HMM Classification Approach for the Stock Market Prediction in Data Mining

Swati Chedwal[1], Dr.Vishal Shrivastava[2]

1(Department of Computer Science & Engineering Arya College of Engineering & IT, Jaipur(Raj)-India)
2(Department of Computer Science & Engineering Arya College of Engineering & IT, Jaipur(Raj)-India)
Corresponding Author: Swati Chedwal

***Abstract** : The data mining is the approach which can extract the useful information from the rough data. The prediction analysis is the approach which can predict future possibilities based on current information. This research work, is related to stock market prediction using approach of classification. In the previous research work, the SVM classifier is applied for the stock market prediction. In the proposed work, to improve the accuracy of stock market prediction the HMM classifier will be applied with the feature extraction algorithm. The proposed algorithm give high accuracy and execution time for the prediction analysis.*
***Keywords:** Stock market, prediction analysis, SVM, HMM*

I.   Introduction

Data mining can be described as "making better use of data". Every human being is increasingly faced with unmanageable amounts of data; hence, data mining or knowledge discovery apparently affects all of us. It is therefore recognized as one of the key research areas. Ideally, we would like to develop techniques for "making better use of any kind of data for any purpose". However, we argue that this goal is too demanding yet [1]. Over the last three decades, increasingly large amounts of historical data have been stored electronically and this volume is expected to continue to grow considerably in the future. Yet despite this wealth of data, many fund managers have been unable to fully capitalize on their value. Recently, data mining techniques and artificial intelligence techniques like decision trees, rough set approach, and artificial neural networks have been applied to this area. Data mining refers to extracting or mining knowledge from large data stores or sets [2]. Some of its functionalities are the discovery of concept or class descriptions, associations and correlations, classification, prediction, clustering, trend analysis, outlier and deviation analysis, and similarity analysis. Data classification can be done in many different methods; one of those methods is the classification by using Decision Tree. It is a graphical representation of all possible outcomes and the paths by which they may be reached. Following the assumption of technical analysis that patterns exist in price data, it is possible in principle to use data mining techniques to discover these patterns in an automated manner. Once these patterns have been discovered, future prices can be predicted [3]. Today, the grand challenge of using a database is to generate useful rules from raw data in a database for users to make decisions, and these rules may be hidden deeply in the raw data of the database. Traditionally, the method of turning data into knowledge relies on manual analysis; this is becoming impractical in many domains as data volumes grow exponentially. The problem with predicting stock prices is that the volume of data is too large and huge. Public companies raise funds by issuing shares to public and institutions. By purchasing shares, an investor acquires partial ownership of company. A stock market is a place where public listed company stocks are traded. Stocks are exchanged among buyer and seller which generate transaction data. Prices are changing as per demand and supply of stocks [4]. All trading data is captured by stock exchange where stock companies are listed. Stock trading data is nonlinear, fluctuating hence highly time variant. A lot of information is hiding in this huge data captured by stock exchange is difficult and time consuming for human being to extract without powerful tools. Thus predicting future price of stock is highly challenging. Stock prices are changing due to demand and supply of stocks in stock market. Based on past history and current information, investors are buying or selling stocks of listed companies. Investors, stock analyst and stock brokers are predicting demand and supply of stocks after studying fundamentals and technical information of stocks [5]. In order to study and analyze stocks data, investor need to go through vast amount of data to discover hidden patterns which is very cumbersome and tiring task. Usually stock analysts analyze stocks based on three vital aspects - external environmental factors, fundamental analysis and technical analysis. Several stock market prediction theories have been proposed by various researchers. Efficient Market Hypothesis was an idea developed in the 1965. EMH states that the price of a security will reflect the whole market information. As soon as there is any information indicating that a

security is underpriced and therefore offers a profit opportunity, investors will immediately buy it and its price will rise up to a fair value. There were three forms of financial market efficiency introduced: weak, semi-strong and strong [6]. In weak-form efficiency, future prices cannot be predicted by analyzing prices from the past; hence the technical analysis was unreliable. In semi-strong-form efficiency, the price will rapidly adjust to publicly available new information; neither fundamental analysis nor technical analysis techniques are able to reliably produce excess returns. And in strong-form efficiency, the price will reflect all information including insider information; no one can earn excess returns. Behavioral finance explains why and how markets might be inefficient. It is to study the influence of psychological, social, cognitive, and emotional factors on the security and the markets. The historical patterns and the current market circumstance can psychologically affect the investor, and their respond will lead to a predictable market trend. Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute [7]. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. Next the algorithm is given a data set not seen before, called prediction set, which contains the same set of attributes, except for the prediction attribute – not yet known. The algorithm analyses the input and produces a prediction. The prediction accuracy defines how "good" the algorithm is. Decision tree is a classification method which yields output as a flowchart-like tree structure. The result from D-Tree is highly interpretable, but the outcome must be represented in categorical data [8]. Mostly, DTree algorithm called "J48" is applied to classify future stock market direction. Naïve Bayes is a simple probabilistic classifier based on Bayes' theorem, with a naive assumption of independence between every pair of features. K-Nearest Neighbor (k-NN) is a non-parametric lazy learning algorithm that predicts class of the object based on the k closest training examples in the feature space. An object is classified by a majority vote of its neighbors; the object will be assigned to the class most common amongst its k nearest neighbors (k = 15 in this study). Every day, large amount of data is generated within different applications which is difficult to be managed. There are large sized databases created to store this huge data [1]. The users can extract this data as per their requirement. For storing the huge amount of data, large sized repositories as well as databases are created. The major issue that arises within these systems is the appropriate retrieval of important data from such huge databases. The basic data mining technology searches or extracts the useful data from huge databases on the basis of similarity amongst them [2]. The users utilize the extracted information as per their requirements. KDD and data mining, both require the analysis of huge data available. The position of required data needs to be determined, for which various intelligent approaches have been developed. In data mining, Depending upon the sizes and quality of databases available, data mining provides new business opportunities. The different kinds of techniques and mechanisms which help in analyzing the historical data such that predictions about the unknown future events can be collected are included within the predictive analytics study. The patterns recognized within the historical business data are exploited depending upon the business aspects of predictive analytics. With the help of these recognized patterns, different risks and opportunities are defined. For providing an assessment of the risk or potential threat, the relationships amongst different factors are captured. Important decisions making steps for a particular business are guided through this. The predictive modeling and forecasting are used in some conditions to describe the predictive analytics. There are three different models which outline the different techniques of forecasting using predictive analytics. Stock market commonly known as public market for a company collects financial resources by trading their company stock with an acceptable price and for listing their stock [3]. Stock Market prediction has always had a certain appeal for researchers. The fundamental and technical approaches summarize elemental trading philosophies in stock market research. In this approach of stock prediction market summary time series approach is used to predict stock markets on the basis of the historical data using charts as the primary tool which defines the technical analysis. Fundamentals analysis is the study depending upon the factors which affects supply and demand. The main process was stock price prediction which analyzed the gathering and interpretation of the data. The major role played in fundamental analysis is news in which the current supply and demand chain is reflected in the market. Due to the examination of larger information, these conventional approaches are becoming lower in rank nowadays due to the computational power increase in a less time [4]. The important issue in the modern financial market states the prediction of stock trend. There is a rapid change in stock markets all time. One of the interesting areas of research is referred to stock market analysis. There is a lot of involvement of investors in the stock market and are keen to gather information about the future of the market for more successful investments. Component inside automatic trader agents are used with an effective market prediction that can also help investors with the trade advices. Future market direction will be provided sometimes with supportive information from prediction system that helps traders indirectly. The incoming news is analyzed by an automatic text classification technique. To increase the prediction accuracy additionally few approaches parameters in numeric for are interrelated to stock price [5]. According to the variation of the results several indicators such as technical, fundamental, and statistical have been proposed. Since none of the combination of techniques has not yet got a

success. The main goal of the prediction research is to focus mainly on the development of intelligent systems. A public market for the trading company stock can be called a stock market. Securities of such kind are recorded on a stock exchange and on private trading too. The involvement trading between two investors is done in stock market that is why it is also called secondary market. On the basis of demand and supply the prices are fixed by share market [6]. Highly stock sold will decrease in price and high price will increase the demand of the stock. Listed companies are allowed to be traded in the market place. One of the most important sources of the company is the stock market for raising the money. The stock market prediction is permitted to be publicly traded. Another important part of selling shares of the ownership of the company in a public market was to increase the additional capital [7]. The price of the shares and the cost o the other assets are a major part of dynamics of an economic activity as shown in the history. Moreover, for a social mood stock market prediction can be an indicator. Country's economic strength and development is a stock market which is considered to be a primary indicator. Lowering share cost tends to be associated with decreased business investment and in return the increased prices tend to increase invest of the firm. There are several techniques in stock market prediction. The computing model designed upon the functions and architecture of human cerebellum is (ANN) [8]. It is also known as connectionist. ANN is a relationship between the inputs and outputs which is classifier and a non-linear statistical data modeling tool. Support vector machine (SVM) is a regression method. The usage of SVM is done in terms of regression and classification analysis. It is a supervised machine learning technique. The other name of SVM is probabilistic binary linear classifier. Hidden Markov Models (HMM) is consisted with some fixed number of states. A time series of multivariate observations is given a probabilistic framework or the modeling. This model has increasingly become a popular in the last several years based on the statistical methods [9]. It used a wide range of applications on the theoretical basis.HMM is defined as a classifier or predictor for speech signal recognition as proposed by the researchers recently. It is also classified into sequence analysis, natural language domains and handwritten characters recognition. For various applications HMM is considered to be a very powerful tool.

II. Literature Review

Kumar, I.et.al (2018) presented a comparative study of this research paper learns algorithms which have been proposed by using the time window of size 1 to 90 of the supervised machine [10]. Five supervised techniques have been developed and stock market trend predicts their comparisons on the basis of their performances. Stock prediction is a difficult and highly complicated task with the impact of various factors on the stock prices. One way to overcome such difficulties Machine Learning Techniques have been applied for the stock price prediction. Further, the accuracy of each algorithm is reduced in the number of technical indicators. In terms of accuracy Naïve Bayes Algorithm shows the best results after reducing the size of dataset to almost half of the original. The consequence comes out to be for the large dataset.

Tejas Mankar, et.al (2018) presented that real world problems are solved by machine learning and artificial intelligence techniques with the usage in conjunction with data mining [11] Huge amounts of time are consumed and highly effective, yielding maximum accuracy have been proved with minimal monetary investment. The study says a good overview of public mood that classifies as positive, negative and neutral tweets. The system will subsequently study factors affecting the public mood. Profits can be earned by taking informed decisions to a potential stock investor who takes a great help from the prediction .Various companies are forecasted using sentiment analysis and twitter API on the collected tweets. Hence, many are discouraged from the fear of loss.

Ze Zhang, et.al (2017) presented the error of BP network is large and the error of the Elman network is slightly larger than the model proposed [12]. The model cannot be fully demonstrated due to low data set. Stock market tends to be unfailing research subject in the financial world. This paper has used self adapting variant PSO algorithm for the optimization of weighs and thresholds of the network. For further prediction it takes stock market open price Elman is given an initial weight and threshold value for training. PSO-Elman network is formed while predicting the model for open price market of stock market based on self adapting variant. Finally, the consequence is some stock prices are verified precisely and BP network and Elman network is compared. Hence the stability and precision models are both superior to the traditional neural network.

Yaojun Wang.et.al (2016) proposed a novel approach to perform quantitative analysis of market segment by using social media mining technology [13]. In this proposed approach, the stock comments information is collected from the social media initially. Emotion vectors are created by performing data preprocessing. The sensitivity of segment value is found to be high for the stock price moment as per the calculations performed for segment value of each stock's emotion vector. It was seen through the conducted experiments and achieved results that in comparison to the SVM model which was not integrated with segment index, the performance of SVM model integrated with segment index was better in terms of prediction accuracy

Meryem Ouahilal.et.al (2016) presented the optimization of stock price prediction with the combination of Support Vector Machine model and Hodrick-Prescott filter under this hybrid approach of predicting stock

price [14]. With the help of Maroc Telecom (IAM) financial time series various experiments has been conducted to assess the performance of this proposed approach. Much attention has been paid to the predicting stock and also it is considered as an important factor to this approach. During the period between 2004 and 2016 every day data was collected. A useful tool such as machine learning techniques was used to provide good accuracy of financial terms prediction. The proposed model had given the better results in terms of stock price predictions on the basis of their experiments.

Tantisripreecha, et.al (2016) presented the online scheme which provided the promised result for KNN, ANN, and Decision Tree [15]. The historical pattern finished the merits of LDA-Online. In this proposed approach numerous of indicators investigated their capabilities on LDA-online in the near future. The study under this approach reflected the new development feature extraction for the stock price prediction in comparison with ANN and Decision Tree. A short total running time was considered under this study. LDA-Online was found with the best and top-level performance. Linear Discriminate Analysis (LDA) was fit by applying features such as the opening, closing, highest and the lowest prices. And through the experiments, the result was taken out that computational time was not a problem hence, LDA was a method of choice.

Xi Zhang, et.al (2018) proposed a Multi-source Multiple Instance model that predicted the movement of the stock market [16]. In this paper, it was also predicted the importance of the information. The event information was captured effectively that was proposed in this novel as an event representation learning process. The effectiveness of this model was demonstrated from the year 2015 and 2016 on the basis of the evaluations on the data. Identification of the crucial information and determination of the importance of each data source was considered interpretable predictions. Hence, the effectiveness of this model was evaluated on the basis of two-year data

III. Research Methodology

This research work is related to stock market prediction. The stock market prices are changing very frequently and technique needs to propose which can predict the price of the stocks. The stock market prediction techniques work in the four phases which are pre-processing, feature extraction and classification. The steps of the proposed algorithm are explained below:-

1. Input dataset and pre-processing:- In this phase, the dataset of stock market will be taken as input. The input dataset will be pre-processed means the missing and redundant data will be removed from the dataset

2. Feature extraction: - In the second phase, the features of the input dataset will be extracted for the final prediction. In this phase, the relationship between the attribute set and target set will be done. The feature extraction, will be used for the final classification

3. Prediction: - In this phase, the technique of classification will be applied for the stock market prediction. The approach of HMM classification will be applied for the stock market prediction. A stochastic model that is used for sequential data is called Hidden Markov Model (HMM). A latent Markov chain and a set of observation probability distributions are the two interrelated mechanisms included within this stochastic process. Each of the probability distributions is associated with a state and a finite number of states are included within a latent Markov chain. The process is assumed to be within a state at every discrete time instant. The probability distribution related to the current state generates an observation. In case is the output alphabet is finite, HMM is known to be discrete and if the output alphabet is not finite, it is known to be continuous. The Markov chain augmentation is used as a base in HMM. A model which describes something related to the probabilities of sequences of random variables is called the Markov chain. A strong assumption which states that if one wants to predict the future in sequence, current state matters the most is made in a Markov chain. There is no impact on the future due to the states previous to the current state.

IV. Result and Discussion

In the last, the performance of the proposed model will be tested using various parameters. The performance analysis parameters are accuracy, execution time

1. Accuracy: Accuracy is defined as the number of points correctly classified divided by total number of points multiplied by 100, as shown in eqn. 1.

$$\text{Accuracy} = \frac{\text{Number of points correctly classified}}{\text{Total Number of points}} * 100 \text{ ---1}$$

2. Execution Time: Execution time is defined as difference of end time when algorithm stops performing and starts time when algorithm starts performing as shown in eqn. 2.

Execution time = End time of algorithm- start of the algorithm --2

The proposed algorithm is implemented in python using anaconda platform. The performance of proposed algorithm will be compared with the existing algorithm called SVM for the stock market prediction. The authentic dataset will be taken as input which is collected from UCI repository for the prediction analysis. The performance of proposed and existing algorithm will be analyzed in terms of accuracy, and execution time.
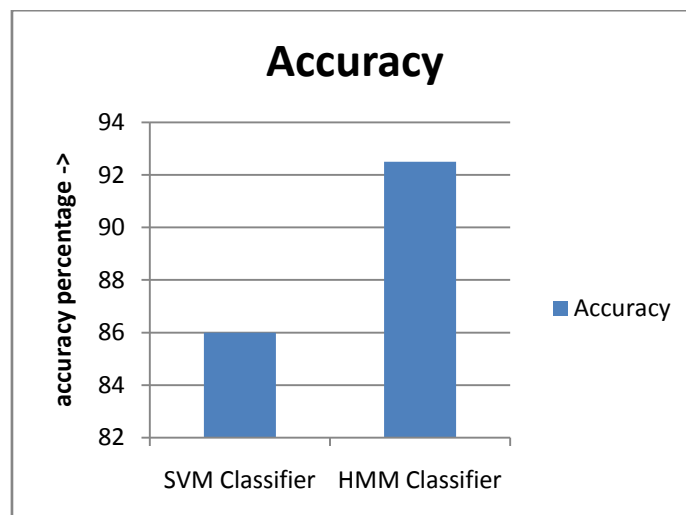
Accuracy

Fig 1: Accuracy Comparison

As shown figure 1, the accuracy of the existing SVM is compared with the HMM classifier. It is analyzed that HMM classifier give high accuracy as compared to SVM classifier for the stock market prediction
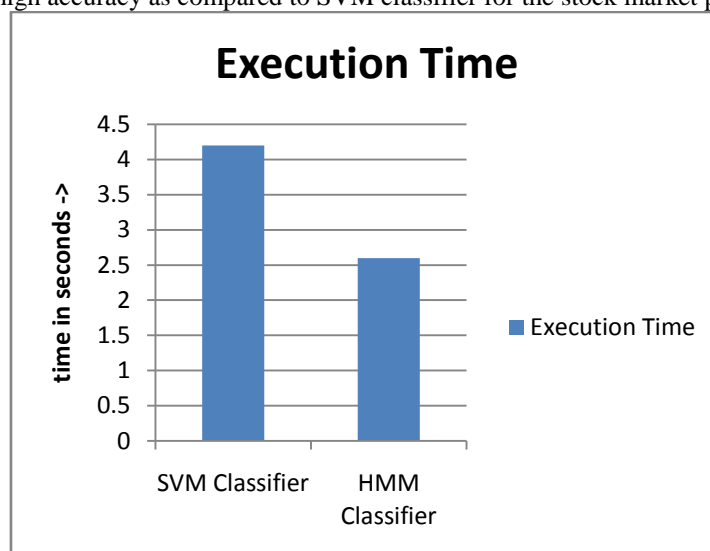
Execution Time

Fig 2: Execution Time

As shown in figure 2, the execution time of the existing SVM classifier is compared with the proposed HMM classifier. The execution time of existing SVM classifier is high as compared to proposed HMM for the stock market prediction.

V.   Conclusion

In this work, approach of stock market prediction is proposed using the stock market prediction. The stock market prediction technique has the three steps which are pre-processing, feature extraction and classification. In the existing technique SVM classifier is applied for the stock market prediction. To increase accuracy for the stock market prediction the SVM classifier will be replaced with HMM classifier. It is analyzed that HMM classifier give high accuracy and low execution time for the stock market prediction as compared to SVM classifier.

References

[1].   R. Suresh, S. R. Harshni, "Data mining and text mining — A survey", 2017 International Conference on Computation of Power, Energy Information and Commuincation (ICCPEIC)

[2].   Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta, "A Comparative Study of Various Clustering Algorithms in Data Mining", International Journal of Engineering Reserch and Applications (IJERA), Vol. 2, Issue 3, pp.1379-1384, 2012

[3].   Pradeep Rai, Shubha Singh, "A Survey of Clustering Techniques", International Journal of Computer Applications, 2010

[4]. Huang Ming, Niu Wenying and Liang Xu, "An improved Decision Tree classification algorithm based on ID3 and the application in score analysis", Software Technol. Inst., Dalian Jiao Tong Univ., Dalian, China, June 2009.

[5]. S. Sneha, "Applications of ANNs in Stock Market Prediction: A Survey," International Journal of Computer Science & Engineering Technology," vol. 2(3), 2011, A review of stock market prediction with Artificial neural network (ANN)

[6]. F. Pegah, "Stock trend prediction using news articles a text mining approach", 2007, Master thesis, Luleå University of Technology

[7]. Mingyue, Q., Cheng, L., & Yu, S. "Application of the Artifical Neural Network in predicting the direction of stock market index", 2016, 10th International Conference on Complex, Intelligent, and Software Intensive Systems.

[8]. M. Mehrara, A. Moeini, M. Ahrari, and A. Ghafari, 2010, "Using Technical Analysis with Neural Network for Prediction Stock Price Index in Tehran Stock Exchange," Middle Eastern Finance and Economics, vol. 6(6), pp. 50-61, 2010.

[9]. Zhang, G., Patuwo, B. E., & Hu, M. Y., 1998, "Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting", 14, 35-62. (97)00044-7

[10]. Kumar, I, Dogra, K Utreja, C & Yadav, P, "A Comparative Study of supervised Machine Learning Algorithms for Stock Market Trend Prediction", 2018, IEEE, Second International Conference on Inventive Communication and Computational Technologies (ICICCT)

[11]. Mankar, T, Hotchandani, T., Madhwani, M., Chidrawar, A, & Lifna, C.(2018), "Stock Market Prediction based on Social Sentiments using Machine Learning" 2018 "International Conference on Smart City and Emerging Technology".

[12]. Ze Zhang, Zhang, Z. Shen, Y, Zhang, G., Song, Y., & Zhu, Y. (2017), " Short-term prediction for opening price of stock market based on self-adapting variant PSO-Elman Neural Network" ,2017, 8th IEEE

[13]. Wang, Y. & Wang Y (2016) "Using social media mining technology to assist in price prediction of stock market. 2016 IEEE International Conference on Big Data Analysis" (ICBDA)

[14]. Ouahilal, M., El Mohajir, M., Chahhou, M, & El Mohajir, B. E. (2016). "Optimizing stock market price prediction using a hybrid approach based on HP filter and Support Vector Machine 2016, 4th IEEE

[15]. Tantisripreecha T & Soonthomphisaj, N (2018) "Stock Market Movement Prediction Using LDA-Online Learning Model". 2018 ACIS "International Conference on Software Engineering, Artificial Intelligence", "Networking and Parallel/Distributed Computing" (SNPD), 19th IEEE

[16]. Zhang, X., Qu, S., Huang, J., Fang, B., & Yu, P. (2018). "Stock Market Prediction -via-Multi-Source Multiple Instance Learning." 2018 IEEE