International Journal for Modern Trends in Science and Technology, 7(03): 176-182, 2021

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0703031
Available online at: http://www.ijmtst.com/vol7issue03.html

Stock Market Prediction: A Systematic Review

Shishir K Sharma

Research Scholar, Department of Computer Science, Kalinga University, Nawa Raipur(C.G.)

To Cite this Article
Shishir K Sharma, "Stock Market Prediction: A Systematic Review", *International Journal for Modern Trends in Science and Technology*, Vol. 07, Issue 03, March 2021, pp.: 176-182.

Article Info
Received on 05-February-2021, Revised on 08-March-2021, Accepted on 12-March-2021, Published on 17-March-2021.

ABSTRACT

Stock market is a promising financial investment which will generate great wealth. However, the volatile nature of the stock exchange makes it a really high risk investment. Thus, tons of researchers have contributed their efforts to forecast the stock exchange pricing and average movement. Researchers have used various methods in applied science and economics in their quests to realize a bit of this volatile information and make great fortune out of the stock exchange investment. Data mining is one of the best suitable technique for stock market prediction. Achieving accurate stock market models can provide investors with tools for making better data-based decisions. These models can help traders to reduce investment risk and select the most profitable stocks. Data mining within the databases is named a way from which the extraction of necessary information are often done from the raw information. With the assistance of the prediction analysis technique provided by the information mining the longer term scenarios regarding to the present information are often predicted. The prediction analysis is that the combination of clustering and classification. In order to supply prediction analysis there are several techniques presented through many researchers. In this review paper, various techniques proposed by various authors are analyzed to know latest trends within the prediction analysis. The prediction analysis techniques have two steps which are feature extraction and classification. The various classification techniques are reviewed in terms of certain parameters and compared in terms of their outcomes.

***KEYWORDS**: Stock Market, Volatile, Classification, Data mining and Prediction.*

I. INTRODUCTION

Data mining is the patterns for analyzing information and the process to extract the interesting knowledge. In data processing, various data processing tools available which are accustomed analyze different types of knowledge. For analyzing the information few applications which is employed by data processing are like making decisions, analysis on market basket, production control, and customer retention, scientific discovers and education systems [1]. Applied to similar cluster and not same sort of data is mentioned clustering during this approach. The clusters are generated by analyzing similar patterns of the input data. While categorizing genes with same functionality and in population gain insight into structures are often inherited in biology for deriving plant and animal taxonomies. In city, similar houses and lands area can be identified by employing clustering in geology. To discover new theories, information clustering can be used to classify all documents available on Web. The unsupervised data clustering classification method creates clusters and objects as these in different clusters are distinct and that are in same cluster are very similar to each other. In data processing, cluster

analysis is taken in to account a standard topic which is applied for the knowledge discovery. The data objects are grouped as a set of disjoint classes which are known as cluster [2].The similarity of objects that belong to one class is high in comparison to the objects that belong to separate classes. The data is extracted from previously existing data sets such that the patterns among them and the future outcomes possible can be determined. Future predictions are not provided through prediction analysis.

The prediction analysis process provides risk assessment forecast and acceptable level of reliable for the applications. This approach thus, helps in predicting the future possibilities. Any kinds of currently available data and historical facts applied to business are analyzed by the predictive models such that the feedbacks of customers related to the products can be understood. This study also helps in recognizing the potential risk and opportunities of this data. Several techniques have been applied by this study for making future business forecasts along with machine learning, statistical modelling and data mining. The information is thus, extracted and then used further for predicting trends and behavioural patterns using predictive analytics. The predictive web analytics are improved by calculating the statistical probabilities of future events online.

In any kind of past, present or future event of interest which is unknown, the predictive analytics is applied. The variables which can be measured and analyzed are used by predictive analytics software applications for predicting the likely behaviour of individuals. For instance, for the potential driving safety variables being used in insurance company, variables like driving record, pricing, age, gender, location, and type of vehicle are considered. High level of expertise is needed in predictive analytics with the statistical methods and ability to build predictive data models. Data engineers help in gathering relevant data and preparing it for analysis. Therefore, with data visualization, dashboards and reports are supported through software developers and business analysts. Clustering methods divided into categories are as follows:

a. ***Partitioning Methods:*** The essential functioning of this method is that the collection of the samples during a path to generate clusters of same objects that are of high similarities. Here, the samples that are dissimilar are grouped under different clusters from similar ones. These methods completely rely on the distance of the samples [3].
b. ***Hierarchical Methods:*** A given dataset of objects are decomposed hierarchically within this technique. There are two types in classification of this method is done with the involvement decomposition. It is divisive and agglomerative methods based upon [4]. Agglomerative technique is that the bottom up technique at which the primary step is that the formation of the separate group. Merging is done when the groups are near to each other.
c. ***Density based Methods:*** In many techniques the distance amongst the objects is taken for the separation of the objects into clusters as a base into clusters. However, these methods can only be helpful while identifying the spherical shaped clusters. It is difficult to obtain arbitrary shaped using the technique of density based clustering.
d. ***Grid based Methods:*** It is known as the generation of grid structure by the quantizing the space of the object to the finite number of cells. This method is independent because it isn't hooked in to the provision of the amount of knowledge objects and also features a high speed.

A. Classification in Data Mining

Within the information mining the prediction of the group membership as an example information are often through with the assistance of the classification technique [5].

Prediction analysis is that the process during which outcome are going to be predicted on the premise of current data. For example, on the basis of current weather information it will be analyzed that day can be either "sunny", "rainy "or "cloudy.

Two steps are followed within this process. They are:

a. ***Model Construction:*** Model construction explains the group of classes of predetermined. Wide numbers of tuples are utilized within the construction of the model referred to as training set. Classification of the principles, decision trees or mathematical formulae/regression is shown during this method.
b. ***Model Usage:*** The second way used in the classification is model usage. In order to classify the test data, the training set is designed of the unknown from the unknown

data for the accuracy analysis [6]. The results of the classification of the model is employed to match in sample test with a label that is known. Test set is not dependent on training set.

B. *SVM CLASSIFIER*

In this study the author proposed SVM classifier for regression, classification and also the overall pattern recognition. Due to its high generalization performance without requiring any prior knowledge to add in it, this classifier is considered to be good in comparison to other classifiers. The performance is even better such as extremely high of the input space dimension. The SVM requires best classification function identification for differentiating of grooming data between the 2 classes. The classification function metric may represent during a geometric manner also [7]. The hyperplane *f(x)* is separated through the linear classification function for the linearly separable dataset. This hyperplane passes through the middle of two classes which can be said to separating them. x_n is classified by testing the sign function of the new data instance function $f(x_n)$; x_n which refers to the positive class if $f(x_n)>0$. This is done after the determination of a replacement function.

Determination of the best function by increasing the margin between the two classes is an important objective of SVM. There are many linear hyper planes because of this fact. Hyper plane is amongst the 2 classes an amount of space or distance present. Margin is closest between the closest data points to a point with a shortest distance on the hyperplane**.** This can further help us in defining the way to extend the margin which can help in selecting only a few hyperplanes for the solution to SVM even when so many hyperplanes are available [8].

For an identification of the target function the aim of the SVM is to produce linear function. Performance of the multivariate analysis can help to increase the SVM. The error models are of quiet help here for the SVRs. Within an epsilon amount the error is defined zero of the differences between real and predicted values. In the off chance, there is a linear growth in the epsilon insensitive error. Through the reduction of Lagrangian, the support vectors are often studied. The insensitivity to the outliers are often of beneficial for the support vector regression. The demerit of SVM is that the computations are not efficient enough. There are many solutions proposed for this. The breakage of one big problem into numerous numbers of smaller problems is one way to solve this issue. There are just some selected variables for the efficient optimization for every problem. Until all the issues are solved eventually, this process keeps working in iterative nature. The problem of learning SVM is to be solved also by recognizing the approximate minimum enclosing a set of instances in the program.

This review paper is based on the prediction analysis which is generally done with the classification techniques. This paper is organized such within the section 1, the introduction of the prediction analysis is given with various classification techniques. In the section 2, the literature survey is written on the prediction analysis.

II. LITERATURE REVIEW

Min Chen, *et al.* presented[9]on the basis of multimodal disease risk prediction (CNN-MDRP) algorithm called a novel convolution neural network. The data was gathered from a hospital including within it, both structured also as unstructured data. In order to make predictions related to the chronic disease that had been spread in several regions, various machine learning algorithms were streamlined here. 94.8% of prediction accuracy was achieved here together with the upper convergence speed as compared to other similar enhanced algorithms.

Akhilesh Kumar Yadav, *et al.* presented an analysis of various analytic tools that are wont to extract information from large datasets like in medical field where an enormous amount of knowledge is accessible [10]. The proposed algorithm has been tested by performing different experiments on it that gives excellent result on real data sets. In comparison with existing simple k-means clustering algorithm using the algorithm results are achieved in real world problem.

Sanjay Chakrabotry, *et al.* (2014) presented clustering tool analysis for the forecasting analysis [11].The meteorology has been performed using proposed incremental K-mean clustering generic methodology. The weather events forecasting and prediction becomes easy using modelled computations. Towards the top section, the authors have performed different experiments to see the proposed approach's correctness.

Chew Li S., *et al.* (2013) presented [12] that the results of a particular university's students have

been recorded to keep a track using Student Performance Analysis System (SPAS). The design and analysis has been performed to predict student's performance using proposed project on their results data. The data mining technique generated rules that are employed by proposed system provide enhanced results in predicting student's performance. The student's grades are wont to classify existing students using classification by data processing technique.

Qasem A., *et al.* (2013) suggested that the information analysis prediction [13] is taken into account as important subject for forecasting stock return. The future data analysis are often predicted through past investigation. The past historical knowledge of experiments has been employed by stock exchange investors to predict better timing to shop for or sell stocks. There are different available data processing techniques amongst which, a decision tree classifier has been employed by authors during this work.

K. Rajalakshmi, *et al.* (2015) presented study related to [14] medical fast growing field authors. In this field every single day, a large amount of data has been generated and to handle this much of large amount of data is not an easy task. By the medical line prediction based systems, optimum results are produced using medical data processing. The K-means algorithm has been wont to analyze different existing diseases. The cost effectiveness and human effects have been reduced using proposed prediction system based data mining.

Bala Sundar V., *et al.* (2012) examined [15] real and artificial datasets that have been used to predict diagnosis of heart diseases with the help of a K-mean clustering technique in order to check its accuracy. The clusters are partitioned into k number of clusters by clustering which is that the part of cluster analysis and every cluster has its observations with nearest mean. The first step is random initialization of whole data,and then a cluster k is assigned to every cluster. The proposed scheme of integration of clustering has been tested and its results show that the highest robustness, and accuracy rate can be achieved using it.

Daljit Kaur, *et al.* (2013) explained [16] that data that contains similar objects has been divided using clustering. The data that contains similar objects is clustered in same group and therefore the dissimilar objects are placed in numerous clusters. The proposed algorithm has been tested and results show that this algorithm is in a position to scale back efforts of numerical calculation and complexity in conjunction with maintaining an easiness of its implementation. The proposed algorithm is also able to solve dead unit problem.

Ming, J., *et al.* (2018) proposed multi-dimensionality and nonlinearity which are the important characteristics of the technical and economic data. It is possible to research for the technical and economic data, the large data and data processing analysis approaches are used. Simplification of the fluctuation pattern and influencing factors of the mineral products price are done [17]. The prediction model of the geological missing data is established on the premise of techniques of geo statistics and artificial neural network. The proposed model helps in providing an analysis and discussion about the regularity of geological data of group boreholes along with their geological data. As per the performance results achieved it is seen that the strength of proposed model is high along with its prediction accuracy.

Sakhare, A.V., *et al.* (2017) presented a survey about the road accident analysis techniques which play an important role in transportation. The description of road accident data analysis is done using various data mining techniques. This paper also studied the k-mean algorithm in proper manner. SOM is employed to form and analyze the clusters [18]. A self organizing technique uses the neural network beside an unsupervised learning method. The developed technique helps in improving the accuracy. The improvement of road transportation system is important to reduce the deaths or injuries of people. The accident reasons can be predicted and the accuracy of analysis can be improved to a greater extent in comparison to the k- means clustering algorithm by applying the proposed approach.

Chauhan, C., *et al.* (2017) presented a review of varied algorithms and techniques which help in identifying the criminals. After several reviews it had been seen that the efficiency of ID3 algorithm was more advanced [19]. When analyzing the experimental data, highly effective classification rules were generated by this algorithm. Detecting the hidden links of networks of co- offenders was done using hidden link algorithms which helped in showing the possible way forward for crime partner. With the application of Bayes theorem, the accuracy of classification techniques was improved to 90%. The information and victim system within which the attacks occurred were analyzed using forensic kit which also helped in generating the file.

It had been concluded that the violent crimes were solved and therefore the accuracy was limited by applying Criminal investigation analysis (CIA) tool.

Anoopkumar M., *et al.* (2016) presented a comprehensive study of the various researches done previously in Educational Data Mining (EDM). For improving the academic performances of students and then improving the effectiveness of institutions, the educational data is analyzed by different techniques. The literature is accumulated and relegated, the preceding work is recognized and then forwarded to the computing educators and professionals by the study explored in this paper. The edification and invigoration of impuissant segment students within the institution, well-fortified advises are given by this research [20]. To ameliorate the pedagogical process, presage the performance of scholars, provide a comparative analysis of precision of data mining algorithms and recognize the maturity of open source implements, these studies have provided good result outcomes.

Lee, E., Jang, *et al.* (2018) proposed an international competition on the game data mining. From one of the major game companies called NCSOFT, the commercial game log data was extracted to propose this technique. The data is made open by the researches for developing and applying previously proposed data mining techniques on the game log data. An action role-playing game from NCSOFT named as Blade & Soul was used to collect data for competition [21]. Around 100 GB of game logs were achieved from 10,000 players within the data. Predicting the possibilities of a player to churn is the major objective of the competition. The two periods in which the business model was modified and a free-to-play model was generated from a monthly subscription helped in defining the time in which the player would churn. Deep learning, tree boosting and linear regression techniques were applied as per the results achieved through the competitions amongst highly ranked competitors.

Table 1: Comparison of various techniques

Authors	**Techniques/ Algorithms**	**Datasets**	**Attributes**	**Tools Used**	**Shortcoming**	**Results**
Min Chen, *et al.*	Naïve Bayesian, KNN and Decision tree	Heart Diseases	79	MATLAB	This classifier has high complexity.	Decision tree performs better in comparison to other classifiers.
Akhilesh Kumar Yadav, *et al.*	Foggy K-mean Algorithm	Lung cancer Data	9	WEKA`	Complexity is high.	Foggy k-mean performs well as compared to K-means
Sanjay Chakrabotry *et al.*	Incremental k-mean clustering Algorithm	Air pollution Data	7	WEKA	Accuracy is less	The accuracy of proposed method is achieved up to 83.3 percent.
Chew Li S. *et al.*	BF Tree classifier	Student's Performance	9	WEKA	Complexity is high which increases the execution time.	BF Tree performs well as compared to other tree classifiers
Qasem A. *et al.*	Decision tree	STOCK Data Prediction	170	WEKA	Accuracy is less which can be increased.	C4.5 classifier performs well as compared to ID3
K.Rajalakshmi	Medical fast growing field	Prediction based systems	3	Python	A large amount of data has been generated and to handle this much of large amount of data	The cost effectiveness and human effects have been reduced using proposed prediction system based data mining.
BalaSundar	real and artificial datasets	to predict diagnosis of heart diseases	5	WEKA	The clusters are partitioned into k number of clusters by clustering which is the part of cluster analysis.	Show that the best robustness, and accuracy rate are often achieved using it.
Daljit Kaur	contains similar objects has been divided using clustering	dissimilar objects	12	Python	algorithm is able to reduce efforts of numerical calculation and complexity	The proposed algorithm is also able to solve dead unit problem.

Ming, J	multi-dimensionality and nonlinearity the Characteristics of the technical	technical and economic data	2	MATLAB	The prediction model of the geological missing data is established on the basis of techniques of geo statistics and artificial neural network.	during the process of mineral development there is a loss of a lot of geological data that decreases
Sakhare	a survey of road accident analysis methods an important role played in transportation	Road Accident Data Analysis	2	WEKA	This paper studied the k- mean algorithm in proper manner. SOM is employed to make and analyze the clusters. A self organizing technique uses the neural network along with an unsupervised learning method.	The accident reasons can be predicted and the accuracy of analysis can be improved to a greater extent in comparison to the k-means clustering algorithm by applying the proposed approach.
Chauhan, C., &Sehgal, S	Review of assorted algorithms and techniques which help in identifying the criminals	Criminal Data Analysis	3	MATLAB	Detecting the hidden links of networks of co- offenders was done using hidden link algorithms which helped in showing the possible future of crime partner.	It was concluded that the violent crimes were solved and the accuracy was limited by applying Criminal investigation analysis (CIA) tool.
Anoopkumar	different Data Mining Methods especially the mostly utilized	comprehensive survey	5	MATLAB	For improving the academic performances of students and then improving the effectiveness of institutions, the educational data is analyzed by different techniques.	To ameliorate the pedagogical process, presage the performance of scholar, provide a comparative analysis of precision of information mining algorithms and recognize the maturity of open source implements, these studies have provided good result outcomes.
Lee, E.	Commercial game log data competition framework was used for game data mining	d tested on the game log data of Blade & Soul of NCSOFT	3	MATLAB	From one of the major game companies called NCSOFT, the commercial game log data was extracted to propose this technique.	Deep learning, tree boosting and linear regression techniques were applied as per the results achieved through the competitions amongst highly ranked competitors.

III. CONCLUSION

This article shows an updated review of the literature on stock market prediction. It is focused on the works between 2010 and 2018 that perform prediction work of the stock market. Future prediction is done from the current information by the prediction analysis which is the technique of data mining. The most popular data for prediction of stock market is historical data.

There has been an increase in popularity of sophisticated machine learning algorithms, such as ensemble models and deep learning. The ensemble models have shown high predictive power, even in some comparative works they have performed better than other techniques such as SVM and ANN. Deep learning models, in general, have not outperformed traditional models. It is possible that the data sets with which these algorithms have been trained is not sufficient to generate an adequate prediction.

Then it is to be expected that future work can be focused on finding new sources of information that complement the technical analysis to predict stock markets. For example, in addition to analyzing sentiment about networks, it can be complemented with the analysis of the topics that are spoken on the networks, which leads to the models having a

more complex idea of what is happening in the world. Additionally, given the success in other types of problems of the free models of feature engineering, in the future, more articles are expected to find optimal technical indicators automatically.

Finally, there are missing comparative works among the fore- casting on developed, emerging, and frontier markets. It is crucial to find what kind of information and which modeling techniques are appropriate for each of the stock markets.

It is important to mention that there are some limitations to this review. First, it may have missed relevant articles that were not indexed in the selected databases. Second, though the definition of the keywords involved much work and several iterations, we could have missed some studies that used less common lan- guage terms to refer stock market prediction. Given the nature of this review, viz. an overview of academic literature, we are aware that much work done in the development of industrial software products is not reflected.

REFERENCES

[1] AbdelghaniBellaachia and ErhanGuven (2010), "Predicting Breast Cancer Survivability Using Data Mining Techniques", Washington DC 20052, vol. 6, 2010, pp. 234-239.

[2] Oyelade, O. J, Oladipupo, O. O and Obagbuwa, I. C (2010), "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", International Journal of Computer Science and Information Security, vol. 7, 2010, pp. 123- 128.

[3] AzharRauf, Mahfooz, Shah Khusro and HumaJaved (2012), "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity", Middle-East Journal of Scientific Research, vol. 12, 2012, pp. 959-963.

[4] Osamor VC, Adebiyi EF, Oyelade JO and Doumbia S (2012), "Reducing the Time Requirement of K-Means Algorithm" PLoS ONE, vol. 7, 2012, pp-56-62.

[5] AzharRauf, Sheeba, SaeedMahfooz, Shah KhusroandHumaJaved (2012), "Enhanced K-Mean Clustering Algorithm toReduce Number of Iterations and Time Complexity," Middle-East Journal of ScientificResearch, vol. 5, 2012, pp. 959-963.

[6] Thair Nu Phyu, "Survey of Classification Techniques in Data Mining", 2009, Proceedings of the International MultiConference of Engineers and Computer Scientists, volume 3, issue 12, pp- 551- 559, IMECS.

[7] Chuan-Yu Chang, Chuan-Wang Chang, Yu-Meng Lin, (2012) "Application of Support Vector Machine for Emotion Classification", 2012 Sixth International Conference on Genetic and Evolutionary Computing, volume 12, issue 5, pp- 103-111.

[8] HimaniBhavsar, Mahesh H. Panchal, (2012) "A Review on Support Vector Machine for Data Classification", 2012, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 1, Issue 10.

[9] Min Chen, YixueHao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang (2017), "Disease Prediction by Machine Learning over Big Data from Healthcare Communities", 2017, IEEE, vol. 15, 2017, pp- 215-227.

[10] Akhilesh Kumar Yadav, DivyaTomar and SonaliAgarwal (2014), "Clustering of Lung Cancer Data Using Foggy K-Means", International Conference on Recent Trends in Information Technology (ICRTIT), vol. 21, 2013, pp.121-126.

[11] Sanjay Chakrabotry, Prof. N.K Nigwani and Lop Dey (2014), "Weather Forecasting using Incremental K-means Clustering", vol. 8, 2014, pp. 142-147.

[12] Chew Li Sa., BtAbang Ibrahim, D.H., DahlianaHossain, E. and bin Hossin, M. (2014), "Student performance analysis system (SPAS)", in Information and Communication Technology for The Muslim World (ICT4M), 2014 The 5th International Conference on, vol.15, 2014, pp.1-6.

[13] Qasem A. Al-Radaideh, Adel Abu Assaf and EmanAlnagi "Predicting Stock Prices Using Data Mining Techniques", the International Arab Conference on Information Technology (ACIT'2013), vol. 23, 2013, pp. 32-38, (2013).

[14] K. Rajalakshmi, Dr. S. S. Dhenakaran and N. Roobin (2015), "Comparative Analysis of K-Means Algorithm in Disease Prediction", International Journal of Science, Engineering and Technology Research (IJSETR), Vol. 4, 2015, pp. 1023-1028.

[15] BalaSundar V, T Devi and N Saravan, (2012) "Development of a Data Clustering Algorithm for Predicting Heart", International Journal of Computer Applications, vol. 48, 2012, pp. 423-428.

[16] DaljitKaur and KiranJyot (2013), "Enhancement in the Performance of K-means Algorithm", International Journal of Computer Science and Communication Engineering, vol. 2 2013, pp. 724-729.

[17] Ming, J., Zhang, L., Sun, J.& Zhang, Y, "Analysis models of technical and economic data of mining enterprises based on big data analysis", International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), 2018, IEEE, 3rd.

[18] Sakhare, A. V., &Kasbe, P. S "A review on road accident data analysis using data mining techniques", International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017.

[19] Chauhan, C., &Sehgal, S, "A review: Crime analysis using data mining techniques and algorithms", International Conference on Computing, Communication and Automation (ICCCA), 2017.

[20] Anoopkumar M, & Rahman, A. M. J. M. Z, "A Review on Data Mining techniques and factors used in Educational Data Mining to predict student amelioration, International Conference on Data Mining and Advanced Computing (SAPIENCE), (2016).

[21] Lee, E., Jang, Y., Yoon, D.-M., Jeon, J., Yang, S., Lee, S, Kim, K.- J, "Game Data Mining Competition on Churn Prediction and Survival Analysis" using Commercial Game Log Data Transactions on Games, IEEE, 2018.

[22] Bustos O., Pomares-Quimbaya A. "Stock Market Movement Forecast: A Systematic Review" Elsevier, 2020.

[23] Chakarverti M., Sharma N. and Divivedi R.R., "Prediction Analysis Techniques of Data Mining: A Review" 2nd International Conference on Advanced Computing and Software Engineering (ICACSE), 2019.