STOCK MARKET ANALYSIS USING MACHINE LEARNING

M. AKHIL KUMAR, D. SUJAN KUMAR
CVR College of Engineering, Hyderabad, Telangana, India
Department of Computer Science and Engineering
akhil4414@gmail.com, dsujankumar@gmail.com

Abstract: ***Many analysts and experts have long been aware of stock market forecasts. Time series prediction has been used extensively to evaluate future stock prices and to analyses, model and direct investor decisions and businesses through financial time series. This work provides an intelligent time series prediction framework which makes use of sliding windows optimization in order to predict stock prices. The algorithm proposed combines the support vector machine (SVM) and Least Squares SVM (LS-SVM). The model proposed is based on the review of inventory historical data and technical indicators. In contrast with the artificial neural network, financial data were used and tested for the proposed model. The findings showed a higher prediction accuracy in the proposed model.***

Keywords: Machine Learning, Least Squares Support Vector Machine, Data Pre-processing, Dataset, Stock Market.

I. INTRODUCTION

One of the most interesting inventions of our time is the financial markets. They have had a major impact on many fields, including industry, education, employment, technology and thus the economy. Over the years, investors and analysts have developed and tested stock price models. However, it is exceedingly difficult to evaluate stock market movements and actions because of complex, nonlinear, non-stop, non-parametric, noisy and chaotic market characteristiques. Accordingly, many highly inter-linked factors, including global, political, psychological and business variables, influence stock markets. Two primary approaches to financial markets research Technical and fundamental analysis Investors have employed these two primary approaches in decision-making for capital markets in order to invest in stocks and gain high profits with low hazards.

The stock price movement typically produces a linear curve over a long period of time. The stocks whose prices are to increase in the near future appear to be purchased by citizens. In the stock market volatility people will not encourage to invest in stocks. Thus, the stock market that can be used in a real-life situation must be correctly forecast. The methods of bursary predictions include time series forecasting, technical analysis, simulation of machine learning and prediction of the variable bursary. Details such as the price opening closing price, the data and many other variables needed for a variable object which is the price in a given day are included in the data sets of the stock market prediction model. Orthodox prediction techniques, such as multivariable regression with a time series model, were employed in the previous model. The stock market forecast is better when it is viewed as an issue with regression but is strong when it is viewed as classification. The goal is to design and quantify future trends of stock value growth through the use of machine learning strategies.

For classification and regression, the Support Vector Machine (SVM) may be used. SVMs have been shown to be used more efficiently in classification-based problem. The SVM method is a point in the n-dimensional space for each data component, with the value of the feature being the value of a specific coordinate, and hence classification is carried out by finding a hyperplane which explicitly differentiates the two classes.

The random forest algorithm follows an ensemble research technique for classification and regression. A random forest takes on an average of the different sub-samples of the data set, which increases the predictive exactness and decreases the over-size of the data set.

II. PROBLEM DEFINITION

The prediction of stock markets is essentially described as an attempt to assess stock value and to provide people with a robust idea of knowing and predicting market prices and stocks. The quarterly financial ratio is usually presented with the data collection. Depending on a single dataset may therefore not be enough for the prediction and may yield an incorrect result. Therefore, we are looking into the study of machine learning with different data sets to forecast the demand and stock patterns.
When intelligent investors realize that numerical time series analysis results closer in predicting stock market behavior, machine learning techniques are employed. This helps financial analysts to forecast and thus behave accordingly the actions of the stock they are worried about.
Historical data from Tata Finance will be included in our framework. Suitable data can be used to assess the movements in stock prices. The prediction model would then inform the next trading day of up or down the stock price flows, and investors will use the prediction model in order to increase profit chances.

III. LITERATURE SURVEY

We collected some information on the stock market

forecast during a literature survey.

1. Machine Learning Survey of stock market forecast

The prediction of the stock market has now become an ever-increasing challenge. Technical analysis is one of the techniques used, but the findings are not always reliable. It is important to establish ways of predicting more precisely. Investments usually take into account all variables that may influence the inventory price, using forecasts derived from the stock prices.

In this case, the methodology used was regression. Since financial inventories at any time produce large quantities of data, a great amount of information needs to be evaluated before a forecast can be made. of the regression strategies has its own benefits and drawbacks compared to its other counterparts. Linear regression was one of the remarkable techniques described. The way linear regression patterns work is that they are often fitted with the least squares method, but they can also be fitted in other ways , for example by reducing the fitness gap in another rule or by decreasing a disabled version of the least square loss function. On the other hand, it is possible to use the least square method in non-linear models.

2. A stock market forecast survey using SVM

Recent studies have shown well that a sample Predictability Test does not work with the most predictive regression models. The explanation was the instability parameter and the uncertainty of the model. The studies have also concluded the conventional strategies promising to resolve this problem. The kernel, decision function, and sparsity of the solution help the vector machine commonly known as SVM. It is used to learn the radial polynomial function and classifier of the multi-layer perceptron. It is a classification and regression training algorithm that operates on a bigger dataset. Many algorithms are present in the market but SVM offers improved efficiency and precision. The correlation analysis between SVM and the bond market shows that equity prices and the market index are interconnected strongly.

3. Pricing of stock management The use of vector support machinery

Financial organizations and traders have made numerous exclusive models to try to beat their clients on the market, but often everyone has become more profitable consistently than normal. Nevertheless, the challenge of stock forecasts is so critical given that only a few prices will support this company by a great deal of dollars.

IV. EXISTING SYSTEM

- The system predict by choosing the appropriate time to obtain highly predictive values.
- If the operating environment is modified, the original system would not work well.
- It don't rely on external environmental incidents, such as news and social media incidents.
- The current framework needs to be interpreted as data, and thus needs to be scaled.
- To avoid confusion and incompleteness of data it does not take advantage of pre-processing techniques.

V. PROPOSED SYSTEM

We concentrate on predicting stock prices using machine learning algorithms as support Vector Machines in this proposed framework. We suggested the "Machine Learning Study of the stock market." The SVM algorithm was used to predict the stock market price. In this method we have been able to train the computer in order to predict the future from the different data points from the past. In order to train the model, we took data from the preceding year. Two machine-learning libraries were primarily used to solve this problem. The first was numpy, which was used for cleaning and manipulating the data and for analytical planning. The other package was used to evaluate actual items and for predicting them. The information we used were collected online from stock markets of previous years from the public database, 80% were used to train the machine and the remainder 20% to test the data. The basic method of the supervised research model is that it learns and reproduces the patterns and relationships in the data from the training set. For data processing, we used the python pandas library to merge various datasets into a data structure. The tuned data framework allowed us to produce the data for extraction of features. The date and the closing price for the same day was the data frame features. All these features were added to train the computer on the least support vector machine and the object variable was estimated, which is the cost of that day. The precision has also been quantified by using the test set predictions and the true values.

VI. METHODOLOGIES

1. Support Vector machine(SVM)

An N-dimensional space which distinguishes the points is the main task of the support machine algorithm. A number of features are defined by N here. There can be several possible hyperplanes

between two groups of data points. The goal is to find a plane with the maximum margin. The algorithm. The optimizing margin refers to the distance between the two types of data points. The advantage of optimizing the margin is that it offers some refinement in order to promote classification of potential data points. Hyperplanes are called decision limits which enable the classification of data points. The data points are assigned to various groups based on their location relative to the hyperplane.

2. *Least Square Supporting Vector Machines*

The least square supporting vector machines (LS-SVM) are least square versions of supporting vector machines (SVMs) that provide a range of similar supervised learning methods for data analysis and pattern recognition and are used for the analysis of classification and regression. Suykens and Vandewalle suggested least squares of SVM classificatory.

The LSSVM procedure incorporates the concept of a quadrat meter of errors for the objective purpose of the SVM norm, which takes into account training data with n samples.

Train = {(xi, yi) = 1, 2, à, n}, where xi is the input data of Rd and yi is the input data of R. To develop and build the following optimum linear decision function, use nonlinear mapping of the sample input space (x) = {< (x1), (xn)}.

$$y(x) = \omega^T \cdot \varphi(x) + \beta$$

By using the structural risk formula

$$R = \frac{1}{2}\|\omega\|^2 + C \cdot R_{emp}$$

One can calculate the weight vector ω and the offset β in Eq (1). In Eq (2), C and R_{emp} denote the penalty factor and the loss function, respectively. In LSSVM, the loss function is always equal to the quadratic loss function $R_{emp} - \sum \xi_i^2$, where ξ_i denotes the degree that the mis-classification sample deviates from the ideal sample. Then, getting the solutions of ω and β is equivalent to solving the following optimization problem

$$\min R = \frac{1}{2} \cdot \|\omega\|^2 + C \cdot \sum_{i=1}^{n} \xi_i^2$$

$$s.t.\ y_i = \omega^T \cdot \varphi(x_i) + \beta + \xi_i,\quad i = 1, \cdots, n$$

Introduce the Lagrangian multiplier $\alpha = \{\alpha_1, \alpha_2, \cdots, \alpha_i\}$ to formula (3) and construct an equation as

$$L(\omega, \beta, \xi_i, \alpha) = \frac{1}{2} \cdot \|\omega\|^2 + C \cdot \sum_{i=1}^{n} \xi_i^2 - \sum_{i=1}^{n} \left(\alpha_i \cdot \left(\omega^T \cdot \varphi(x_i) + \beta + \xi_i - y_i\right)\right)$$

Calculate the derivative of every factor in the formula (4). For the purpose of eliminating ω and ξ, introduce the equations $\omega = \sum_{i=1}^{n} \alpha_i \varphi(x_i)$ and $2C \cdot \xi_i = \alpha_i$ to (4). Then one can obtain a new formula described as

$$y_i = \sum_{j=1}^{n} \left(\alpha_j \cdot < \varphi(x_j), \varphi(x_i) >\right) + \beta + \frac{\alpha_i}{2C}$$

Suppose that there exists a kernel function $K(x_i, x_j) = <\varphi(x_j), \varphi(x_i)>$ which satisfies the Mecer condition. Then the formula (5) can be described as

$$y_i = \sum_{j=1}^{n} \left(\alpha_j \cdot K(x_i, x_j)\right) + \beta + \frac{\alpha_i}{2C}$$

Solve the formula (6) to get the model parameters β and $[\alpha_1, \cdots, \alpha_n]^T$. Then the decision function of LSSVM is shown as

$$y(x) = \operatorname{sgn}\left[\sum_{i=1}^{n} \left(\alpha_j \cdot K(x, x_i)\right) + \beta\right]$$

The various types of kernel function K(xi, xj) have already become well known to lead to different LSSVM 's output and the polynomial kernel, sigmoid kernel and RBF are now used commonly. The experiment shows that the RBF is more general than the above functions. The RBF is commonly referred to as

$$K(x, x_i) = \exp\left(-\frac{|x - x_i|^2}{2\sigma^2}\right)$$

Therefore we chose the formula (8) as the kernel for building the LSSVM classification of the fault of the power supply. However, it is necessary to optimize the factor in kernel function and the penalty factor. This paper proposes to improve machine efficiently.

VII. SYSTEM ARCHITECTURE

The first step is that the raw data will be transformed to processed data. This is achieved with the extraction of features as several attributes are obtained in raw data, but only some of them are useful to predict. The first step is to delete the key attributes of the raw data set from the entire list of attributes. The extraction role begins with the initial calculated data and produces derivative values or functions. These features are designed to make the

following learning and generalization simpler and more detailed. Functional extraction is a method of reduction in dimensionality, in which the initial set of raw variables are reduced to appropriate features for ease of management and still present the first information collection accurately and entirely.

The extraction method follows a classification procedure in which the data obtained after the extraction of features are divided into two separate parts. Classification is the question of identifying which groups are included in a new discovery. The training data set is used to train the model, while the test data are used to assess the model's precision. The division is done in a way which preserves a higher proportion of training data than the test data.

A set of random decisions trees is used to evaluate the data. The random forest algorithm a community of decision-makers looks for unique attributes of the results, based on the total number of decision-making trees in the forest. This is called division of data. Since our proposed system's ultimate aim here is to forecast inventory prices by analysis of the inventories.

Fig 1: System Architecture

VIII. EXPERIMENTAL RESULTS

The CSV file contains the raw data from which our results will be released. The rise and decrease of stock prices was explained of eleven columns or eleven attributes. Certain of these characteristics are

(1) HIGH, which imply the highest stock value in the past year.

(2) OPENP is the value of the stock at the beginning of the trading day, and

(3) OPENP is the stock price at which the stock is priced before the trading day ends;

(4) OPENP is the stock value at the very beginning of the trading period. Other attributes are YCP, LTS, TRADE, VOLUME and VALUE, but in our findings the above-mentioned four play a vital part.

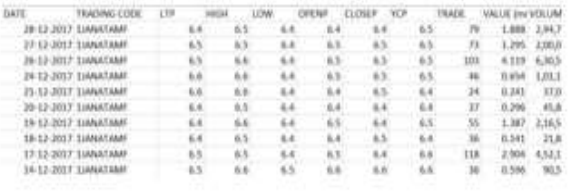

Fig 2 Raw Data

The data in our CSV file are depicted in pictures. This file contains 121608 records of this nature. The dataset contains more than 10 trade codes and some of the documents do not have relevant data that can help us train the system, so it is reasonable to treat the raw data. This gives us more advanced data set that can now be used to train the computer.

	DATE	TRADING CODE	LTP	HIGH	LOW	OPENP	CLOSEP	YCP	TRADE	VALUE (mn)	VOLUME
0	2018-06-19	1JANATAMF	6.2	6.3	6.1	6.2	6.2	6.2	58	0.757	122741
1	2018-06-19	1STPRIMF1MF	11.2	11.2	10.9	11.0	11.1	10.9	145	2.640	238810
2	2018-06-19	AAMRANET	80.1	80.4	78.5	78.5	79.7	78.3	545	15.488	195035
3	2018-06-19	AAMRATECH	30.9	31.0	30.7	31.0	30.9	31.0	186	5.100	164899
4	2018-06-19	ABB1STMF	8.1	8.1	5.9	6.0	8.1	8.0	109	11.214	1857588

Fig 3 head ()

This is the product of the head application. Since we use the library of pandas to analyses data, the first five rows are returned. The default value for the number of rows returned here is five, unless otherwise defined. Trading code is not important in the processed data collection, so we delete strip(s) and substitute all trading codes with "GP" values.

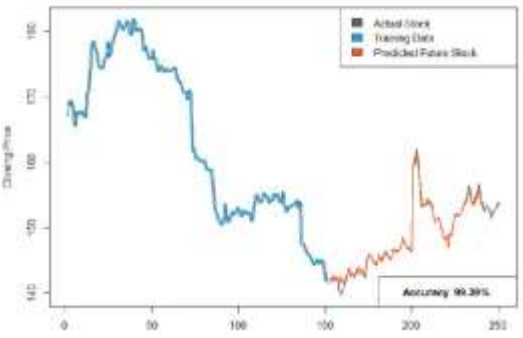

Fig 4 Time series plot

This is a time series plot developed using the library "matplotlib.pyplot." The drawing is of 'CLOSEP' and 'DATE' attributes. This demonstrates the latest trend to close stock prices as time varies over a two-year period. The following figure is the candle stick plot created by the "mpl finance" library. The candle stick plot for 'DATE,' 'OPENP,' 'Large,' 'LOW,' 'CLOSEP' has been produced.

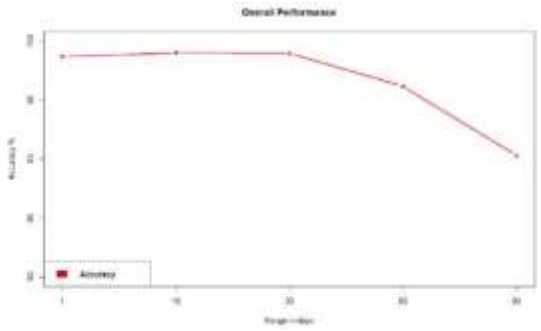

Fig 5: Overall Performance

The following step included the configuration of the feature, goal and train size. We import and match with the training data using the sklearn libraries. The uncertainty matrix obtained is shown below after the model is trained with the data and test data are performed via the trained model.

	precision	recall	f1-score	suppor
-1.0	0.76	0.93	0.84	2
1.0	0.85	0.58	0.69	1
micro avg	0.79	0.79	0.79	4
macro avg	0.81	0.75	0.76	4
weighted avg	0.80	0.79	0.78	4

Confusion Matrix

A confusion matrix indicates that the classification model has the number of correct and incorrect predictions in relation to the actual results (target value) of the data. NxN is the matrix, where N is the number of destinations (classes).

These models are typically assessed with the data in the matrix. A 2x2 confusion matrix for two groups (Positive and Negative) appears in the following table.

		Actual	
		Positive	Negative
Prediction	Positive	True Positive	False Positive
	Negative	False Negative	True Negative
	Accuracy (ACC) = (Σ True positive + Σ True negative) / Σ Total population	True positive rate (TPR) = Σ True positive / Σ Condition positive	False positive rate (FPR) = Σ False positive / Σ Condition negative

Confusion Matrix

IX. CONCLUSION

Support vector machine for predicting the course of financial movement. We saw that the Support Vector Machine provided us with better results of both these algorithms. SVM is a promising financial prediction method. In the forecast of regular motion route, SVM is superior to other individual classification methods. For financial predictors and traders, this is a simple message which can lead to a capital gain. However, the strengths and disadvantages of each approach are different. The key components defined by the SVM and internal and external financial factors are used for prediction in this model. The choice of the indicator function can also improve / reduce the predictive system's accuracy considerably. Even a specific machine learning algorithm may be more suitable for a particular type of stock, say Technical inventory, while the same algorithm could offer less accuracy when predicting other types of inventory, say Energy stocks.

REFERENCES

[1]A.Beattie,"7 Controversial Investing Theories," [Online]. Available: http://www.investopedia.com/articles/financial--theory/controversial--financial--theories.asp.

[2] L. QianYu and F. ShaoRong, "Stock market forecasting research based on Neural Network and Pattern Matching," 2010. [Online]. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5592759.

[3]Wikipedia,"List of S&P 500 Companies," [Online]. Available: https://en.wikipedia.org/wiki/List_of_S%26P_500_companies.

[4] Wikipedia, "R," [Online]. Available: https://en.wikipedia.org/wiki/R_(programming_language).

[5] Wei Huang, Yoshiteru Nakamori, Shou-Yang Wang, "Forecasting stock market movement direction with support vector machine", Computers & Operations Research, Volume 32, Issue 10, October 2005, Pages 2513–2522.