International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072

Analysis for Enhanced Forecast of Expense Movement in Stock Exchange

Mrs. R.Pavithra[1], Mrs. S.Vaijayanthi[2], Mr. R.Kannan[3]

[1]Student, Department of Computer Science and Engineering, Gojan School of Business and Technology, Chennai.
[2]Assistant Professor, Department of Computer Science and Engineering, Gojan School of Business and Technology, Chennai.
[3]Assistant Professor, Department of Computer Science and Engineering, Gojan School of Business and Technology, Chennai.

ABSTRACT - *The Share market process is unstable and is affected by many criteria's. Hence the Stock market analysis is one of the major exertions in finance and business sector. Technical analysis is done by customer sentiments using historical data's by applying machine learning. This model can then be used to make predictions in stock values. The system is trained by using machine learning algorithm. Then the correlation between the stock values is analyzed. This model can make future analysis about share values. It can be shown that this method is able to predict the stock performance. In this system we applied prediction techniques approach in order to predict stock prices for a sample companies. In this analysis, a set of original technical, fundamental and derived data's are used in prediction of future values. The original data is based on previous stock data and the fundamental data ensures the companies' activity and the perception of market values. Here data's are classified and clustered are done by data mining techniques.*

Keywords: Sentiment Analysis, HDFS, Stock market prediction, Machine Learning, SVM Algorithm.

I INTRODUCTION

Now a day's Academia and Industry people are working on large amount of data, in petabyte, and they are using technique of Map Reduce for data analysis. The input for such framework is very large and main requirement for theses inputs are that all the files cannot be kept on single node. After putting all data on single machine, we have to process it parallel. Hadoop is a framework which enables applications to work on large amounts of data on clusters with thousands of nodes. A distributed file system (HDFS) stores the data on these nodes, enabling a high bandwidth across the cluster. Hadoop also implements a parallel computational algorithm, Map Reduce, which divides the main task into small chunks and these work in parallel known as mapping, and all the results are combined into a final output, the reduce stage. This paper is based Hadoop Based Stock forecasting using neural networks. Stock Market prediction has high profit and risk features and it gives parallel of accuracy, the main issue about such data are, and these are very complex nonlinear function and can only be learnt by a data mining method. We have tried to utilize distributing capability of Hadoop ecosystem which is parallel too. Map-Reduce for managing training of large datasets on the neural network. Our experimental results basically show the speedup achieved by increasing number of processors to the hadoop cluster for an artificial neural network. . To analyze the large volume of data and to process it, is difficult and challenging and there are different methods. Hadoop is a very fast way for massively parallel processing. Hadoop analyze the scattered data and predict the future trends and business intelligence solutions which would benefit the enterprise and client all together. In this system[4] we develop a broadly applicable parallel programming method. We adapt Google's map-reduce paradigm to demonstrate this parallel speed up technique on feed neural network.

Map Reduce: Map reduce is the high level programming system. This helps in doing the computation of the problem in parallel using all the connected machines so that the output, results are obtained in efficient manner. DFS also provides data replication up to three times to avoid data loss in case of media failures. The Master node stores the huge data HDFS and runs parallel computations on all the data i.e. Map Reduce.

1. The Name Node coordinates and monitors the data storage function (HDFS), while the Job

Tracker coordinates the parallel processing of data using Map Reduce.

2. Slave node do the exact work of storing the data and running the computations. Master nodes give the instructions to their Slave node. Each slave runs both a Data node and a Task Tracker daemon that communicate with their respective Master nodes.
3. The Data node is a slave to the Name node.
4. The Task Tracker act as a slave for the Job Tracker.

HDFS: HDFS is a distributed file system that provides a limited interface for managing the file system to allow it to scale and provide high throughput. HDFS creates multiple replicas of each data block and distributed them on computers throughout a cluster to enable reliable and rapid access. When a file is loaded into HDFS, it is replicated and fragmented into "blocks" of data, which are stored across the cluster nodes; the cluster nodes are also called the Data Node.[8] The Name Node is responsible for storage and management of metadata, so that when Map Reduce or another execution framework calls for the data, the Name Node informs it where the data is needed resides.

Variety - Big Data can be generated from various fields. The category of Big Data belongs to essential fact that needs to be known by the data analyst technique.

Velocity - The term ‗velocity 'in the context refers to the speed of generation or processing of data. Since the speed of generation of this data is very high it increases the level of difficulty to process it.

Complexity - Data management is a very complex process, especially when large volumes of data come from different sources. These data need to be linked, connected and correlated so that it can grasp the information that is supposed to be conveyed by these data. And hence the complexity of Big-Data

II OBJECTIVE OF STOCK ANALYSIS

The stock market shows the variation of the market economy, and receives millions of investors 'attention from the time of opening development each day. The stock market is characterized by high-risk, high-yield, hence investors are concerned about the analysis of the stock market and trying to forecast the trend of the stock market. However, stock market[6] is affected by various factors like the politics, economy , along with the complexity of its internal law, such as price changes in the non-linear, and so on therefore the traditional mathematical statistical methods to predict the stock market has not yielded suitable results. Thus, it is very suitable for the analysis of stock data.

SUPERVISED LEARNING APPROACH

- Supervised learning approach is used to build dictionary which is time consuming, because of initial level of manual work. They are used index tracking algorithm.
- Setting some threshold values words are added to either positive dictionary of negative dictionary. This approach is not suitable for real time analytics until the dictionary is complete.
- Stock market data used for stock market prediction [3], data set which is used for train the model is very less.
- In existing system they are find the sentiment of the user comments and predict stock market status.

III HADOOP SYSTEM

- This paper proposes a major information display for securities exchange forecast machine learning calculations.
- In rehearse; the machine will be prepared utilizing the profound learning technique.
- The information gathered from the online sources utilizing money markets programming interface. That information will be unstructured arrangement. That will clean utilizing the preprocess method.
- The cleaning information will be jumped into various parts and assembled into comparable data's. That every one of the information dealt with by the Hadoop system.
- The gathering information going to the calculation and foresee the element stock promoting report.

IV RELATED WORK

Ranking decision making units is an important applications in data envelopment analysis (DEA).The power of individual appreciativeness is prepared by developing a methodology that combines cross evaluation, preference voting and ordered weighted averaging. We used an ordered weighted averaging (OWA) operator to calculate the aggregated scores of 15 baseball players.Suggestion Discovery done by ranking via maximum appreciative cross-efficiency leads to ranking patterns whose structure is completely independent from the decision maker's optimistic level.

Hedge fund managers are exempt from much of the regulation faced by other investment managers, and typically employ dynamic trading strategies with frequent rebalancing, and often make extensive use of derivatives, short positions and leverage. Primarily aimed at institutional investors and high net worth individuals, hedge funds have recently become more widely accessible through the emergence of 'funds of funds,' which hold portfolios of hedge fund investments that are sold to a wider investor base. These funds provide a broad exposure to the hedge fund sector and diversify the risks associated with an investment in individual funds. the EWMA model, which is the most parsimonious of the dynamic models, offers a superior risk-return trade-off and, moreover, does so with rebalancing costs that are no higher than those of the static models. In the second out-of-sample period, which is characterised by much greater volatility and generally unfavourable conditions for the hedge fund industry, dynamic models again tend to outperform static models, providing a superior risk-return trade-off, although the differences are less marked than during the favourable conditions of the first out-of-sample period

V EXTERNAL INTERFACE REQUIREMENTS:

USER INTERFACE:

Graphical user interfaces are now the established standard. Nowadays, software is operated by graphical controls and symbolic images that are often designed to be objects from the 'real world'. It's normal for a user to employ his or her mouse and keyboard as a control device but touch screens are now becoming more popular. With the graphical user interface, icons have also moved into the digital world — such as the desktop, individual windows, and the trash can. The desired elements can be selected using the mouse or by tapping on the touch screen.

SOFTWARE INTERFACES

Hadoop framework is an open source data processing and storage of big data applications running in clustered interfaces. It is at the center of a growing ecosystem of big data technologies that are primarily used to support advanced analytics initiatives, including predictive analytics, data mining and machine learning applications[12]. Hadoop handles various forms of structured and unstructured data, giving users more flexibility for collecting, processing and analyzing data than relational databases and data warehouses provide.

A web application software is a software programmer which is stored on the Server and accessed via a web browser (Chrome, Firefox, IE, Edge, Safari, etc.). A more technical definition. 'A software programmer which is developed using web technologies (HTML, CSS, JS, etc.) and accessed via a web browser is called a web application'.

COMMUNICATION INTERFACES

HTTP is the used communication protocol. Load balancing algorithm is used in order to achieve better performance parameters such as response time and Data processing time. A proper communication interface to access internet is required.

VI SYSTEM FEATURES

Some of the important features of Hadoop features

- HDFS (Hadoop Distributed File System)
- Map Reduce

HDFS (Hadoop Distributed File System)

The Hadoop Distributed File System (HDFS) is store the primary datas by using Hadoop applications. It employs Name Node and Data Node architecture to implement a

distributed file system that provides high- performance access to data across highly scalable Hadoop clusters.

The core of Apache Hadoop consists of a storage part HDFS and a processing part MapReduce. Hadoop splits data files into large blocks and distributes them to the nodes in the cluster. Data process can be done by Hadoop MapReduce which transfers packaged code for nodes in parallel, based on the data each node needs to process their own Cluster. The HDFS file system, a subproject of the Apache project—is a distributed, highly fault-tolerance system designed to run on low-cost commodity hardware[11]. HDFS give high throughput efficiency to application data and it is suitable to apply large data sets.

MAP REDUCE

MapReduce is the heart of Apache Hadoop. It is a programming paradigm that enables massive scalability across hundreds or thousands of servers Hadoop cluster [11]. The MapReduce concept is easily understandable for those who are familiar with clustered scale--out data processing solutions.The term "MapReduce" actually refers to two separate and distinct tasks that Hadoop programs perform. The first is the map job, here set of data is taken and it is converted into another set of data, where individual elements are broken down into tuples (key/value pairs).The job reduce takes the input from the output of the map and join those minimum set of tuples is comeout to the data set of tuple[2]. The sequence implies from MapReduce, while performed the job reduce after the map job is take place.

VII ALGORITHM DESCRIPTION

SUPPORT VECTOR MACHINE

A Support Vector Machine (SVM) is a discriminativ classifier is formally separating hyper plane. In otherwise, given labeled training data .optimal hyper plane which categorizes new examples.

The Support Vector Machines[1] is the best understood with a simple example. We will have two categories: red and blue, and our data has two features: x and y. We want a classifier that, given a pair of (x,y)coordinates, outputs if it's either red or blue. We plot our already labeled training data on a plane: A support vector machine takes these data points and outputs the hyper plane (which in two dimensions it's simply a line) that best separates the categories. This line is the decision boundary: anything that falls to one side of it we will classify as blue, and anything that falls to the other as red. But, what exactly is the best hyper plane. For SVM, it's the one that maximizes the margins from both categories. In other words: the hyper plane (remember it's a line in this case) whose distance to the nearest element of each category is the largest.

candidateSV = {closest pair from opposite classes} while there are violating points do
Find a violator
candidateSV = candidateSV S violator
if any αp < 0 due to addition of c to S then
candidateSV = candidateSV \ p
repeat till all such points are pruned
end if
end while

SVM is generally used for text categorization. It can achieve good performance in high-dimensional feature space. An SVM algorithm[3] represents the examples as points in space, mapped so that the examples of the different categories are separated by a clear margin as wide as possible. The basic idea is to find the hyper plane which is represented as the vector w which separates document vector in one class from the vectors in other class.

SYSTEM DESIGN

SYSTEM DESIGN ARCHITECTURE

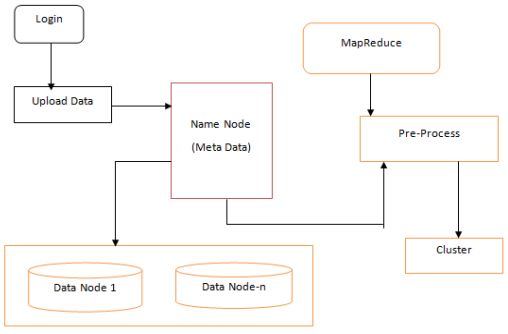

Fig 1

MODULE DESCRIPTION:

- Data uploading
- Preprocessing
- Data clustering
- Svm classification
- Report Prediction

VIII MODULE DESCRIPTION

DATA UPLOADING

The readied informational index will store the Hadoop document framework. HDFS occurrences are partitioned into two segments: the name node, which keeps up metadata to track the arrangement of physical information over the Hadoop case and data nodes, which really store the information. The information stacking to the hdfs utilizing hdfs url way. The transferred information will be kept up by the name hub and information hubs. The informational index traits are kept up in the name hub like record name, measure, get to consent, and so forth. The crude information kept up by the information hubs. The information hub controlled by the name node. The transferred information can't change any qualities in light of the fact that hdfs have compose once perused many time property.

PREPROCESSING

Information preprocessing is an information mining method that includes changing crude information into a reasonable arrangement. True information is frequently inadequate, conflicting, as well as ailing in specific practices or drifts, and is probably going to contain numerous blunders. Information preprocessing is a demonstrated strategy for settling such issues. The transferred information recover from the hdfs. The recovered information going to the MapReduce calculation and information will be composed into organized configuration. In this procedure have expelling the unusable qualities from the informational indexes. The information diminishment process is lessened portrayal of the information in an information distribution center.

DATA CLUSTERING

To gather those information into those bunches whose stock information class has been as of now characterized. In this way it develops a procedure to anticipate the promoting of the up and coming days. This one procedure gathering the information and ought to be made out of focuses isolated by little separations, in respect to the separations between groups. The information will gathering in light of the value, open, high, low, shut and time. In this bunching will apply the mapreduce with svm approach. It give more productive in high volume information bunching process.

PREDICTION

The cluster value will be different ranges. Those values are gathered and compared to each other's. Finally we will get the low and high result based on the calculation. The predicted values will be given the graphical representation graph.

ALOGORITHMS/TECHNIQUES

SUPPORT VECTOR MACHINE

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. The Supervised learning data, algorithm outputs an optimal hyperplane which gives new examples.

The basics of Support Vector Machines and how it works are understandable. Let's imagine we have two categories: red and blue, and our data has two features: x and y. We want a classifier that, given a pair of (x,y)coordinates, outputs if it's either red or blue. We plot our already labeled training data on a plane: A support vector machine takes these data points and outputs the hyperplane (which in two dimensions it's simply a line) that best separates the categories. This line is the decision boundary: anything that falls to one side of it we will classify as blue, and anything that falls to the other as red. But, what exactly is the best hyperplane?[7] For SVM, it's the one that maximizes the margins from both categories. In other words: the hyperlane (remember it's a line in this case) whose distance to the nearest element of each category is the largest.

candidateSV = {closest pair from opposite classes} while there are violating points do
Find a violator
candidateSV = candidateSV S violator
if any αp < 0 due to addition of c to S then
candidateSV = candidateSV \ p
repeat till all such points are pruned
end if
end while

SVM is generally used for text categorization. It can achieve good performance in high-dimensional feature space. An SVM algorithm represents the examples as points in space, mapped so that the examples of the different categories are separated by a clear margin as wide as possible. The basic idea is to find the hyperplane which is represented as the vector w which separates document vector in one class from the vectors in other class.

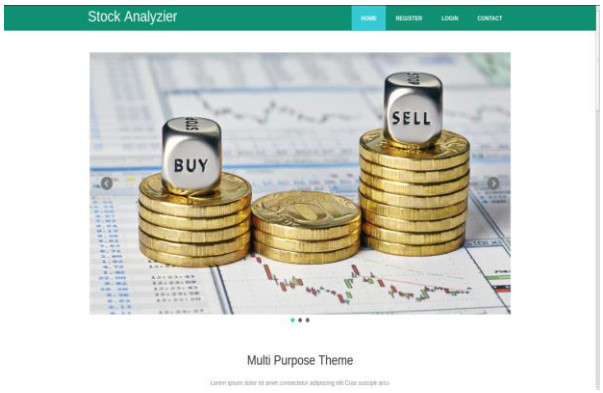

Fig 2 Home Page

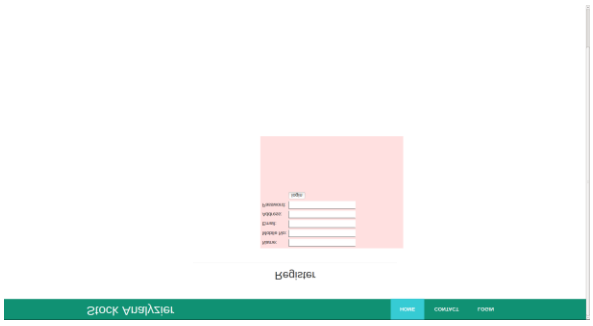

Fig 3 Register Page

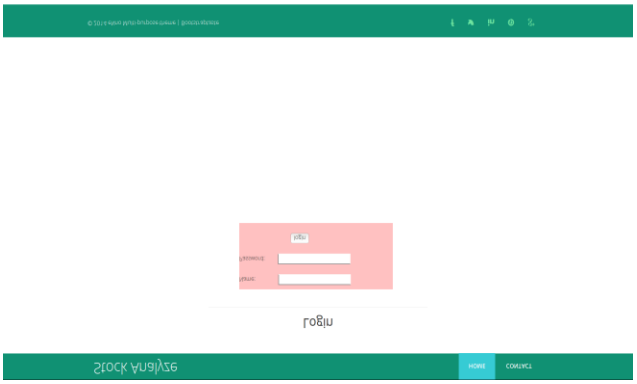

Fig 4

Login Page

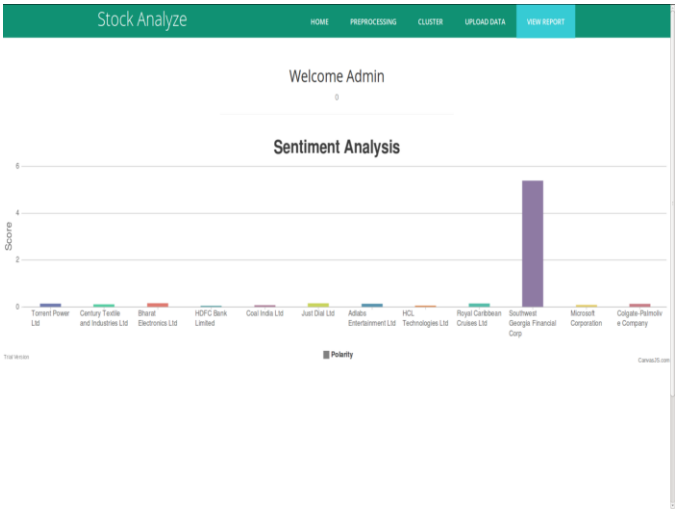

Fig 5

Final analysis

So the purposes can be described as:

- Describe the sequence from one activity to another.
- Describe the parallel, branched and concurrent flow of the system.

IX IMPLEMENTATONS AND TESTING

Implementation is the most crucial stage in achieving a successful system and giving the confidence that the new system is workable and affective. It is easy to handle and provides less changes in the system.

Each program is tested individually at time of development using the data and has verified that this program linked together in the way specified in the programs specification, CPU and its environment is tested. The system that has been developed , accepted and proved to be satisfactory for them to implementation. A simple operating procedure is included so that they can understand the different functions clearly and quickly. Initially as a first step the executable from of the application is to be created and loaded in the common server machine which is accessible to the entire and the server is too connected to a network. The last stage is to document the entire system with components and the operating procedures of the system.

In the Implementation the stage which are using I the project when the theoretical design is convert into a working system. It will considered as the critical stage in achieving a successful level of confidence that the new system will work effectively.

The implementation stage involves planning, investigation of the entire system and its constraints based on implementation, designing of methods to achieve change over and evaluation of their methods.

Validation is achieved through a series of black box and White box tests that demonstrates conformity with its special requirements. After validation test has been conducted, depending on the criteria the condition exists.

X CONCLUSION

The stock advertising information is expanding day by day with the information will be produced in various associations. The information will gathered and stacked into the hdfs utilizing the Hadoop system. The put away information broke down utilizing mapreduce calculation and SVM calculation used to characterize and grouping process. The mining method to foresee the stock promoting status in view of utilizing the authentic information like value, low, high, open and close everything utilizing the verifiable information.

REFERNCES

[1] W. Cho and M. Shaw, "Portfolio selection model for enhancing information technology synergy," IEEE Trans. Eng. Manag., vol. 60, no. 4, pp. 739–749, Nov. 2013.
[2] P. C. Pendharkar, "A data envelopment analysis-based approach for data preprocessing," IEEE Trans. Knowl. Data Eng., vol. 17, no. 10, pp. 1379–1388, Oct. 2005. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2005.155
[3] F. Zhang, "High-frequency trading, stock volatility, and price discovery," Social Sci. Res. Netw. Work. Paper Series, Oct. 2010. [Online]. Available: http://ssrn.com/abstract=1691679
[4] D. Dash.Wu, "Supplier selection: A hybrid model using DEA, decision tree and neural network," Expert Syst. Appl., vol. 36, no. 5, pp. 9105–9112, 2009. [Online]. Available:http://www.sciencedirect.com/science/article/pii/S095741740800910X
[5] J. Doyle and R. Green, "Efficiency and cross-efficiency in DEA: Derivations, meanings and uses," J. Oper. Res. Soc., vol. 45, no. 5, pp. 567–578, 1994.
[6] S. Lim, K. W. Oh, and J. Zhu, "Use of DEA cross-efficiency evaluation in portfolio selection: An application to korean stock market," Eur.J. Oper. Res., vol. 236, no. 1, pp. 361–368, 2014.
[7] A. A. Kirilenko, A. P. Kyle, M. Samadi, and T. Tuzun, "The flash crash: The impact of high frequency trading on an electronic market," Social Sci. Res. Netw. Work. Paper Series, Oct. 2010. [Online]. Available: http://ssrn.com/abstract=1686004
[8]A. Charnes, W. Cooper, and E. Rhodes, "Measuring the efficiency of decision making units," Eur. J. Oper. Res., vol. 2, no. 6, pp. 429– 444, 1978. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0377221778901388
[9] R. D. Harris and M. Mazibas, "Dynamic hedge fund portfolio construction," Int. Rev. Financial Anal., vol. 19, no. 5, pp. 351–357, 2010.
[10] A. J. Patton, "Volatility forecast comparison using imperfect volatility proxies," J. Econom., vol. 160, no. 1, pp. 246–256, 2011
[11] C. Dose and S. Cincotti, "Clustering of financial time series with application to index and enhanced index tracking portfolio," Physica A, Stat. Mech. Appl., vol. 355, no. 1, pp. 145–151, 2005.
[12] R. Roll, "Amean/variance analysis of tracking error," J.PortfolioManage.,vol. 18, no. 4, pp. 13–22, 1992.