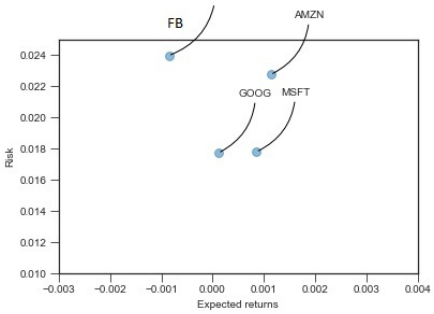

Fig 3. Risk analysis of the four companies

The risk value is analyzed from the predicted value and this provides the information when the higher risk to invest in the respective stock.The risk is analyzed for the four companies and shown in the Fig. (3).The risk value is higher for the FB and second higher risk is measured for the Amazon. The Google has lower risk to invest in its stock and Microsoft has the second lowest risk in the stock value.

Table 1. The parameter measures for the different companies

	Google	FB	MSFT	AMZN
Mean squared error	0.2162	0.2183	0.2141	0.2116
Root Mean squared error	0.4633	0.4653	0.4691	0.4608
Mean Absolute error	0.0164	0.0148	0.0161	0.0121

The different parameter such as Mean Square Error (MSE), Root Mean squared error (RMSE), and Mean Absolute error (MAE) measured from the output. The error value calculated for the four company's stock values. The three different parameters measured for the four company's SMP performance and shown in the Table 1.The error value obtained as low for the stock prediction. This shows that the SVR-Monte Carlo has lower error value in the stock prediction analysis.

V. CONCLUSION

The efficient SMP is in demand among the BI and various studies have been conducted for the SMP. The SMP is the difficult task due to various factors involves in the stock. This research involved in the SMP for one year using SVR and Monte Carlo method. The RBF used as the kernel in the SVR due to its simplicity and capacity and Monte Carlo is used to forecast the stock price. The risk factors are analyzed in this method for the different companies based on the twitter data. The stock data of three companies such as Google, FB, Amazon and Microsoft collected at the time of one year. The different parameters like MSE, MAE and RMSE measured from the outcome of the SVR-Monte Carlo method. The RMSE value of the proposed method in the Google stock prediction is 0.4633.The future work will be involved in using the different types of data, including the twitter feed of the stock to increase the performance of the stock prediction.

REFERENCES

[1] S. S. Ahila, and K. L. Shunmuganathan, "Role of agent Technology in Web Usage Mining: homomorphic encryption based recommendation for E-commerce applications," *Wireless Personal Communications*, vol. 87, no. 2, pp. 499-512, 2016.

[2] T. Tamilselvi, and G. Tholkappia Arasu, "Handling high web access utility mining using intelligent hybrid hill climbing algorithm based tree construction," *Cluster Computing*, pp. 1-11, 2018.

[3] D. G. Lorentzen, "Webometrics benefitting from web mining? An investigation of methods and applications of two research fields," *Scientometrics*, vol. 99, no. 2, pp. 409-445, 2014.

[4] Young-Joo Lee, and Ji-Young Park, "Identification of future signal based on the quantitative and qualitative text mining: a case study on ethical issues in artificial intelligence," *Quality & Quantity*, vol. 52, no. 2, pp. 653-667, 2018.

[5] V. Medvedev, O. Kurasova, J. Bernatavičienė, P. Treigys, V. Marcinkevičius, and G. Dzemyda,"A new web-based solution for modelling data mining processes," *Simulation Modelling Practice and Theory*, vol. 76, pp. 34-46, 2017.

[6] M. Vijaya Kamal, and D. Vasumathi, "Business Intelligence & Geo Tracking-A Novel Mining Technique to Identify Alerts and Pattern Analysis," *Computational and Business Intelligence (ISCBI), 2014 2nd International Symposium on*. IEEE, 2014.

[7] Moro Sérgio, Pao Cortez, and Paulo Rita, "Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation." *Expert Systems with Applications*, vol. 42, no. 3, pp. 1314-1324, 2015.

[8] M. A. Aufaure, R. Chiky, O. Curé, H. Khrouf, and G. Kepeklian, "From Business Intelligence to semantic data stream management," *Future Generation Computer Systems*, vol. 63, pp. 100-107, 2016.

[9] Mihaela Filofteia Tutunea, "Business Intelligence Solutions for Mobile Devices–An Overview," *Procedia Economics and Finance*, vol. 27, pp. 160-169, 2015.

[10] N. Pushpalatha, and S. Sai Satyanarayana Reddy, "Towards an extensible web usage mining framework for actionable knowledge," *Inventive Communication and Computational Technologies (ICICCT), 2017 International Conference on*. IEEE, 2017.

[11] Zhang, X., Qu, S., Huang, J., Fang, B. and Yu, P., 2018. Stock market prediction via multi-source multiple instance learning. *IEEE Access*, *6*, pp.50720-50728.

[12] Oliveira, N., Cortez, P. and Areal, N., 2017. The impact of microblogging data for stock market prediction: using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications*, *73*, pp.125-144.

[13] Weng, B., Ahmed, M.A. and Megahed, F.M., 2017. Stock market one-day ahead movement prediction using disparate data sources. *Expert Systems with Applications*, *79*, pp.153-163.

[14] Zhang, X., Zhang, Y., Wang, S., Yao, Y., Fang, B. and Philip, S.Y., 2018. Improving stock market prediction via heterogeneous information fusion. *Knowledge-Based Systems*, *143*, pp.236-247.

[15] Kazem, A., Sharifi, E., Hussain, F.K., Saberi, M. and Hussain, O.K., 2013. Support vector regression with chaos-based firefly algorithm for stock market price forecasting. *Applied soft computing*, *13*(2), pp.947-958.

AUTHORS PROFILE

M. V Kamal, who could complete B.E in CSE from Gulbarga University, and M. Tech in Software Engineering from JNTU Hyderabad has been pursuing Ph.D in Data Mining from the JNT University, Hyderabad. He is having 18 years of experience in academia. He has published several papers in both National and International Papers and attended several National and International Conferences and organized Seminars etc. His area of interest includes Data Mining and Information Security.

D Vasumathi, completed her B.Tech, and M.Tech from Jawaharlal Nehru Technological University Hyderabad. She did her Ph.D (Research) in the area of Data Mining from JNT University, Hyderabad. Presently she is working as Professor in Dept. of CSE, JNTUCEH and having more than 25 years of experience in teaching. She is a member for several professional bodies like CSI, IEEE and ISTE. She had presented and published several papers in National and International Conferences and also in IEEE Explorer. She was chair for several conferences. She did for Editorial board member for several papers of National and International events. Her area of interest includes, Data warehousing and Data Mining, Information Security and Information Retrieval Systems.