An Efficient System to Predict and Analyze Stock Data u sing Hadoop Techniques

Article *in* International Journal of Recent Technology and Engineering (IJRTE) · October 2019

DOI: 10.35940/ijrte.B1824.078219

Suja Cherukullapurath Mana
Sathyabama Institute of Science and Technology

An Efficient System to Predict and Analyze Stock Data using Hadoop Techniques

Jithina Jose, Suja Cherukullapurath Mana, B Keerthi Samhitha

Abstract: Stocks, they assume significant job in keeping up the capital inflow of an organization or to keep up the business of the nation. As indicated by certain reports, 1.46 billion exchanges are done every day in NYSE. Increment of open enthusiasm on the promoting came about to expand the stock exchanges flawlessly. Because of the expansion of number of exchanges one can abuse the information and the examples which exists in the information by applying present day strategies like HDFS(Hadoop Distributed File System) .Taking the upside of size of the information and by allocating information to dispersed frameworks one can accomplish the plots from the information and when this procedure is done powerfully it can give an estimation of future examples.

Index Terms: Stock prediction, Linear regression, Logistic regression, Genetic algorithm.

I. INTRODUCTION

Positive trading helps an investor to increase his investment which helps a company to raise more funds. The raise in stock price of a company depends upon the research done using invested money and vice versa. The investment and the positive trading are interlinked. One can predict the market with the past patterns and data points which have occurred in similar consequences in the history of stocks. Implementing stock prediction techniques and publishing the results will encourage the investors to decide when to invest at that particular instance. Increasing the work in this field will help the investors in countries like India where only 2.50 percent of the total population are into investing because of un-predicted consequences we change the un-predicted nature of this field we can enhance the percent of investors. The data of the stocks market is collected online in the form of files where each file ranges from 0.5 to 1 GB size which perfectly suits for MapReduce techniques no additional software is required for this process. The data is uploaded to HDFS and to plot this data K-mean algorithm is implement in java.

II. RELATED WORK

Historically there were many stock predictions done using ML(Machine Learning),ANNs(Artificial Neural Networks) and GA(Genetic Algorithm). The commonly used prediction in ANN is the feed forward network which utilizes the backward propagation of errors which updates the weights of network and these are referred as Backpropagation Networks. Another is time recurrent neural network(RNN) or called as time delay neural network(TDNN) which is appropriate for predicting stocks.

A. Tobias Preis:

Another way of prediction is by using trends on social networks and search engines. Tobias Preis et al. [8] Showed a method for identifying online precursors for stock market moves. His way of strategies is based on search volume on Google trends. According to their scientific report the analysis is done on 98 terms of varying financial relevancies. According to Helen Susannah Moat, Tobias Preis demonstrated a link between number views of stock related terms on English Wikipedia and the stock moves are related to each other.

B. Text mining:

The headlines on Google Finance and Yahoo! Finance were used in Text mining process to forecast the Stocks Predictions [9] . But the methodology used in their process are been criticized by the market heads.

C. Aspect Structuring:

Also referred as Jacaruso Aspect Structuring(JAS) [10] is a trend Forecasting method which uses geopolitical time series datasets. This method addresses challenge of high dimensional data in which variables are immeasurable to be accounted which makes it a difficult method of forecasting. This method identifies the variable which is a primary influence for all the remaining factors. This observes the trend changes and significance of that variable with time. However this method cannot elucidate the multiple variety background factors.

III. PROPOSED SYSTEM

A. *Algorithms Used*

1. **Linear Regression:**

Linear regression predictive modelling is considered to minimize error by making predictions with minimal error for expandable datasets [11]. Which is perfectly suitable for ever expanding datasets like stocks. The representation is done with a line which represents the best fits for the input variables and the output variables. For linear regression to predict well we should remove the similar values and the noise

Revised Manuscript Received on July 08, 2019.

Jithina Jose,Assistant Professor, School of computing, Sathyabama Institute of science and Technology, Chennai,India.

Suja Cherukullapurath Mana, Assistant Professor,School of computing , Sathyabama Institute of science and Technology, Chennai,India.

B Keerthi Samhitha, Assistant Professor,School of computing , Sathyabama Institute of science and Technology, Chennai,India.

from the data. Linear regression is a simple and fast.

2. Logistic Regression:

This algorithm is from the field of statics. Logistic Regression is like regression done linearly to find the coefficient values and weights for each variable which is given as a input. Unlike linear regression [11] the prediction is transformed to logistic function using non-linear functions.

The logistic function converts values into bit values that is 0 or 1 and when plotted looks like big S. The data instance is divided into class 0 or class 1 which is useful to give more rational to the prediction.

3. Genetic Algorithm:

Genetic algorithms (GAs) are adaptive methods which can be used for solving optimisation and search problems. As their name suggests, genetic algorithms [11] are based on the genetic processes of real biological organisms. As in nature, populations evolve according to the principles of natural selection, also known as “survival of the fittest”. By this, genetic algorithms, if being suitably encoded, are able to “evolve” solutions to real world problems. 10 Genetic algorithms work with a population of “individuals”, represented by chromosomes, which are a possible solution to the given problem. In the process of “evolving”, each such chromosome is assigned a fitness score, that represents how good the chromosome is as a solution to the problem. The individuals with higher fitness are then more likely to be selected for reproducing, which is represented by the crossover phase, after which the obtained offspring, i.e. their chromosomes can be mutated in the mutation phase, as it is common in nature. Since the evolving of populations in genetic algorithms, works as natural selection, the final solution should be a chromosome with high fitness score, meaning a suitable solution to the problem. This is based on the fact that chromosomes with higher fitness value are more likely to be chosen for reproduction, whereas the other ones simply die out. For the purposes of our project, genetic algorithm is used in order to find such fittest individual (chromosome) that can be further used in real-time to identify correct patterns so that we could act upon them. What we are interested in is finding parameters for each of the aforementioned strategies, which are tuned in a way that they are not too large, such that it is realistic to detect patterns with those parameters. Hence, it is very important to create a suitable fitness function to evaluate how good a chromosome is. Additionally, it is also important to set up the constants that are part of the genetic algorithm.

Constant Description:

NUM_OF_GENERATIONS Number of generations of the individuals (chromosomes) which will go through selection, crossover and mutation phase after which the solution (best chromosome of the last generation) is obtained

NUM_OF_CHROMOSOMES Number of chromosomes in the population

SELECTIVITY The percentage of chromosomes that survive the selection phase

NUM_ELITE Number of chromosomes with highest fitness value that certainly survive the selection phase

CROSSOVER_PROBABILITY The percentage of chromosomes in the population that are chosen for breeding

MUTATION_PROBABILITY The percentage of genes in all chromosomes of the population that are mutated

The first population is generated randomly. After the execution of all phases for the number of generations we set, the fittest chromosome is obtained. The process is exactly the same for each strategy.

Why choose GA for this project?

There are a lot of machine learning techniques which can be used for algorithmic tradingstrategies. Among those, we chose Genetic Algorithm for several reasons. Besides it being verysuccessful in identifying graphical patterns, there are additional advantages such as the
following:

- It is very useful for complex, and not well defined problems
- It easily solves problems with multiple solutions
- Any bad solutions do not affect the end solution in a negative way, as they are simplybeing discarded
- It is simple to understand and implement
- It works by its own internal rules, hence does not need to know any rules of the problem
- Data from recent history have more impact on the final result without any additional cost.

However, it also has certain drawbacks, such as that there is no guarantee that it will find the global optimal solution, and has longer running time. 11 As the advantages for using genetic algorithms outweigh the shortcomings, it was a natural choice for our project.

DATA:

An order book is the list of orders that a stock exchange uses to record the interests of buyers and sellers. An engine is used to determine which orders can be fulfilled, i.e. which trades can be executed. For our project, we are using Tick Data derived from Order Book Data. Due to the high price needed to obtain Order Book Data, we were limited in terms of data that we can use. We managed to find datasets containing Order Book Data from Microsoft for 42 consecutive trading days, which we obtained from the website www.tradingphysics.com. The order book represents the data from the NASDAQ stock exchange. The format of the data that we have retrieved is described below:

Timestamp Milliseconds after midnight

Ticker Stock identifier

Order Unique order ID

T Message type. Allowed values:

Shares

- “B” - Add buy order
- “S” - Add sell order
- “E” - Execute outstanding order in part
- “C” - Cancel outstanding order in part
- “F” - Execute outstanding order in full
- “D” - Delete outstanding order in full
- “X” - Bulk volume for the cross event
- “T” - Execute non-displayed order

The quantity for the number of shares for the order for all message types, except 0 for the message types “F” and “D”

PREPROCESSING OF DATA:

In order to use the data, we needed to pre-process it and transfer it to tick data. Tick data means executed transactions. We processed the downloaded data and created new data in the following format:

Column Description

Tick ID Unique Tick ID for the file

Timestamp Milliseconds after midnight

T Message type (not used)
Shares Number of traded shares
Price The order price, the last 4 digits are decimal digits. The number needs to be divided with 10000 to convert into the currency value

IV. IMPLEMENTATION (APPLICATION FLOW)

A. *Organisation of implementation*

The organization of our implementation is split in three parts. The first part is the Genetic Algorithm itself. We have already explained how it works in the previous section. The second part is the training phase over the data that we have chosen for training, and the last part is the evaluation phase over the data that we have chosen for evaluation (different than the data for training). In order to test certain strategy, one needs to extend the Fitness Function explained earlier and implement the calculate fitness method in his/her own way. Before starting with the training and evaluation, the parameters for the genetic algorithm have to be defined. After the GA parameters are defined by the user, the data has to be split in training and evaluation data. When everything is setup, one can run the algorithm which will find the parameters for the strategy that is being trained.

B. *Training*:

The training part of the algorithm is executed in such way that each generation is trained over a set of days (window for generation training). In our case, the window contains 5 consecutive days. This window is moved after certain number of generations have been trained over it, for us this number is 4. More precisely, after every 4th generation we move the window for training. The total number of days for training for our project is 30 days. In the setup of the training part, we also define the starting amount of money for the strategy, for us this is 3000 dollars. After the training is finished we output the best chromosome and start the evaluation phase with the parameters derived from the training. We can see on figures 15, 16 and 17 how one chromosome is evaluated. In each strategy we have a pattern detection phase, which is different for each strategy. After the pattern is detected, we have a phase in which we act (buy => sell or sell => buy).

C. *Real time simulation – evaluation*

The evaluation part of the algorithm is the part where we test the parameters derived from the training phase. The evaluation part can also be a real time simulation. This is possible because of the way the data is received - it doesn’t matter if the data is read from a file or read from somewhere else.

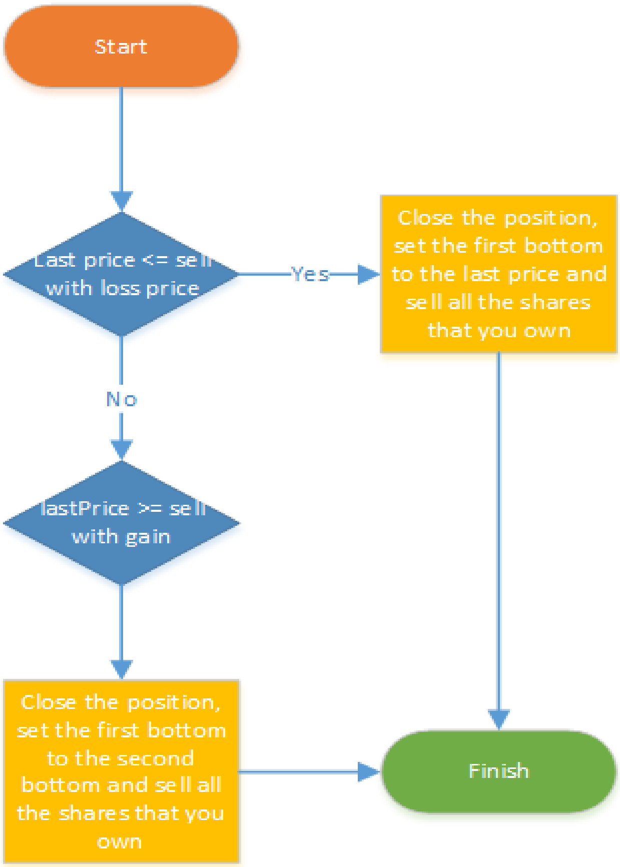
Start
Last price <= sell with loss price
Yes
Close the position, set the first bottom to the last price and sell all the shares that you own
No
lastPrice >= sell with gain
Close the position, set the first bottom to the second bottom and sell all the shares that you own
Finish

Fig. Flow for open position

While we receive data we try to detect the desired pattern and act upon it. One evaluation for the Double Bottom strategy is depicted on the pictures below. When we start the evaluation (trading) we check if we are in an open position (this means that we have bought shares and now we need to sell them). If we are in an open position we check if we can sell or not, otherwise we enter the pattern detection phase where we look for the pattern (in this case for Double Bottom).

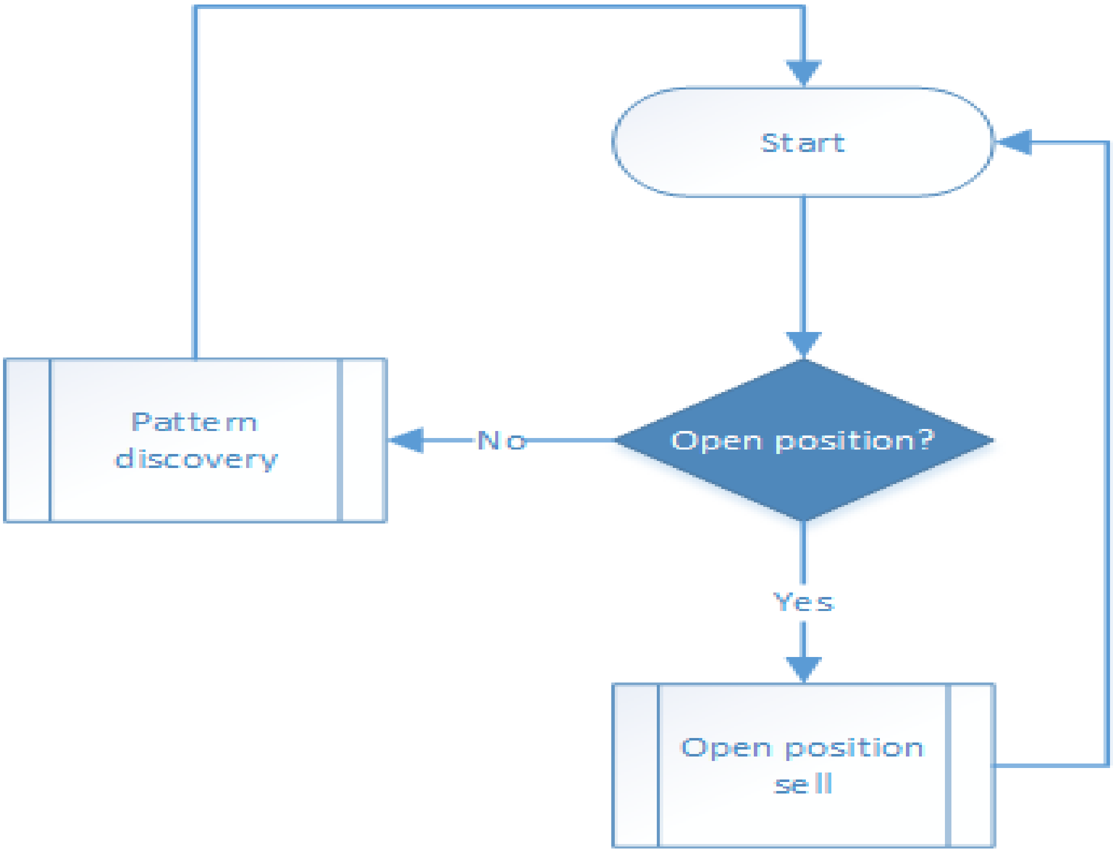

Fig. Main for trading

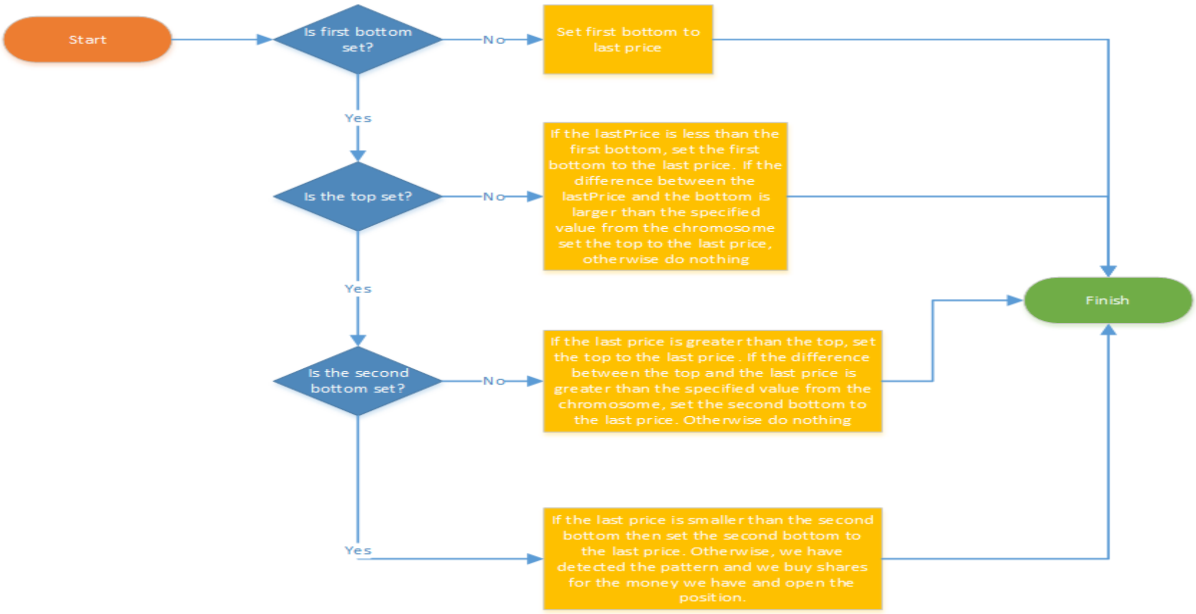
Start
Is first bottom set?
No
Set first bottom to last price
Yes
Is the top set?
No
If the lastPrice is less than the first bottom, set the first bottom to the last price. If the difference between the lastPrice and the bottom is larger than the specified value from the chromosome set the top to the last price, otherwise do nothing
Yes
Is the second bottom set?
No
If the last price is greater than the top, set the top to the last price. If the difference between the top and the last price is greater than the specified value from the chromosome, set the second bottom to the last price. Otherwise do nothing
Finish
Yes
If the last price is smaller than the second bottom then set the second bottom to the last price. Otherwise, we have detected the pattern and we buy shares for the money we have and open the position.

Fig. Flow of pattern detection

V. DISCUSSION OF RESULTS

In the evaluation of our algorithm for the strategies that we have implemented we observed the following results. We have retrained and evaluated each strategy 10 times. Every training had the following parameters Starting number of chromosomes was 100 Training was performed on 30 consecutive days 100 generations of chromosomes were created Each generation was evaluated over a window of 4 consecutive days After every 4th generation the window was moved for one day Starting amount of money was 3000 dollars When each training phase finishes, we start the evaluation of the best chromosome over the rest of the data (12 consecutive days). In average for every strategy we received the following results (The average is calculated over 10 runs): Double Bottom - amount 3125, number of transactions 139.8 Double Top - amount 3378, number of transactions 109.8 Rectangle - amount 3078, number of transactions 69.2 Head and Shoulders - amount 3240, number of transactions 6 Random - amount 17, number of transactions 318.3.

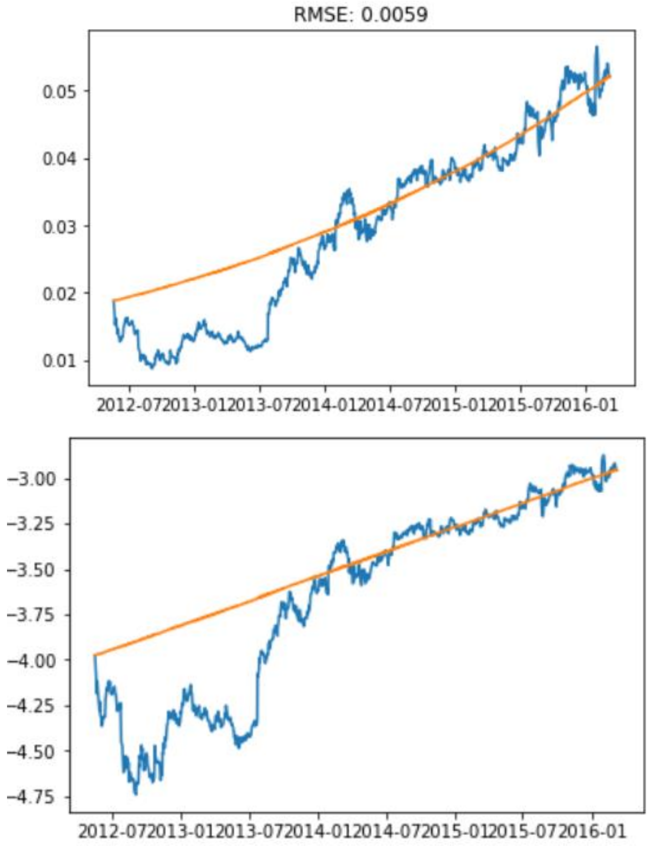

Fig. Average profit percentage after trading for 12 days

VI. CONCLUSION

As the area of algorithmic trading is gaining more and more attention, order book data which is the list of orders that a stock exchange uses to record the interests of buyers and sellers, is a great source of data that can be exploit. For this project, we divided our dataset (order book data for Microsoft for 42 consecutive days) into two sets, one for training and one for evaluation, respectively. For the training part, we used genetic algorithms in order to obtain the fittest chromosomes with parameters defining the patterns for each of the implemented trading strategies, Double Bottom, Double Top, Rectangle and Head and Shoulders. In the evaluation phase, we used each of those chromosomes in a real time simulation in order to evaluate how well they perform on unseen data. The results show that using each of those strategies, we obtained profit, with the highest profit of ~13% of the given starting amount. Furthermore, the results were compared with the results of a random algorithm that buys and sells with a certain probability, such that it makes transactions that are the average number of transactions made by the other implemented algorithmic strategies. The comparison shows that the random algorithm loses on average, and it performs much worse than the other strategies. It follows that the implementation of these four strategies, and their combination with genetic algorithms, is a promising starting point for investigating algorithmic trading strategies for real life implementation. Although our implementation is not perfectly realistic, as it makes some simplified assumptions, it is able to show that Double Bottom, Double Top, Rectangle and Head and Shoulders perform well in predicting the future trend. Furthermore, genetic algorithms turned out to be a good choice for this project.

ACKNOWLEDGMENT

We would like to thank our under graduated students C.H. Tejeswara Reddy , T. Sesha Sai Reddy for helping the successful completion of this project.

REFERENCES

1. Ian J. Goodfellow, Jean Pouget-Abadie , Mehdi Mirza, Bing Xu, David Warde-Farley, SherjilOzair , Aaron Courville, YoshuaBengio. (2014). Generative Adversarial Nets.InarXiv
2. G.Geetha, R.Samuel Selvaraj (2011). Prediction of monthly rainfall in chennai using back propagation neural network model. In International Journal of Engineering Science and Technology.
3. Diederik P. Kingma, Jimmy Lei Ba (2017). Adam: A Method For Stochastic Optimization. In arXiv
4. Y. Le Cun (1988). A Theoretical Framework for Back Propagation. In Connectionist Models Summer School.
5. Antonia Creswell, Tom White, Vincent Dumoulin, KaiArulkumaran, Biswa Sengupta, Anil A Bharath (2017). Generative Adversarial Networks: An Overview. In arXiv.
6. Mislan, Haviluddin, SigitHardwinarto, Sumaryono, Marlon Aipassa (2015). Rainfall Monthly Prediction Based on Artificial Neural Network: A Case Study in Tenggarong Station, East Kalimantan. Indonesia. In International Conference on Computer Science and Computational Intelligence.
7. Aakash Parmar, Kinjal Mistree, Mithaila Sompura (2017). Machine Learning Techniques For Rainfall Prediction: A Review. In International Conference on Innovations in information Embedded and Communication Systems (ICIIECS).
8. Preis, T., Moat, H. S., Stanley, H. E. & Bishop, S. R. Quantifying the advantage of looking forward. *Sci. Rep.* **2**, 350 (2012).
9. J. Garcke, M. Griebel (Eds.), Intraday foreign exchange rate forecasting using sparse grids, Sparse grids and applications, Springer, Berlin Heidelberg (2013), pp. 81-105
10.https://en.m.wikipedia.org/wiki/Stock_market_prediction
11. S.B. Kotsiantis, Supervised machine learning: a review of classification techniques, Informatica, 31 (2007), pp. 249-268
12. Apiletti D, Baralis E, Cerquitelli T, Garza P, Pulvirenti F, Michiardi P (2017) A parallel MapReduce algorithm to efficiently support itemset mining on high dimensional data. Big Data Res 10:53–69
13. S. C. Mana, "A Feature Based Comparison Study of Big Data Scheduling Algorithms," *2018 International Conference on Computer, Communication, and Signal Processing (ICCCSP)*, Chennai, 2018, pp. 1-3.doi: 10.1109/ICCCSP.2018.8452837

AUTHORS PROFILE

Ms. Jithina Jose is working as an assistant professor in sathyabama institute of science and technology. Her research interests are in the field of Big data wireless sensor networks, machine learning and networks.

Ms. Suja C.M is working as an assistant professor in sathyabama institute of science and technology. Her research interests are in the field of Big Data, machine learning and artificial intelligence.

.

Ms. B Keerthi Samhitha is working as an assistant professor in sathyabama institute of science and technology. Her research interests are in the field of Big Data, machine learning and image processing.