Stock Market Prediction Using Machine Learning Approach: A Review

Nargis I khan¹, Prof. Nitin Janwe²

¹Student, Department of Computer Science and Engineering Rajiv Gandhi College of Engineering Research & Technology, Chandrapur, India, 442401
²Associate Professor, Department of Computer Science and Engineering Rajiv Gandhi College of Engineering Research & Technology, Chandrapur, India, 442401

Abstract – Now a day’s, the prediction of stock market prices and conditions has become a major researched topic amongst the data scientists, investment bankers, and stock brokers. As, the behavior of stock market is very non-linear and volatile in nature, it makes a very high-risk investment. Consequently, a lot of researchers have contributed their efforts to forecast the stock market pricing and average movement. Researchers have used various methods in computer science and economics in their quests to gain a piece of this volatile information and make great fortune out of the stock market investment. Data mining and machine learning approaches can incorporate into Business Intelligence (BI) systems to help users for decision support in many real-life applications. This paper presents the brief survey of application of machine learning in stock market prediction and investigates various techniques for the stock market prediction using Artificial Neural Network (ANN) and Support Vector Machine (SVM). ANN is non-linear and non-parametric classifier which is viable for forecasting of stock prices. SVM focuses on marginal values rather than average values for the classification prediction model. The aim of this paper is to provide a review of the application of machine learning in stock market prediction to determine what can be done in the future.

KEYWORDS – Stock market; machine learning; ANN; SVM.

1. INTRODUCTION

Predicting the future has always been an adventurous and attractive task for the probing individuals. This kind of prediction becomes more fascinating when it involves money and risk like predicting stock market. Goal behind making any financial investment is to achieve above average return for invested money while maintaining certain level of involved risks but as the stock market is a very complex, volatile and non-linear dynamical system, stock market prediction has become a tough challenge for researchers and investors. Lot of Research has been done on stock market prediction by researchers of different fields including the business and computer science. Researchers have tried different approaches for market prediction including different techniques and algorithm and different combination of attributes. The attributes that makes a prediction model depends upon the factor on which market performance can depend. Different methods have been developed to forecast the behavior of stock market prices based on historical data. As per recent survey Stock prices prediction with conventional statistical methods has proven to be less effective because of non-linear characteristics of financial time series. On the contrary, machine learning-based methods, such as Artificial Neural Networks (ANNs), support vector machine (SVM) and data mining system offer useful tools for forecasting noise environments like stock market.
The reminder of the paper is organized as follows, in section 1 we discussed about the stock market. In section 2 we discussed about Machine learning algorithms, ANN and SVM. In section 3 we briefly examine the previous work. In section 4 we described the motivation and objectives. Finally, in section 5 we summarize our conclusion.

1.1 STOCK MARKET

Stock market is an organized and regulated financial market where securities such as bonds, notes and shares are bought and sold at price governed by the forces of demand and supply. Stock market serves as, market where corporations, governments, municipalities and other incorporated bodies can raise capital by channeling saving of the investors into productive ventures. Also, market where investor can sell their securities to other investors for cash, thus reducing the risk of investment and maintaining liquidity in the system.

Stock is a share in the ownership of the company that means you own a part of company. Stocks are partial ownership of businesses, which can be traded in the stock market. If company ownership is divides in 100 parts, and we are the investor purchase one part which is equal to 1 share then we own 1% of that company. Stock exchanges uses a trading system which is order driven automated matching system. Stocks prices are defined on the basis of at any time how many buyers and sellers. For the same stock in the market. If numbers of buyers are more than sellers then stock prices become high and if number of sellers higher than buyers, then stock prices become low. If order does not find the match, then it remains in the system and waiting for the fresh orders or updating of previous orders which are already present. Previously stock market was known as stock exchange denotes a place where actual buying and selling takes place. With computerized trading and electronic communication networks like NASDAQ and BATs, manual involvement has been reduced. Automated trading platforms uses computer algorithms which gives high frequency trading thus forecasting model requires a robust technique which gives accurate prediction of prices so that profit can be maximized.

2. MACHINE LEARNING ALGORITHMS

This section contains Introduction to machine learning and work to evaluation of the machine learning techniques. Some of the algorithms we have considered are as follows:

Machine learning is a concept that provides systems the ability to learn automatically and improve from experience without being explicitly programmed. Machine learning concepts focuses on the development of computer programs that can access data and use it learn for themselves. The learning process begins with observing the data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the sample data that we provide. The primary goal is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly. Machine learning algorithms are often categorized as unsupervised and supervised.

A) Unsupervised learning algorithm

In unsupervised learning, the training of machine is done using information which is neither classified nor labeled and allowing the algorithm to work on that information without guidance. The ultimate goal of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data. Different from supervised learning, no teacher is provided that means no training will be given to the machine. That’s why machine is restricted to find the hidden structure in unlabeled data by our-self. Unsupervised learning classified into two categories of algorithms:

1. Clustering: A clustering problem is where we want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.

2. Association: An association learning problem is where we want to discover rules that describe huge portions of your data, such as people that buy X also tend to buy Y.

B) Supervised learning algorithm
In Supervised learning, as the name indicates there is a supervisor as teacher. Learning in which we train the machine using data which is well labeled that means some data is already tagged with correct answer is known as Supervised learning. Next, machine is provided with new set of examples (data) so that supervised learning algorithm analyses the training data (set of training examples) and produces a correct outcome from labeled data. Supervised learning algorithms increases consistently with the data. It is a type of inductive learning. Supervised learning classified into two categories of algorithms:

1. **Classification**: A classification problem is where the output variable is a category, such as “Red” or “blue” or “disease” and “no disease”.

2. **Regression**: A regression problem is where the output variable is a real value, such as “dollars” or “weight”.

2.1. **Artificial neural networks (ANN)**

An artificial neural network (ANN) is a computing model designed upon the functions and architecture of human cerebellum. ANN is also known as connectionist. ANN is a classifier and also a non-linear statistical data modeling tool which models the entangled relationship between the inputs and outputs or it classifies data into patterns. ANN is widely accepted owing its ability to attain and generalize adaptability, parallelism of data processing, robustness, and fault tolerance. ANN proved it to be a universal function approximate and is suitable for modeling tool in various fields like data validation, sales forecasting, medicine, customer research, price forecasting. ANN provides a prediction model which acts as desirable tools that not only show a direction of stock price movement, but also indicates the most supposable price value of the stock. ANN use data samples instead of complete data sets to arrive at solutions, which saves both time and money. ANNs are considered fairly simple mathematical models to improve effectiveness of the available the data analysis technologies. ANNs includes three layers. These layers are connected to each other. The first layer consists of input neurons. Where, neurons send data on to the second layer, which then sends the output neurons to the third layer. Training of artificial neural network includes selecting from allowed models for which there are several associated algorithms.

![Figure 1: Artificial Neural Network](image)

2.2. **Support vector machine (SVM)**

Vapnik and his co-workers in 1963 developed Support vector machine (SVM). SVM is a supervised machine learning algorithm, generally used for regression and classification analysis. It is also known as probabilistic binary linear classifier. It can also perform non-linear classification with the usage of different types of kernels. SVM can be used to solve numerous real-world problems such as text and hypertext categorization, classification of images, permutation test, weather forecasting, sales forecasting etc. SVM is widely accepted due to its remarkable advantages such as effective in high dimensional spaces, memory efficient, versatile as uses different kernel functions for different decision functions. Kernel selection is an important task in support vector machines.

SVMs emerges from statistical learning theory; the aim being to solve only the problem of interest without solving a more difficult problem as an intermediate step. The foundation of SVMs is on the structural risk minimization principle, closely related to regularization theory. This principle incorporates capacity control to prevent over-fitting and thus is a partial solution to the bias-variance...
trade-off dilemma. Two important elements in the implementation of SVM are the techniques of mathematical programming and kernel functions. The parameters are found by solving a quadratic programming problem with linear equality and inequality constraints; rather than by solving a non-convex, unconstrained optimization problem. SVM algorithm developed by Vapnik are based on statistical learning theory. SVM can be used for both classification and regression task. In classification case we try to find an optimal hyper plane that separates two classes. In order to find an optimal hyper plane, we need to minimize the norm of the vector w, which defines the separating hyper plane. This is same as maximizing the margin between two classes.

In SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Next, we perform classification by finding the hyper-plane that differentiates the two classes very well. Below figure shows the 2-dimensional case where the data points are linearly separable.

![Support Vector Machine](image)

**Figure 2: Support Vector Machine**

### 3. LITERATURE SURVEY

In this section we will review basically two data mining techniques i.e., ANN and SVM for forecasting stock prices. In recent survey it has been observed that including indicators as inputs along with dataset in the system gives better prediction results. This section will present some of the works that has been carried out by researchers in adopting and applying ANN and SVM in financial stock market prediction.

#### 3.1. ARTIFICIAL NEURAL NETWORK (ANN)

It has been shown that NN is able to predict the volatility and non-linear stock market prices due to its learning, mapping, generalizing and self-organizing characteristics. In stock market forecasting the most common NN architecture used is Feedforward NN. This is the simplest NN, in which the information of feedforward NN only moves in one direction. The information moves from the input layer to the hidden layer(s) and from the hidden layer(s), it moves to the output layer. In this literature, different sets of input variables are used to predict stock returns. Some researchers have used different input variables to predict the same set of stock return data. Some used input data from a single time series where others considered the inclusion of heterogeneous market information and macro-economic variables. On the other hand some even pre-processed these input data sets before feeding it to the ANN for forecasting.

In paper [1] Author has applied a multilayer feedforward NN with backpropagation technique to predict the direction of Istanbul stock exchange (ISE) market index value. In this study, 6 ANN models were applied to the system model with different numbers of hidden layers (1, 2 and 4), using an ANN software package. These models were trained and tested on historical stock datasets gathered for the period of July 1, 2001 to February 28, 2003 for training and testing. ANN models i.e, Multi-layer perceptron (MLP) and generalized feed forward network (GFF) performances can be measured by the coefficient of determination or the mean relative percentage error. The ANN performance has been compared with Moving Averages (MA) approach for a period of (5 and 10 days in this study). Study shows that, the Total of 8 sets of predictions, that result from the application of 6 ANN models and two MA were performed and compared using the coefficients of determination for ANN models and using mean relative percentage errors for all of the models. Based on the findings of this study, the prediction models based on ANNs were more accurate than the ones based on MAs. From the ANN models, GFF network model was found to be more appropriate for the prediction. In [2]
The best accuracy with multiple dicting nearest stock prices. With the help of results, the UK FTSE 100 Index, similar to predict the stock market prediction. They provided valuable insight in implementing NN in results, the rest were used for testing. 5000 days data, 1000 days were used for training and the backpropagation three layers feedforward NN with 10 input neurons, a hidden layer with five neurons and one output neuron with tan sigmoid and linear transfer functions in the hidden and output layers. The best accuracy of 89.65% and an average accuracy of 69.72% in their predictions have been achieved.

Authors have also organized and carry out experiment to compare ANN with linear model in predicting the stock market price. In [8] Author Carried out a comparative study on the volatility of the BSE Sensex 30 and NSE Nifty 50 using statistical models such as GARCH, EGARCH, GJR GARCH, IGARCH with MLP models.

Similarly, Feedforward NN was also used in both [9] and [10] to forecast the stock market index. Paper [9] predicted trading signals of the AORD one day ahead based on the current day’s close price of the US S&P 500 Index, French CAC 40 Index, the UK FTSE 100 Index, German DAX Index and AORD. NN models were examined in the forecasting study which is the feedforward and probabilistic NNs. As per results feedforward NN outperformed probabilistic NN both in the classification/misclassification rate and trading simulations. Paper [10] forecasted the S&P 500 index using MLP and probabilistic NN. The data set were composed of daily closing S&P 500 index and also the currency exchange rate of Yen, Pound and Mark from Feb 1994 to Sept 1995. The results showed that probabilistic NN performed slightly better than MLP. Another similar work is in [11] which predicted both the trend of stock price and value of stock price using feedforward NN and radial basis NN with backpropagation. Bhavcopy (dataset repository for NSE) copied data from the National Stock Exchange (NSE) for November 2005 has been used. As per Experimental results the feedforward NN using backpropagation is better for trend prediction where it achieved almost 100% accuracy compared to 80% accuracy attained by the radial basis NN. Even so, the radial basis NN performed better in stock price prediction by achieving higher percentage of accuracy than the
feedforward NN. Similarly [12] predicted the daily returns of the BSE Sensex using a MLP network with backpropagation as the learning algorithm. The historical data set used was the daily index values of the BSE Sensex from 16th Jan 1980 to 26th December 1997. The predicted results for daily returns of BSE Sensex are found to be satisfactory and also showed that the previous day value heavily influenced the predictive model. [13] Combines a multilayer feedforward NN with backpropagation to predict the direction of Istanbul stock exchange national 100 indices. 1905 observation data were used in this study. Technical Parameters used for the prediction involves highest, lowest, closing price, exchange rate with USD and the response rates. Results show that the accuracy of 74.51% in predicting the stock market direction is obtained. [14] Compared the backpropagation NN and genetic based backpropagation NN to predict the stock price of the day. It has been observed that Genetic Algorithm based backpropagation NN predicted more accurate price than the backpropagation NN. For training the data set used is the Maruti stock from Jan 2004 to Dec 2006 and 2nd Jan 2007 to 30th Mar 2007 for testing. They stated that the accuracy achieved by Genetic Algorithm’s based backpropagation NN system was 98.31% while conventional backpropagation NN only achieved an accuracy of 93.22%. [15] Uses genetic algorithm with artificial neural network measuring the qualitative effect on the stock market using fuzzy inference rules. [16] Applied feature discretization using GA and ANN for prediction to getting better performance than GALT-ANN model.

[17] Predicted Taiwan stock exchange by Using EDA based local linear wavelet neural network. In [18][19] ANN is compared with logistic regression Technique and shows that ANN outperforms. In [20] The Bayesian ANN model is compared with ARIMA model and fusion model with weighted average and also the result shows that regularized Bayesian network gives an average of 98% fit for the future stock prices. [21] Shows the Comparison between single stage approaches consisting of ANN, SVR, RF and also the fusion of these approaches i.e., SVR-ANN, SVR-SVR, SVR-RF and according to different performance measures SVR-ANN model performed best in prediction. A wide variety of applications and studies have been tested and applied with NNs, including diagnosis of diseases, speech recognition, data mining, image processing, forecasting, robot control, and many others.

3.2. Support Vector Machine (SVM)

Support vector machine can be applied to forecast the stock exchange by using some company specific parameters such as price per earnings ratio of stock, net income, net revenue diluted earnings per share etc. SVM is acutely explicit kind of learning algorithm indicated as the capacity control of decision function and usage of kernel function. The maximal-margin classifier shows how SVM works. SVM can be used to learn polynomial, multi-layer perceptron (MLP) and radial basis function (RBF) classifiers. SVM works upon the principle of structural risk minimization that leads to the prevention of over-fitting problem. SVM can be used to solve numerous real world problems such as text and hyper text categorization, classification of images, permutation test, weather forecasting, sales forecasting etc. SVM is widely accepted due to its remarkable advantages such as effective in high dimensional spaces, memory efficient, versatile as uses different kernel functions for different decision functions.

In [22] Authors has used Support vector Machine for predicting stock market. They present a theoretical and empirical framework to apply the Support Vector Machines strategy to predict the stock market. Four Parameters are company-specific and six macroeconomic factors that may influence the stock trend has selected for further stock multivariate analysis. They applied Support Vector Machine for analyzing the relationship of these factors and predicting the stock performance. According to experimental results SVM is a powerful predictive tool for stock predictions in the financial market. The base paper which is used by authors for demonstrating the result is [23] Where investigation of predictability of financial movement direction with SVM by forecasting the weekly movement with those of linear discriminate analysis, quadratic discriminate analysis and Elam Back propagation Neural Network has been done. As per experimental result SVM out forms the other classification method. In [24] support vector machine technique is combined with variable selection method for stock market prediction. A variable selection is applied to select the
significant variables to feed to the SVM model. 11 variables have been used in the process of variable selection implemented by cross-correlation based method and by peeling method. Accordingly three models have been proposed. Based on experimental results, both hybridized models shows improvement in the prediction accuracy compared to SVM model. The trend prediction accuracy is presented in table 1. The accuracy increases from 82.5% to 84.166% by using two different variable selection methods combine with SVM.

Paper [25] proposed a model in which they integrates particle swarm optimization (PSO) algorithm and LS-SVM for stock market prediction using financial technical indicators. Comparative study has been done on LS-SVM, LS-SVM optimized by PSO and ANN-BP. Levenberg-marquardt(LM) algorithm is used as a benchmark to compare LS-SVM and LS-SVM-PSO. According to results LS-SVM optimised with PSO outperforms with lowest error value as compared to LS-SVM, while ANN-BP is the worst one. Indicators include relative strength index, money flow index, exponential moving average, stochastic oscillator and moving average convergence/divergence.

Paper [26] Studies and compare four machine learning algorithms that are Single layer perceptron, multilayer perceptron, Radial basis function and Support vector machine. 9 parameters(i.e oil rates, gold rates, silver rates, FEX, SMA, ARIMA, KIBOR, NEWS, Twitter) has been used as inputs for prediction model. The results of all four algorithms over training set and testing is presented in table 2. According to results shown SVM performed best on training set while MLP algorithm did well on test data set. As test data is completely new and unseen instances. Therefore, MLP seems to be more efficient in predicting the stock market. MLP outperforms the remaining three algorithms.

In [27] The SVM forecasting model has been compared with back-propagation neural network (BPN) and Case-based reasoning (CBR) and the result shows that SVM is the best model out of these three models.

In [28] 12 technical indicators has taken as input for SVM, ANN, random forest (RF) forecasting model and SVM outperforms after comparison. [29] Support vector machine has been compared with ANN, ANFIS and SVM outperforms. Support vector machines has been applied successfully in many problems such as speech recognition, signal recognition, text categorization, gene selection, intrusion detection, spam filtering, forecasting, medical image classification, classification. Several papers have been published assessing performance of SVM against some statistical and machine learning algorithms in financial applications.

4. MOTIVATION & OBJECTIVE

When and how to invest money is a problem that many individuals face. Although the large amount of money and interest embedded into this problem, there is still no definitive answer on when and where money should be invested. Because to this, most people either keep their money in the bank or hand it off to someone else to manage. This study looks to develop an automated trading system that can be used to make more money off trading securities than the traditional buy-and-hold strategy and to reduce the risk involved in making these investments. This is small idea to build up the project for predicting stock market value so the people get an idea about the tomorrow stock market value and safe to take decision at the time of investment. In recent years, there have been a growing number of studies looking at the direction of movements of various kinds of financial instruments. Both academic researchers and practitioners have made tremendous efforts to predict the future movements of stock market index or its return and devise financial trading strategies to translate the forecasts into profits.

The main objective of this study is to give a brief review on how to predict the stock market current condition with reference to historical stock market values. The given Algorithms will help the share holders to predict the financial condition of a company. In market so many algorithm are present for stock market prediction but, ANN and SVM provides the stock value with efficiently within a less amount of time.
5. CONCLUSIONS

This paper provides a review on machine learning techniques to predict stock prices. In this review, we have surveyed an articles that have used artificial neural network (ANN) and support vector machine (SVM) as a forecasting model for stock prices. This study pinpoints different stock markets, different forecasting methodologies and model comparisons. From the survey, we can conclude that machine learning techniques outperform other conventional models in many cases. And also, data processing gives better prediction results and provides better prediction direction. Every methodology has its advantages and limitations. Although ANN has shown acceptable results, many researchers are still trying to improve the accuracy of the stock market prediction by using a hybrid method, and by focusing more on external factors in order to generate more accurate prediction. SVM in many cases performs the best over ANN; SVM does not give a problem of over fitting as in case of ANN. According to experimental results, SVMs provide a promising alternative tool to ANN for forecasting stock market. It is a known fact that the dynamic stock market world is non-linear, volatile and subject to influences by so many external factors. This paper can be used as an introductory material to those who are interested to work on stock market prediction using ANN and SVM.

ACKNOWLEDGMENT

I am grateful to my guide Prof. Nitin Janwe for his unwavering support and clear guidance. Also, my sincere thanks to all authors, whose papers in the area of stock market prediction are published in various conferences proceeding and journals.
REFERENCES


Stock Market, Wikipedia.