DATA MINING TECHNIQUES AND ITS APPLICATION IN SECURITY MARKET

V NIRANJAN REDDY, K JAGAN MOHAN, T C VENKATA SIVA
ASSISTANT PROFESSOR [1,2,3]
anjanreddymsc@gmail.com, mohan.kjagan56@gmail.com, tcsiva222@gmail.com
Department of mathematics,
Sri Venkateswara Institute of Technology,
N.H 44, Hampapuram, Rapthadu, Anantapuramu, Andhra Pradesh 515722D

Abstract:- A number of financial indicators are used to calculate the stock price. However, we do not have a simple formula that can predict the precise behaviour of a stock price. But there aren't many factors that we're aware of that affect stock prices. We are talking about fundamentals, market sentiment, and the company's situation here.
When used correctly, data mining may help with stock market forecasting. We may examine patterns using various tactics and methodologies, and these patterns will tell us what the stock market trend will be. These methods and approaches, in turn, cause various decision-making algorithms to initiate stock transaction orders. The stock market's past and current performance may be used to foretell how the market will act in the future.

By calculating a few financial indicators, data mining methods may be utilised to examine past stock values and forecast their future trends.

1. Introduction

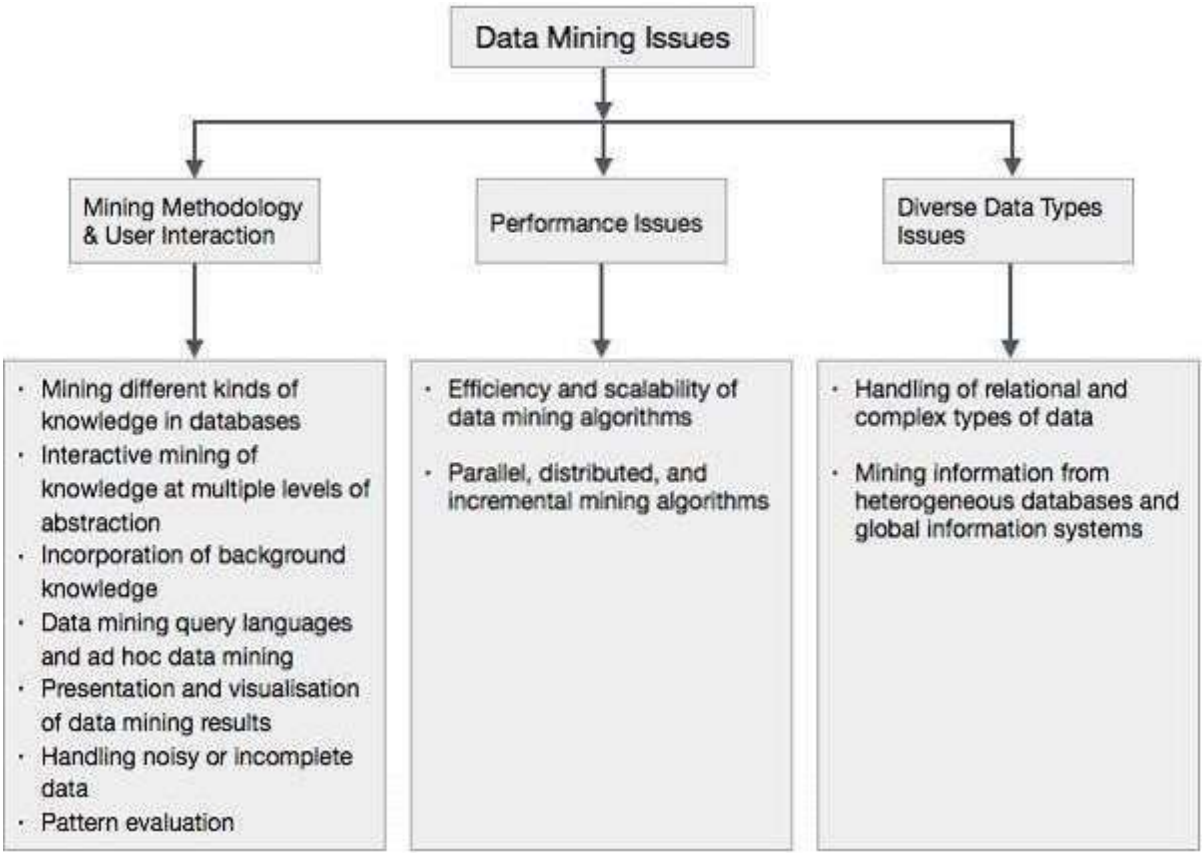
The process of discovering patterns in large datasets is known as data mining. To rephrase, data mining is the process of gaining insight from datasets. Data mining is the process of mechanically exploring massive data sets for insights that transcend conventional statistical analysis. Data mining may also be defined as the practice of sifting through massive data sets in search of patterns and links in order to find analytical solutions to issues. Businesses are able to foresee trends thanks to data mining techniques. Segmenting data and assessing the likelihood of future occurrences are two of data mining's primary applications for complex mathematical algorithms. Data mining is an emerging discipline that combines elements of statistics and computer science. Its overarching objective is to intelligently extract useful information from datasets and organise it in a way that can be understood and used. The "knowledge discovery in databases" (KDD) method includes data mining as one of its analytical steps. Data pre-processing, administration of databases, considerations of models and inference, metrics for interest and complexity, discovery of structures, post-processing, visualisation, and online updating are all part of it, in addition to the raw analysis stage.

Data mining's most important features are:

• Rendering useful insights from massive datasets and databases • Automated pattern recognition • Outcome prediction

Steps in the typical definition of knowledge discovery in databases (KDD) include:

1. Data Selection—Date Selection is the first step in retrieving data from the database that is relevant to the analysis activity. Prior to selecting data, it is sometimes necessary to alter and consolidate the data. In data mining, data kinds, sources, problems with systems, data mining functions and methods, and data mining coupling are all part of the data selection process.
Data Integration is a preprocessing method that unites disparate data sets from several sources into a unified database. Data cleansing is necessary because data integration might lead to inconsistent data. Step three, "transformation," involves doing summary or aggregate operations on the data in order to make it more suitable for mining.
4. Data mining - Data mining isn't a picnic, what with all the complicated algorithms utilised and the fact that data isn't usually in one central location. It requires integration from a wide variety of disparate data sources. Additionally, these elements cause a few problems.
You may see the main problems shown in the figure below.



1. A data warehouse's ability to aid in management's decision-making is shown by its interpretation and assessment capabilities.
Data Warehouse Is Subject Oriented - Instead of providing information about the organization's continuous activities, it gives us information about certain subjects. Product, client, vendor, sales, income, etc. are all examples of such topics. Data modelling and analysis for decision-making, rather than continuing operations, are the data warehouse's primary goals.
Data from several sources, including relational databases and flat files, is combined to form an integrated data warehouse. Data analysis is made more effective with this connection.
The data stored in a data warehouse is associated with a certain time frame, which is called the time variance. From a historical perspective, the data stored in a data warehouse offers valuable insights.
When new data is added to a non-volatile storage, the old data is not erased. The data warehouse does not reflect the regular changes in the operational database since they are maintained separate.

3. Methods for Classifying Data Mining

It is possible to classify data mining techniques as supervised or unsupervised. There is no explicit identification of a target variable in unsupervised approaches. The data mining algorithm, on the other hand, looks for structures and patterns in all the variables. According to Larose (2005) and 90, clustering is the prevailing unsupervised data mining technique. The learning job is referred to be unsupervised when the data is not labelled and every instance does not have a specific class label. Applying a clustering technique may help us determine which instances belong together, allowing us to create natural groups (Olafsson et al., 2008:1439). In order to better manage risks and investments, clustering methods may be used to discover reliable dependencies (Zhang and Zhou, 2004: 514).

Classification is a typical learning job in data mining. Classification, in contrast to clustering, allows for unsupervised learning. Predicted attributes are those in the database that indicate the class of a tuple; remaining attributes are referred to as predicting attributes (Sumathi & Sivanandam, 2006: 577). Among the many straightforward classifiers available, the Bayes technique stands out. Using the class label as input, this technique learns the conditional probability of each characteristic. The next step in classification is to use Bayes rule to determine the likelihood of each class value for each occurrence and then to forecast the class value with the greatest probability (Olafsson et al., 2008:1437). Using

graphical models, Bayes ethos is able to depict relationships among attribute subsets.

Decision Trees, Neural Networks, Rule Induction, Nearest Neighbours, and Genetic Algorithms are only a few of the data mining techniques included in another description. Kovalerchuk and Vityaev (2000) state that the lists of components of contemporary data mining tools provide the most realistic description, which is less formal. Numerous products are available, such as Intelligent Miner by IBM, SAS Enterprise Miner by SAS Corporation, Recon by Lockheed Corporation, MineSet by Silicon Graphics, Relational Data Miner by Tandem, Knowledge Seeker by Angoss Software, Darwin by Thinking Machines Corporation, ASIC by NeoVista Software, Clementine by ISL Decision Systems, Inc., Data Mind Data Cruncher by DataMind Corporation, Brain Maker by California Scientific Software, and Wiz Why by WizSoft Corporation (Kovalerchuk and Vityaev, 2000:16).
4. Market for Security

Public marketplaces for the issuance, purchase, and sale of stocks traded on stock exchanges or over-the-counter are collectively known as the stock market. The stock market is a marketplace where investors may purchase and sell shares, often known as stocks, which represent a fractional ownership in a corporation.

Predicting the path of the stock market may help both short-term investors with recommendations and long-term owners with early warnings of financial trouble. The first consideration when choosing a forecasting technique is the accuracy of the prediction. Making smart stock choices that are good investments is no easy feat. The subject of whether or not stock premiums or returns are predictable is one of the most researched topics in economics and finance. It was believed by financial economists up until the mid-1980s that returns could not be predicted, at least not in a meaningful economic sense, and that stock

The level of market volatility remains relatively constant. Because it lacks linearity and parametric data, the stock market is notoriously difficult to model accurately.

Other scholars have made use of technical analysis methods, which include formulating trading rules using stock price and volume history as inputs. The goal of technical analysis, which draws on a variety of tools and techniques, is to predict how stock prices and volume will change in the future. The idea behind it is that we can learn a lot about where the market is going in the future by looking at past prices. This leads one to believe that there are discernible patterns and trends in product prices.

Rough set method, decision trees, and artificial neural networks are some of the data mining and artificial intelligence approaches that have recently been applied to this subject. The process of obtaining useful information from massive databases is known as data mining. The Decision Tree Method is used in this work. The capacity control of the decision function is a defining feature of the Decision Tree Method, a subset of learning algorithms. The research delves into the topic of financial movement forecasting using Decision Tree.

5. The Study's Methodology

To forecast the stock market, and more specifically the purchase or sale of individual shares, we use a handful of data mining and clustering approaches in this article.

1. **Decision Tree**

There is a root node, some branches, and some leaf nodes in a decision tree. A class label is stored in each leaf node, an attribute test in each internal node, and the result of a test in each branch. A common and effective method for categorization and forecasting, decision trees are widely used. Data miners often utilise decision trees, a kind of predictive model, to depict both regression and classifier models. A tree's root node is its very first node. Having a decision tree has the following advantages: You don't need to know anything about the subject matter.

Reasons why decision trees are popular include: • Their ease of understanding; • The simplicity and speed of their learning and categorization processes; and • Their programability. • After training, it outperforms Neural Networks in terms of performance. • Trees are interpretable and offer a visual representation of data. • Information about which data impacts the prediction can be found at the top of the tree.

What is the Problem?

It is common for financial experts to have little idea how the stock market works, even yet they invest in it. Trading is proving to be a challenge for them since they lack the knowledge necessary to make informed stock purchases and sales. All the stock market data you could ever want is at your fingertips in the modern world. Manually or independently analysing all this data is really challenging. Therefore, the procedure must be automated. For this, data mining methods are useful. Aware that numerical time series research yields near findings, astute investors use machine learning methods to forecast stock market behaviour. By doing so, they will be able to anticipate how the stock in question would behave, enabling them to take appropriate action. We shall be using BSE India's historical data as input to our system. To determine the patterns in stock prices, suitable data would be used. In this way, investors will be able to maximise their chances of making a profit by acting upon the prediction model's alerts on the following trading day's up or down movement of the stock price.

Collect Input Data

The following procedures are used to extract data from the NIFTY 50:

1. The NIFTY 50 Companies are being considered for our proposal.
2. NSE India is the place to go for a list of NIFTY 50 firms.
To get NSE data, step three is to use the stock ticker from step a.
Using the company's stock data from the last two years, the system will
5. Next, we split the data into two sets: training and testing. The training set will include 75% of the data, while the testing set will contain 25%.

Solve the Problem

Decision trees with technical indicators are one of the supervised learning approaches that will be used to develop our model and solve the issue.

Here are the measures we will take to fix the issue:
1. Get a stock's NSE statistics for the last two years.
Two, figure out what the RSI, EMA, MACD, SMI, and other technical indicators are worth.
3. Use these indicators and training data to train the model.
4. Get the model tested using some test data.
5. Use a variety of assessment methods to assess our system.

Technical Indicator Details -

Indicator of Relative Strength - RSI One technical momentum indicator is the relative strength index (RSI), which attempts to identify whether an asset is overbought or oversold by comparing the size of recent gains to losses. To determine it, we use the following

the formula: relative strength index (RSI) = 100 - 100 / (1 + RS*) Where RS is calculated as the average of the up closes for x days divided by the average of the down closes for x days.

2. The EMA, or exponential moving average Compared to a standard moving average, the most recent data points have greater weight in an exponential moving average (EMA). The term "exponentially weighted moving average" may describe the exponential moving average as well.

• MACD, or Moving Average Convergence Divergence The link between two price moving averages may be shown by the trend-following momentum indicator known as moving average convergence divergence (MACD). To get the MACD, take the 26-day EMA and subtract it from the 12-day EMA.

Additionally, the MACD is overlayed with a nine-day exponential moving average (EMA) known as the "signal line" to serve as a trigger for buy and sell signals.

STMI (Stochastic Momentum Index) Technical momentum indicators like the Stochastic oscillator compare the closing price of a securities to its price range over a certain time period. Altering the time period or calculating a moving average of the output might make the oscillator less sensitive to changes in the market. This metric is determined using this formula: The expression %K = 100 [(C - L14) / (H14 - L14)] indicates that. The formula takes the most recent closing price (C), the lowest price (L14) from the 14 trading sessions before, and the highest price (H14) from the same 14-day period into account. %D is the average of %K across three periods.

CCI stands for the Commodity Channel Index. In technical analysis, an oscillator may assist you spot whether a certain investment vehicle is overbought or oversold. Originally devised by Donald Lambert, the Commodity Channel Index measures the correlation between an asset's price, its moving average (MA), and the standard deviations (D) from the MA. This may be calculated using the formula: Price minus MA divided by 0.015 times D is the CCI.

Mortgages with Additional Security—CMO Principal repayments in collateralized mortgage obligations (CMOs) are structured in a way that is both time- and risk-based, with maturities and risk categories serving as determining factors. Collateralized mortgage obligations are special purpose entities that hold the mortgages from which they receive cash flows and are designated to receive mortgage repayments. The mortgages are classified according to their risk profile and are used as collateral. Investors get their money according to the portion of the mortgages they've put their money into, which is referred to as a tranche, and this distribution is dependent on a set of regulations that have already been specified.

Change Rate (ROC) One technical indicator is the price rate of change (ROC), which compares the most current price to the price "n" periods ago and displays the percentage change between the two. The following formula is used to determine it:

As a percentage, the formula is (today's closing price minus "n" periods' closing prices) divided by "n" periods' closing prices.

Algorithm Design

First, gather the necessary information: [Updated Date, Open, High, Low, Close, Volume] Second Step: Determine All Necessary Indicators
1. The Relative Strength Index (RSI): In the case of a 5-period RSI, a value of 0 indicates that prices fell across all five periods. No benefits could be quantified. To have RSI equal to 100, set the Average Loss to zero. In other words, prices rose throughout each of the five decades. It was not possible to quantify the losses.
2. The EMA Crossover is a method for gauging price movement by subtracting the EMA Large Period from the EMA Short Period, where the shorter the period, the higher the opening price.
Chapter Three: CCI
2. ROC 3. CMO
6. WPR 7. ADX
Step 3: Determine the Up/Down prediction variable.
Construct the decision tree using the data obtained in the previous phases (Step 4). Fifth, cut down the tree to get rid of data overfitting.
The sixth step is to feed the trees test data.

Generate Output

Complete the following steps to get the desired result:
a) Using the stock's symbol, get data from the last three years.
c) Supply the system with the data.
c) Educate the programme.
d) The output will be predicted by the system.

The best way to demonstrate correctness?

We plan to split the dataset into two parts when we have it:
In order to construct predictive models, a training set is used.
A portion of the dataset used to evaluate a model's potential future performance is called a test set or unseen instances. Over fitting is likely to blame if a model fits the training set much better than the test set.

Discussion and analysis of data

1. Generating Output In order to produce a result, the programme takes in input data and runs the prediction algorithm on it. Confusion matrices, accuracy, root-mean-square-error, area under the curve, and ROC curves make up the output. All the stocks that were chosen provide this result.

Part two: talking Stocks from Nifty 50 firms were chosen for the system's training and testing. After downloading two years' worth of data from NSE, 75% is used for training purposes and 25% is reserved for testing purposes. The system was trained and tested using a large number of indicator functions and permutations of them. We used the indicator functions that produced the most accurate predictions out of all the ones we tried. When it comes to predicting the chosen stocks, the method works well.

Conclusion

The use of decision trees for directional financial movement prediction is the focus of this article. One kind of instrument that shows promise for financial forecasting is the Decision Tree. When compared to other individual categorization systems, Decision Tree provides the most accurate daily movement direction forecasts. This is a direct message for those who predict or trade in the financial markets, and it might result in a profit. The advantages and disadvantages of each approach should be carefully considered.

2. Grouping Approaches to Clustering Clustering is the process of arranging comparable things into clusters, where each cluster is distinct from the others. Attribute values, which sometimes include distance measurements, provide the basis of the similarities and differences. Some of the methods used for clustering include hierarchical, partitioning-based,

density-based, model-based, and grid-based approaches. Methods such as Partitioning, Hierarchical, Density-based, and Model-based approaches were used in this study.

1. Approaches to partitioning From a dataset with x data items, the partitioning technique generates k clusters or partitions. The programme begins by arbitrarily dividing the dataset into many halves, and then it iteratively moves the data items to a new division. A centroid, which is the average of all data items in a given partition, represents each partition. Each data object must include precisely one cluster, and every division must at least contain one data object. There are other clustering approaches available, such as k-means, k-medoids, "Partitioning Around Medoids" (PAM), "Clustering Large Applications based upon randomised Search" (CLARANS), "Clustering LARge Applications" (CLARA), and many more *1+. The emphasis of this work is on K-means.

One well-known unsupervised method is K-Means. The data items are divided into "k" clusters, where k is the number of clusters and should be explicitly stated beforehand. A centroid need to be present in every cluster. Taking into account the distance from each data point to the centroid. The data point is placed in the cluster with the centroid that is closest to it. Afterwards, with every cycle, the

Recalculating the distance reassigns data points to alternative clusters. When the centroid's position remains unchanged, the procedure terminates [4].

Method [4]:
Data entered: The quantity of clusters that were formed: k The quantity of items allocated: x The result: The similarity function "k" is used to produce clusters.
What to Do:a. With regard to the similarity function, allocate the remaining data points to each cluster; b. Next, randomly choose k data entities and use them as the initial cluster centroids. b. Use the cluster mean to update the centroid of each cluster.
iii)Until there is no further change

2. Thermodynamic methods Unwrapping clusters of any form is the objective of density-based approaches. The key point is that there should be at least a certain range of points in any given area for all data belonging to a certain category. To remove the "noise" outlier data, this technique may be used. A review of "Density Based Spatial Clustering of Applications with Noise" (DBSCAN) is presented in this study.

One key distinction between density-based and partitioning techniques is that density, rather than distance, is used to determine classification priorities [1].

1. Pick a point q at random (algorithm[9]) ii. Using "MinPts" and "H," get all points from q that are density-reachable. iii) To create a cluster, q must be a core point.
the programme moves on to the next database point as no sites are density-reachable from q if q is a boundary point.
v. Proceed as before until all data points have been processed.

3. Methods Based on Hierarchies The hierarchical clustering method builds a tree out of the nodes in a cluster and produces clusters according to the data hierarchy. The agglomerative and divisive methods of this kind of clustering algorithm are named after the distinct ways in which their respective hierarchies decompose. If it's agglomerative hierarchical clustering, a kind of bottom-up hierarchical decomposition, the method works by grouping related items into larger ones, one by one, until all of the objects in the set belong to the same cluster or until some other stopping condition is satisfied. While top-down hierarchical decomposition is known as divisive hierarchical clustering, the majority of hierarchical clustering algorithms fall into this category, which differs only in its definition of inter-cluster similarity; this algorithm is also discussed in this study. The items start off in a single cluster, unlike in agglomerative hierarchical clustering, and then they split into smaller and smaller clusters until they either hit a termination condition or each object belongs to its own cluster.

The BIRCH and CURE algorithms are examples of hierarchical clustering techniques. The problem with hierarchical approaches is that they cannot be undone after a merge or split has been done.

and are unable to swap clustering items with one another. Because of the helpful strong restrictions, we won't have to worry about the various combinations' choices, and the computing cost will be less. But the technology can't fix a

bad choice, which is one of its major flaws. The clustering findings could be of poor quality if the merge or split choice is made in a single step without proper consideration. In addition, this clustering approach is not very scalable as merging or splitting decisions need checking and estimating numbers of items or clusters. [1]

Method 4: Model-Based Approach A probability-based model is the basis of a model-based method. Parameters inside probability models are often determined using Maximum Likelihood Estimation (MLE). The goal of the model-based approach is to find the best possible mathematical models that match the provided data. It finds out a description of each group's characteristics, where each group stands in for a class.

5. Stock Prediction Using Regression Techniques To forecast a result from an input, regression is used. Linear regression is the most basic kind of regression analysis, while multiple regression is the most sophisticated. Simple linear regression refers to the use of a single descriptive variable, while multiple regression describes the use of numerous descriptive variables.

Fitting a Linear Model Predicting the connection between a dependent and independent variable is the goal of the statical approach known as Linear Regression. Where X is the independent variable, Y is a constant, and W is the slope of the regression line, generalisations are expressed as V=Y+WX.

Numerous Regressions The relationship between the scalar dependent variable "V" and one or more descriptive variables denoted by "U" may be represented using multiple regression, a method. As a function of other variables, it estimates how much one variable will be worth in the future.
The volume is equal to the sum of the weights, w0, y0, w1–wn, yn–wn+ Ɣ.

In this context, V represents the dependent variable, w0 to wn denote the coefficients, y1 to yn denote the independent variables, and Ɣ denotes the random error. The stock price of the company is expected to be predicted using the multiple regression method in this study.

Information **Detail**

The necessary information for Nifty 50 is retrieved from the National Stock Exchange (NSE). Six months was the time frame that was considered for each company's dataset. Open, close, high, low, previous close, and average price are the variables that have been used for clustering and categorization.

Experimental Result

Here we provide a concise summary of the outcome that was achieved by using the clustering and regression methods that were examined in this study.

This model takes data from the National Stock Exchange (NSE) and runs it through a number of clustering algorithms, including K-means in the partitioning technique, agglomerative in the hierarchical technique, EM in the model-based technique, and DBSCAN in the density-based method. We used the validation index to determine the algorithms that performed better. We employ the C-Index, Jaccard Index, Rand Index, and Silhouette-Index as our validation indices. Comparing the index results, the agglomerative and DBSCAN algorithms do poorly in comparison to the K-means and EM algorithms. In order to forecast the future value of its stock, the corporations using these algorithms use a regression approach known as multiple regression.

Finding a clustering algorithm that can use the aforementioned methods to produce the top firms for investment is the primary goal.

Conclusion

Here, we use NSE stock data to do clustering, and the result is a list of the top firms. After that, we use validation indices such the c-index, Jaccard index, rand index, and silhouette index to compare methods that are based on partitioning, hierarchical structures, models, and density. When compared to hierarchical and density-based techniques, K-mean salgorithm in partitioning-based techniques and EM algorithm in model-based techniques perform better. Multiple regression, one of the regression techniques for predicting future stock prices, is then applied to the clustered result. It is possible to build an online stock prediction system in the future by combining multiple regression with a partitioning-based or model-based approach.

5. Moving Forward and Concluding Remarks

This research proposes a method for generating buy/sell recommendations for stocks using decision tree classifier and clustering algorithms applied to stock price history. Investors may benefit from this suggested model since it analyses stock price history to extract predictive information, which can then be used to make informed stock investment decisions. Due to the fact that the stock market is sensitive to a wide range of variables, such as political events, macroeconomic circumstances, and investor expectations, the suggested model did not provide entirely accurate results.

There is a lot of space for improvement and testing in the suggested model, thus future work should focus on analysing it across all stock market businesses. Potentially fruitful avenues for further research include assessing a wider range of learning methods, including but not limited to neural networks, genetic algorithms, and association rules.

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 Match
[3] Wikipedia, "List of S&P 500 Companies," [Online]. Available: https://en.wikipedia.org/wiki/List_of_S%26P_500_companies.
[4] Wikipedia, "R,"
[5] Wikipedia, "RStudio," [Online]. Available: https://en.wikipedia.org/wiki/RStudio.
[6] Wikipedia, "Eclipse," [Online]. Available:

https://en.wikipedia.org/wiki/Eclipse_(software).

[7] Wikipedia, "Java," [Online]. Available: https://en.wikipedia.org/wiki/Java_(programming_language).

[8] SVMS, "Support Vector Machines vs Artificial Neural Networks," [Online]. Available: http://www.svms.org/anns.html.

[9] Stock Market Prediction, [Online]. Available, Pawan Kumbhare, Rohit Makhija, HiteshRaichandani.

[10] Clustering and Regression Techniques for Stock Prediction-Bini B.S. & Tessy Mathew [11]Hailong Chen ,Chunli Liu Research and Application of Cluster Analysis Algorithm,2013 2nd International Conference on Measurement,pp.575-579.

[12]Joel Joseph and Indratmo, Visualizing Stock Market Data with Self-Organizing Map,2013 Proceedings of theTwenty-SixthInternational Florida Artificial Intelligence Research Society Conference,pp.488-491.

[13]S.R. Nanda, B. Mahanty, M.K. Tiwari, Clustering Indian stock market data for portfolio management, 2010, pp.8793–8798.

[14]Shalini S Singh, N C Chauhan, K-meansv/s K-medoids:A Comparative Study, National Conference on Recent Trends in EngineeringTechnology,13-14May 2011.

[15]M.Suresh Babu, Dr. N.Geethanjali, Prof B.Satyanarayana, Clustering Approach to Stock Market Prediction, Int. J. Advanced Networking and Applications Volume: 03, Issue: 04, Pages:1281-1291(2012).

[16] Parul Agarwal, M. Afshar Alam, Ranjit Biswas, Issues, Challenges and Tools of Clustering Algorithms, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 2, May 2011 ISSN (Online):1694-0814.

[17] Saeid Fallahpour 1, Mohammad Hendijani Zadeh & Eisa Norouzian Lakvan, Use of Clustering Approach For Portfolio Management, Inter- national SAMANM Journal of Finance and Accounting ISSN2308-2356January 2014, Vol. 2, No. 1

[18] Anil K.Jain,Data clustering:50years beyond k-means,2009,pp.651-666.