International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

An Implementations of Clustering Technique in Data Mining to Analyse Big Data in Finance

[1]Dr. Rajesh Gargi, [2]Sheenu Sachdeva, [3]Pooja Majoka
[1]Principal of JCDM College of Engineering, Sirsa, Haryana, India
[2]Assistant Professor, Department of CSE, JCDM College of Engineering, Sirsa, Haryana, India
[3]M.Tech. Scholar, Department of CSE, JCDM College of Engineering, Sirsa, Haryana, India

ARTICLEINFO

Article History:

Accepted: 01 June 2023
Published: 12 June 2023

Publication Issue
Volume 9, Issue 3
May-June-2023

Page Number
448-454

ABSTRACT

This research work is proposed to manage Big Data in Finance using Clustering and optimization mechanism. There are different knowledge discovery techniques, applications and process models that are applicable to deal with big data. There are several researches related to Big Data in Finance. A review of existing researches is stated here. Different researches are proposed research considering K-different clustering mechanism. These algorithms have been applied to various real-life applications running in serial, parallel, and high-performing computing environments. But these researches and traditional data mining techniques have their own limitations. The aim of researches is to judge the efficiency of different data mining algorithms on dataset and determine the optimum clustering algorithm. The performance analysis depends on many factors encompassing test mode, distance function and parameters. There are several researches which used K-MEAN clustering and optimization mechanism. The issues related to Kmean clustering would be resolved in this research. Research has introduced the more effective cluster mechanism to classify the data set. Therefore it is essential to propose a data mining technique to deal with big data in Finance. This research work would be helpful to known about data mining of big data. Data related to Equity and Mutual fund is considered in proposed model. After getting data, it is classified in to different clusters. Same data is stored in a cluster. Each cluster has different type of data set but the data within a cluster have similarly. There are different factors such as face value, market cap, promoters holding , fi holding, domestic holding, 52 week low, 52 week high , dividend, P/E are considered to analyze the better result in future. Using this proposed module, it will be possible to determine the shares or funds that will provide more profit.

Keywords : Kmean, Big Data, PSO

I. INTRODUCTION

Optimization of share price in order to consider whether to invest in particular share or not is theme of research. Research has been provided on large cap, mid cap & small cap fund which are available in large, medium and small amount for th purpose of investment. It has been observed that share in large cap are more reliable but provides limited return. However share in case of mid cap are found relatively more risky than large cap but return might be high.

Big Data:Big data has been determined as a term that is used to denote a huge data set of data sets. Such big data are very huge in size and difficult to manage. The existing data processing applications that are used to manage the big data are insufficient. It is not easy to analysis, capture, and searches the big data. Several issues are there such as sharing, storage, distribution of the big data. In addition, visualization, querying, and updating with data security are also some challenges. There are various algorithms and techniques as Classification, Clustering, Regression, Artificial Intelligence, Neural Networks etc. Association Rules, Decision Trees, Genetic Algorithm, etc. have been applied in discovery of knowledge from databases.

If there is the accuracy in big data, there will be easy to made confident decision and better decisions. After that there will be the increments in operational efficiency and decrement in cost reduction.

Clustering: Cluster Analysis Or Clustering Is The Task Of Grouping A Set Of Objects In Such A Way That Objects In The Same Group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).Clustering, in the context of databases, refers to the ability of several servers or instances to connect to a single database. An instance is the collection of memory and processes that interacts with a database, which is the set of physical files that actually store data. Clustering is the process of grouping a set of objects into classes of similar objects. In simple words, the aim is to segregate groups with similar traits and assign them into clusters. Documents within a cluster should be similar. Documents from different clusters should be dissimilar.

K Mean Clustering: The k-means clustering algorithm is a data mining and machine learning tool used to cluster observations into groups of related observations without any prior knowledge of those relationships. By sampling, the algorithm attempts to show in which category, or cluster, the data belong to, with the number of clusters being defined by the value k. The k-means algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging, biometrics, and related fields. The advantage of k-means clustering is that it tells about your data (using its unsupervised form) rather than you having to instruct the algorithm about the data at the start (using the supervised form of the algorithm).It is sometimes referred to as Lloyd's Algorithm because the standard algorithm was first proposed by Stuart Lloyd in 1957.

Optimization Techniques

PSO:It becomes a method of assessment. It exists in the form of method which is quite simple in use and implemented on regular basis. It was already assessed that such type of assessement methods discovers best possible solution in a very efficient manner. In the field of information technology, one can define this technique as a method that is able to optimize any considered problem. It is observed that in PSO based model, the efforts are put one by one for enhancing the performance of candidate solution. It deals with any issue of population related to candidate solutions. Here around in search-space, the dubbed particles move. This technique performs on the basis of arithmetical rule above position and velocity of particle. Its domestic well known location make a huge impact on its movement. Exactly, moved in the direction of its well known positions in the search-space. This location is updated in the form of better positions. Thses location can be easily identified by other particles. This is expected to move the swarm toward the best solutions PSO is a met heuristic as it makes few or no

assumptions about the problem being optimized. However, Meta Iheuristics such as PSO do not guarantee an optimal solution is ever found. In the present scenrio, it becomes most important and useful met heuristics because it showed success of various optimization problems after applied on. It is a self-organized model. It specified the activeness of this complicated systems. In order to take care of optimization problems, in a cooperative and smart structure it use an extremely streamlined model of social conduct.

MVO: It is a new type of invention. It is an effective maximization method which gets encouragement from environment. Mirjalili et al invented this. For putting this in to operation, two customized factors was kept in mind by them. This method is invented by using three ideology of cosmology. In addition to this form, it also becomes famous in new form of meta-heuristic optimization method. It efficiently figures out those problems which are related to OPF. It is a method which gets continuous motivation from living body & social science stand point. In working of this method different ideology of cosmology are bring in to use. In addition to idea of white & black hole, concept of wormhole is also used in this method. One of most important strong point of this method is that it will find out fast rate of intersection. For this purpose it use roulette wheel selection. In addition to this, this algorithm is able to deal with regular & discrete optimization issues.

II. Literature Review

D. Asir Antony Gnana Singh et. al. [6] proposed the research work on efficiency evaluaton on clustering concepts. Clustering has been determined as a process in which a group of abstract objects has been converted into classes of similar objects. Several methods clustering methods are there such as Partitioning technique, Hierarchical technique, Density-based technique, Grid-Based technique, Model-dependent technique, and constraint-based technique.

Kalyani M Ravalet. al. [7]has explained the method of data mining.

V.Ramesh, et. al. [8] analyzed the presentation evaluation of DM methods. There will be less risk. Cluster in big data has been known as a group of objects. In a cluster the objects are of same class. In the different words, it can be said that the similar kind of objects have been grouped in a cluster. Similarly the mismatched objects have been grouped in a separate cluster..

NeelamadhabPadhyet. al. [9] offered the review of the Applications of Data Mining with its future Scope. It has been considered and applied in the implementation. In the research work the issues related to empty cluster in K-mean

David Jensenet. al. [10] discussed the data mining with social networks. Clustering evaluation has been applied in many applications. Such applications are market research, pattern recognition, and data analysis. K-means clustering has been considered as partitioning technique. Aarti Sharmaet. al. [11] offered the application of data mining.

Kun-Ming Yuet. al. [12] explained a perfect frequent patterns mining algorithm.

Che-Yu Lin et. al. [13] proposed research work on open candidate slicing frequent pattern mining algorithm. they proposed on Hadoop acceleration in an open flow-based cluster. This paper presents details of their preliminary study of how Hadoop can control its network resources using Reduces get delayed because of inadequate bandwidth among them. It degrades the cluster performance.

In 2014, C.Vorapongkitipun et al. [14] wrote on improving performance of small-file accessing in Hadoop. To maximize efficiency, name node stores the entire metadata of HDFS in its main memory. With too many small files, name node can be running out of memory. In this paper, they propose a mechanism based on Hadoop Archive (HAR), called New Hadoop Archive (NHAR), to improve the memory utilization

for metadata and enhance the efficiency of accessing small files in HDFS.

In 2016, C. Verma et al. [15] discussed the big Data representation for grade analysis through Hadoop framework. Big Data is well known dataset. It is displaying volume, velocity features. It also considers variety in an OR relationship. They are enabling capturing, storing and subsequently analysis of Big Data.

In 2014, A. Siretskiy et al. [16] stated the HTSEQ - Hadoop with Extending HTSEQ for massively parallel sequencing data analysis using Hadoop. Hadoop has been considered as convenient framework. It has been enabling scalable distributed data analysis. They use the Hadoop-streaming library which allows the components to run in both Hadoop and regular Linux systems.

In 2017, K. Rattanaopas et al. [17] wrote on Improving Hadoop Map Reduce performance with data compression: A study using word count job. It has map reduce engine for distributing data to each node in cluster. Research has discussed some famous Hadoop's compression codecs for example; deflate, gzip, bzip2 and snappy. An over-all compression in map reduce, Hadoop uses a compressed input file which is gzip and bzip2. This research goal is to improve a computing performance of word count job using a different Hadoop compression option.

In 2012, A. B. Patel et al. [18] reviewed the Addressing of big data problem using Hadoop and Map Reduce. Hadoop cluster has been required to execute and analyze big data. It has map reduce engine for distributing data to each node in cluster. Compression is required as it is not just increasing space of storage. It also improves performance to calculate job.

In 2018P. R. Merla et al. [19] presented the Data analysis using Hadoop map reduce environment. The research work has provided the evaluation of YouTube data. They have done this with the use of Hadoop map reduce system on the base of cloud platform AWS.

In 2017.J. Kaur et al. [20] presented the Image processing on multimode Hadoop cluster. Research represents that data produced over internet is increasing day by day at exponential rate. The classical methods are insufficient for processing is termed as 'Big Data'. There are various tools in Hadoop to analyze the textual data such as Pig, base, etc.

In 2016A. Bhardwaj et al. [21] did research on Analyzing Big Data with Hadoop cluster in hd insight azure Cloud. Presently Cloud dependent Hadoop is gaining huge interest. It is offering ready to use Hadoop cluster environment in order to process Big Data. It has been excluding the operational issues in case of on-site hardware investment, IT support, and installing, configuring of Hadoop components such as HDFS and Map Reduce.

Problem Statement

In the traditional work there are different clustering algorithms which have advantages and disadvantages. The present research work is discussing such limitations. There are a variety of algorithms which are used in Finance systems for clustering such as hierarchical, partitioned; density based clustering according to the factors: methodology, structure, model, application or suitability, usefulness. But it is analyzed the tradition Finance systems are not sufficient and have their own limitations. Therefore it has become essential to propose an innovative and fast Finance System that would be efficient to deal with big data. During study of different clustering mechanism the Kmean clustering with big data has been found significant mechanism to cluster the data. But it has certain limitation. There is remains the issue of empty clusters after clustering in Kmean clustering mechanism. This leads to wastage of space due to empty cluster creation. Existing researches covers clustering algorithms, benefits and its applications. The performance analysis depends on many factors encompassing test mode, distance function and parameters.However there are several researches in field of Artificial intelligence. The issue with existing

research is that limited work has been done to resolve real life issues. The tradition work has not implemented optimization mechanism in order to predict optimal value of shares. Moreover there is need to give more exposure to multi verse optimization mechanism.

III.Research Methodology

Research methodology is a method which is followed to conduct research work of a topic. There are different research methodologies. Quantitative researches are systematical investigation on defined topic whereas qualitative researches provide study of research subject. These researches are descriptive & apply reasoning. The Researchers could use qualitative & quantitative research methodology together in their research work. The experiments based researches are systematic, scientific approach & provide a result whereas survey based research provides us review on a topic.

IV. Result after simulation

The convergence curve plotted in case PSO is shown below

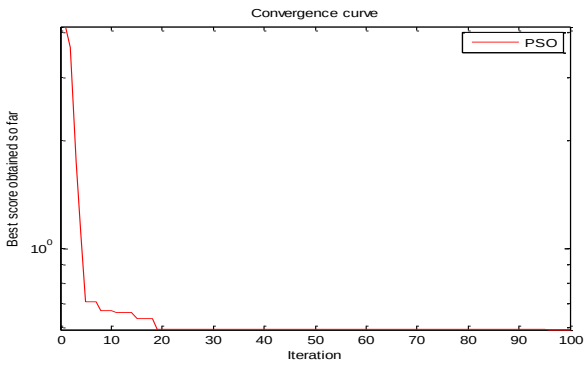

Fig 5.1 Convergence cure in case of PSO

The optimum solution and objective value is showing in following window

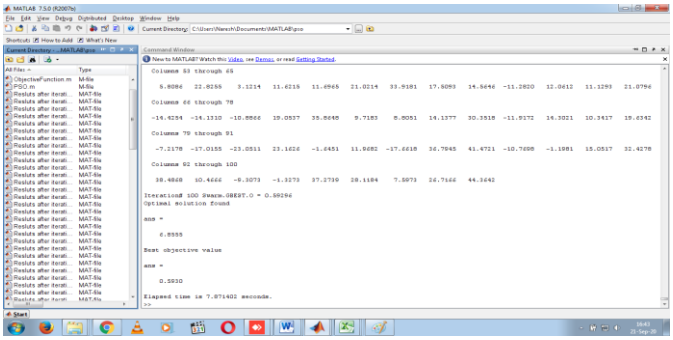

Fig 5.2 Optimization using PSO

As result have shown the optimum solution is 6.8555
Best objective value 0.5930
Elapsed time is 7.871402 seconds.

Get optimal solution using MVO

After getting optimal solution using MVO following results are produced.
At iteration 50 the best universes fitness is 0.666
At iteration 100 the best universes fitness is 0.59404
At iteration 150 the best universes fitness is 0.59404
At iteration 200 the best universes fitness is 0.59404
At iteration 250 the best universes fitness is 0.59302
At iteration 300 the best universes fitness is 0.59302
At iteration 350 the best universes fitness is 0.59302
At iteration 400 the best universes fitness is 0.59297
At iteration 450 the best universes fitness is 0.59297
At iteration 500 the best universes fitness is 0.59296
The best solution obtained by MVO is : 6.8561
The best optimal value of the objective funciton found by MVO is : 0.59296
Elapsed time is 5.949338 seconds.

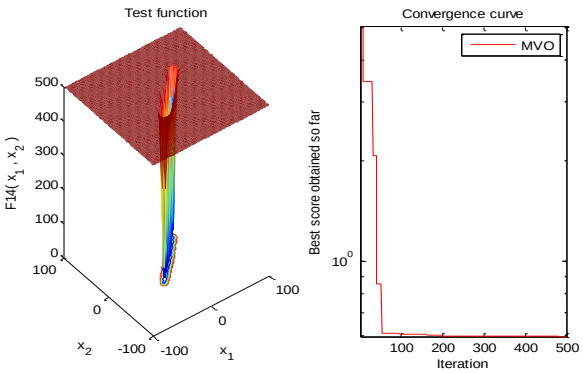

Fig 5.2 Convergence curve in case of MVO

V. Conclusion and Future Scope

Conclusion

The proposed work made use of optimization techniques in order to filter out the best share to be invested considering current value, cluster of 52 week high, 52 week low and average value. The optimization is helpful in prediction of best price among given data set. However the PSO is frequently used for simulating the optimized result. But during this research, it has been concluded that the performance of MVO is fast as compare to PSO. The data set of share has been taken from NSE/BSE site is preprocessed and then PSO and MVO are applied respectively to find the best solution. The filtering has been made considering optimized value. Moreover the research made comparison of performance in both cases.

Scope Of Research

Research would be used to study the clustering of big data in Finance. This work would be helpful to know the need of clustering along with challenges during implementation. Challenges during implementation of Kmean for big data must be considered in future work. There are different factors such as face value, market cap, promoters holding, fi holding, domestic holding, 52 week low, 52 week high, dividend, P/E that would be helpful to analyze the better result in future. Using this proposed module, it will be possible to determine the shares or funds that will provide maximum profit.

This research has played significant role in prediction of share price considering various factors. Such researches are beneficial for fund houses and investors who are paying a lot to trace the status of share to be bought. The research has provided flexible as well as scalable approach to predict the best value and filter out the suitable script. In future such mechanism could be helpful in case of crypto currency trading also where the value of currency fluctuates and investors need to know which currency should be Bought In Present Scenario.

VI. REFERENCES

[1]. T. Soni Madhulatha. AN OVERVIEW ON CLUSTERING METHODS. IOSR Journal of Engineering Apr. 2012, Vol. 2(4) pp: 719-725.

[2]. W.Sarada, Dr.P.V.Kumar, A REVIEW ON CLUSTERING TECHNIQUES AND THEIR COMPARISON , International Journal of Advanced Research in Computer Engineering &Technology (IJARCET) Volume 2 Issue 11, November 2013

[3]. Bhoj Raj Sharmaa.Clustering Algorithms: Study and Performance Evaluation Using Weka Tool.International Journal of Current Engineering and Technology ISSN 2277 - 4106 © 2013.

[4]. Shweta Srivastava. Clustering Techniques Analysis for Microarray Data. International Journal of Computer Science and Mobile Computing A Monthly Journal of Computer Science and Information Technology IJCSMC, Vol. 3, Issue. 5, May 2014

[5]. Muhammad Husain Zafar.A Clustering Based Study of Classification Algorithms.International Journal of Database Theory and Application Vol.8, No.1 (2015), pp.11-22.

[6]. Asir Antony Gnana Singh. Performance Analysis on Clustering Approaches for Gene ExpressionData.International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 2, February 2016 .

[7]. Kalyani M Raval. Data Mining Techniques

[8]. V.Ramesh, P.Parkavi, P.Yasodha. Performance Analysis of DM Techniques for Placement Chance Prediction .

[9]. Neelamadhab Padhy1, Dr.Pragnyaban Mishra .Survey of Data Mining Applications and Feature Scope .June 2012.

[10].David Jensen and Jennifer Neville. Data Mining in Social Networks.

[11].Aarti Sharma, Rahul Sharma,Vivek Kr. Sharma,VishalShrivatava. Application of Data Mining – A Survey Paper.2014.

[12].Kun-Ming Yu, Che-Yu Lin, Wen Ouyang Jiayi Zhou. An OpenCL Candidate Slicing Frequent Pattern Mining algorithm on graphic processing units. 2011 IEEE International Conference on Systems, Man, and Cybernetics.

[13].Che-Yu Lin, Kun-Ming Yu, Wen Ouyang Jiayi Zhou. An OpenCL Candidate Slicing Frequent Pattern Mining algorithm on graphic processing units. 2011 IEEE International Conference on Systems, Man, and Cybernetics.

[14].C. Vorapongkitipun and N. Nupairoj, “Improving performance of small-file accessing in Hadoop,” 2014 11th Int. Jt. Conf. Comput. Sci. Softw. Eng. "Human Factors Comput. Sci. Softw. Eng. - e-Science High Perform. Comput. eHPC, JCSSE 2014, pp. 200–205, 2014.

[15].C. Verma and R. Pandey, “Big Data representation for grade analysis through Hadoop framework,” Proc. 2016 6th Int. Conf. - Cloud Syst. Big Data Eng. Conflu. 2016, pp. 312–315, 2016.

[16].A. Siretskiy and O. Spjuth, “HTSeq-Hadoop: Extending HTSeq for massively parallel sequencing data analysis using Hadoop,” Proc. - 2014 IEEE 10th Int. Conf. eScience, eScience 2014, vol. 1, pp. 317–323, 2014

[17].K. Rattanaopas and S. Kaewkeeree, “Improving Hadoop MapReduce performance with data compression: A study using wordcount job,” ECTI-CON 2017 - 2017 14th Int. Conf. Electr. Eng. Comput. Telecommun. Inf. Technol., pp. 564–567, 2017.

[18].P. R. Merla and Y. Liang, “Data analysis using hadoop MapReduce environment,” Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017, vol. 2018–Janua, pp. 4783–4785, 2018.

[19].J. Kaur, K. Sachdeva, and G. Singh, “Image processing on multinode hadoop cluster,” 2017 Int. Conf. Electr. Electron. Commun. Comput. Optim. Tech., pp. 21–26, 2017.

[20].A. Bhardwaj, V. K. Singh, Vanraj, and Y. Narayan, “Analyzing BigData with Hadoop cluster in HDInsight azure Cloud,” 12th IEEE Int. Conf. Electron. Energy, Environ. Commun. Comput. Control (E3-C3), INDICON 2015, 2016.

[21].M. Bhandarkar, “MapReduce programming with apache Hadoop,” 2010 IEEE Int. Symp. Parallel Distrib. Process., p. 1, 2010.

[22].A. Q. Mohammed and R. Bharati, “An efficient technique to improve resources utilization for hadoop MapReduce in heterogeneous system,” ICCT 2017 - Int. Conf. Intell. Commun. Comput. Tech., vol. 2018–January, pp. 12–16, 2018.

[23].X. Qiao, D. Shi & F. Xu, "Optimal pricing strategy & economic effect of product sharing based on analysis of B2C sharing platform," 2019 16th International Conference on Service Systems & Service Management (ICSSSM), Shenzhen, China, 2019, pp. 1-6, doi: 10.1109/ICSSSM.2019.8887720.

[24].E. Turkedjiev, M. Angelova & K. Busawon, "Validation of Artificial Neural Network Model for Share Price UK Banking Sector Short-Term Trading," 2013 UKSim 15th International Conference on Computer Modelling & Simulation, Cambridge, 2013, pp. 6-11, doi: 10.1109/UKSim.2013.31.

[25].M. Wu & Q. Yu, "Empirical Analysis of Impact Factors of A Shares & H Shares of Price Differences," 2010 International Conference on E-Product E-Service & E-Entertainment, Henan, 2010, pp. 1-4, doi: 10.1109/ICEEE.2010.5661626.

[26].L. Guoyi & L. Renzhong, "An Analysis on Correlation of Factors Affecting Bank Share Prices," 2009 International Conference on Information Management, Innovation Management & Industrial Engineering, Xi'an, 2009, pp. 394-397, doi: 10.1109/ICIII.2009.404.

[27].X. Fang & T. Bai, "Share Price Prediction Using Wavelet Transform & Ant Colony Algorithm for Parameters Optimization in SVM," 2009 WRI Global Congress on Intelligent Systems, Xiamen, 2009, pp. 288-292, doi: 10.1109/GCIS.2009.85.

[28].F. Wang, L. Duan & J. Niu, "Optimal Pricing of User-Initiated Data-Plan Sharing in a Roaming Market," in IEEE Transactions on Wireless Communications, vol. 17, no. 9, pp. 5929-5944, Sept. 2018, doi: 10.1109/TWC.2018.2851578.

[29].N. Trivedi, P. Jangir, N. Jangir, S. A. Parmar, M. Bhoye & A. Kumar, "Voltage stability enhancement & voltage deviation minimization using multi-verse optimizer algorithm," 2016 International Conference on Circuit, Power & Computing Technologies (ICCPCT), Nagercoil, 2016, pp. 1-5, doi: 10.1109/ICCPCT.2016.7530136.

[30].M. Anbarasi, K. S. Sendhil Kumar, R. Balamurugan & Thejasswini, "Disease Prediction using Hybrid Optimization Methods based on Tuning Parameters," 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2020, pp. 643-648, doi: 10.1109/Confluence47617.2020.9058029.

Cite this article as :

Dr. Rajesh Gargi, Sheenu Sachdeva, Pooja Majoka, "An Implementations of Clustering Technique in Data Mining to Analyse Big Data in Finance", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 3, pp.448-454, May-June-2023.