iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 4 **Issue:** VI **Month of publication:** June 2016

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 4 Issue VI, June 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Bird Eye Review: Distributed Approach for Advancement of K Means Clustering Algorithm

Nishu Agrawal[1], S R Tandan[2]
[1]M.Tech Scholar, Computer science Department, Dr C V Raman University
[2]Assistant Professor, Computer science Department, Dr C V Raman University

Abstract— Clustering is the process of assigning similar objects in to one group, this group termed as clusters. To assign a similar data object in to one cluster, a well known K means clustering technique is used. In this paper we have review the various research papers to know the existing work done in the field of K means clustering data mining algorithm. We analyse that the performance of existing K means clustering algorithm for large data set is very inefficient. So to improve the performance of K means clustering algorithm, parallelization of K means clustering using multi core computation is the new area of interest for our research work.

Keywords— Clustering, Multi-core, K means clustering, parallelization

I. INTRODUCTION

Data mining is the latest technique that extracts or mine precious, knowledgeable efficient, effective, useful and meaningful data from large repository database. In data mining application variety of technique are available for data analysis such as association rule, classification, clustering. Clustering is one of the most researched areas for data analysis and has received much attention from data mining community. Clustering plays an important role in data mining applications such as scientific data exploration, information retrieval and text mining, web analysis, bioinformatics, and many others. In clustering data elements having same properties are placed in one group or cluster. So many clustering techniques are used. But the k means clustering method is more popular because it has ability to deal with large input dataset and it is very simple method.K means clustering is one of the most widely used clustering algorithm in which clusters are formed with the help of Centroids. It is partition based clustering method aimed to partition p data points in k cluster (p>k) in which each data points belong to cluster with nearest mean. The mean distance is calculated using Euclidean distance. The basic procedure of K means is shown in figure:

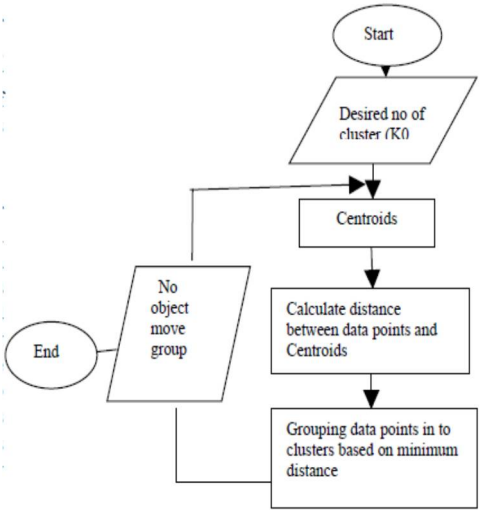


Figure 1: Procedure of k means clustering

But the computational time of conventional K means clustering method for large dataset is more, since the efficiency of conventional k means is very less. Our main focus in this research work is to propose a parallelization of well known K means clustering algorithm. This research work will improve efficiency of well known conventional K means clustering data mining algorithm by using parallel processing over large data set.

Volume 4 Issue VI, June 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

II. LITERATURE REVIEW

We have reviewed the existing work, done by researchers and presented in tabular form.

S.NO	TITLE	AUTHOR	PROPOSED WORK	FUTURE SCOPE/CONCLUSION	PUBLICATION	YEAR
1	Parallel K means clustering algorithm on NOWS	Sanpawat Kantabutra and Alva L. Couch	Suggested complexity and expensitivity of serial K means algorithm when applied to large dataset. Applying Parallel computing theory showed an improvement by factor of O (k/2) i.e. k no. of desired cluster.	Proposed algorithm will increase their speed if the communication time will reduce; this is achieved by grouping more than two subset on one machine.	Technical Journal	Jan-feb 2000
2	Parallel K means clustering algorithm on DNA dataset	Fazilah Othman,Rosni Abdullah,Nur Aini ,Abdul Rashid	Parallelized version of k means has been implemented based on inherent parallelism in Distance calculation and Centroids updates Distance calculations operate asynchronously and in parallel for all data points and node perform Centroids update parallely. Each participated node p handle n/p data points. Proposed algorithm applied on DNA dataset. Experimental result demonstrate proposed algorithm work well on large dataset	Work can be ported to high performance cluster of sun machine.	Springler-Verlag berlin Heideberg	2004
3	A Modified *k*-means Algorithm to Avoid Empty Clusters	Malay K. Pakhira	Presented a modified Version of the *k*-means algorithm, terms as m_k means algorithm. Modified version reduces the problem of empty cluster efficiently. Modify the center vector updation procedure of the basic k-means that denies the possibility of empty		International Journal of Recent Trends in Engineering	May 2009

			clusters. Proposed algorithm will maintain the performance of means as well as reduce the problem of empty cluster.			
4	An improved K-mean clustering algorithm	Zhiyi Fang, Chunfei Zhang	Advantage and disadvantage of traditional k- mean clustering algorithm is discussed. Provided improved K means clustering algorithm by enhancing the initial focal point and determination of k value. Simulation result *showed* that K means clustering has been improved as well as its vulnerability to noise is reduced.	Still further improvement of K means clustering will be required.	Journal of Information & Computational Science	2013
5	A hybrid approach to support the k means clustering method	T.hitendra Sharma P.viswanath and R.eswara reeddy	To speed up the k means clustering method author proposed prototype based hybrid based approach. First partition the dataset into small grouplet that represent by prototype. Further prototype set is partition into k cluster using modified k means, in iterative process the empty cluster eliminate Replace prototype with corresponding set of dataset partition.	Hybrid aprproach can be work with large dataset.	International Journal of Machine Learning and Cybernetics (IJMLC)	April 2013
6	Validation and verification of map reduce program model for parallel K means algorithm on Hadoop cluster	Amresh Kumar, Kiran M, Saikat Mukherjee	Design parallel K means based on map reduce program. Performance of map reduce application has been analysed with respect to the execution time and no. of nodes and dataset. The model of map reduce program, verified and validated that no. of node is inversely proportional to the	Further enhancement will increase the performance of Pk means algorithm by using different pre-processing steps.	International Journal of Computer Applications	May 2013

Volume 4 Issue VI, June 2016
ISSN: 2321-9653

			execution time. The result of pK means depend on the hadoop cluster size.			
7	A survey on clustering principles with k means clustering algorithm using different methods in detail	Manpreet kaur,usuir kaur	Discussed the execution time of k means clustering. Analyzed that traditional method take more time for execution. Author proposed a method of ranking and query redirection to reduce execution time. (i)Ranking method eliminates the occurrences of data item or the objects Hence optimize the result of means clustering. (ii) Using query redirection user find better result corresponding to queries in less execution time. Suggest two principle i.e. using query content and using documents click		International Journal of Computer Science and Mobile Computing	May 2013
8	A review of k means algorithm	Jyoti Yadav, Monika Sharma	Demonstrate the limitation of simple k means. Discussed the three dissimilar modified k means algorithm. i) Eliminate the requirement of input the value of k in advance. ii) Reduce complexity and eliminate dead unit problem. iii) used data structure that stored current iteration data that used in next iteration	Third algorithm(data structure) can be used in First algorithm to store information so time complexity reduce and result will be in optimal state.	International Journal of Engineering Trends and Technology (IJETT)	July 2013

9	Analysis And Study Of K-Means Clustering	Sudhir Singh and Nasib Singh Gill	Observed that actual K-mean algorithm takes lot of time when it is applied on a large database. That's why they proposed clustering concept to provide quick and efficient clustering technique on large data set. cluster no are generated automatically Proposed a possibility for finding global optimum and improve the problem of outliers. It is known that how many time the proposed k mean algorithm iterates Their experimental result demonstrated that scheme can improve the direct K-means algorithm. This paper also explains the time complexity of K-means.	Suggest the concept of Nearest Neighbour Clustering Algorithm to improve the compactness of clusters.	International Journal of Engineering Research & Technology (IJERT)	July 2013
10	Comparative study of k means algorithm by different measure	Kahkashan kouser,sunita	Applied k means algorithm on fisher IRIS dataset. Demonstrate three different distance function metrics such as Euclidian distance, Manhattan distance and	Conclude that the chebyshey distance accuracy is better than others but in chebyshey the no	International journal of and innovative research in computing and communicatio	Nov 2013

			chebyshey distance. to obtain result no of iterations, accuracy, mean absolute error. Euclidian distance is used in k means with its variant.	of iteration is more.	n technology	
11	Enhancing K-means Clustering Algorithm and Proposed Parallel K-means Clustering for Large Data set	Dr.Urmila R. Pol	Suggested parallel K means clustering using MPI. Assumed share nothing architecture and master slave processor model. Proposed original K means algorithm variation. No of cluster are fixed.i.e 3	This paper Conclude that Parallelization is best for data mining algorithm or large dataset.	International Journal of Advanced Research in Computer Science and Software Engineering	May 2014.
12	Performance based analysis between K means and fuzzy c mean clustering algorithm for connection oriented data	T.Velmurugan	Analysed K means algorithm and fuzzy c mean algorithm based on their computational time. K means being simple runs faster whereas fuzzy c mean is complex and hence run slower. Lesser execution time makes K means algorithm better than fuzzy c mean. K means algorithm has even distribution of data points but Fuzzy c mean has some variation in the distribution.	K means has vulnerable to noise.	elsevier.asc	June 2014
13	A survey on clustering algorithm and k means	Megha Mandloi	Studied all clustering algorithm. Compared k means algorithm with other clustering algorithm. Elaborate the advantage and disadvantage of different clustering method Partition based clustering handle small as well as large dataset,	Despite of all the demerits of means it is more popular because its simplicity to deal with large dataset. Two challenges in large scale clustering i.e. how to integrate	IJRTEM International Journal of Research in Engineering Technology and Management	July 2014

			Hierarchical clustering method handle small dataset. Density based clustering handle spatial data.	data feature and how to reduce computation cost Many improvement s will be done on k means to meet this challenges.		
14	Comparative analysis of parallel K means and fuzzy c mean cluster algorithm	Juby Mathew, Dr. R Vijaya Kumar	Compared the performance of the Parallel K means and parallel fuzzy c means clustering algorithms. Author use two metrics to compare the p performance : (i) Evaluation based on the execution time (ii) classification error percentage and efficiency.	Researcher concludes the computational time of parallel K means is better than the parallel FCM algorithm.	International Journal of Science, Engineering and Technology Research	Sep 2014
15	An Enhanced clustering algorithm by comparative study on K means algorithm	Sonia Guliani,Alpana Vijay Rajoriya	Proposed enhanced K mean clustering algorithm to reduce the error generated in existing K means algorithm. Enhanced algorithm applied on traffic dataset to find the cause of accident. Author compared the enhanced algorithm with existing algorithm. To check the error of algorithm real traffic dataset is used. Mean and standard deviation has calculate for each cluster, Accuracy of algorithm improved.	K means clustering can be used for mining high dimensionality dataset	International Journal of Engineering Research & Technology (IJERT)	June-2015

Volume 4 Issue VI, June 2016
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

III. MUTICORE PROCESSING

A multi core processor is a single computing element with more than one independent Processing units (PU) which is also called as "cores". The multi core can the program instructions simultaneously hence increasing the overall speed of computer programs. Nowadays all personal computers having more than one core .so the motivation behind our research work is to take the advantage of multi core. The more core there are more task can execute at the same time. We can improve the efficiency of conventional k means by applying multi core computation. The following figure shows that the system having four core i.e. CPU1, CPU2, CPU3, CPU4

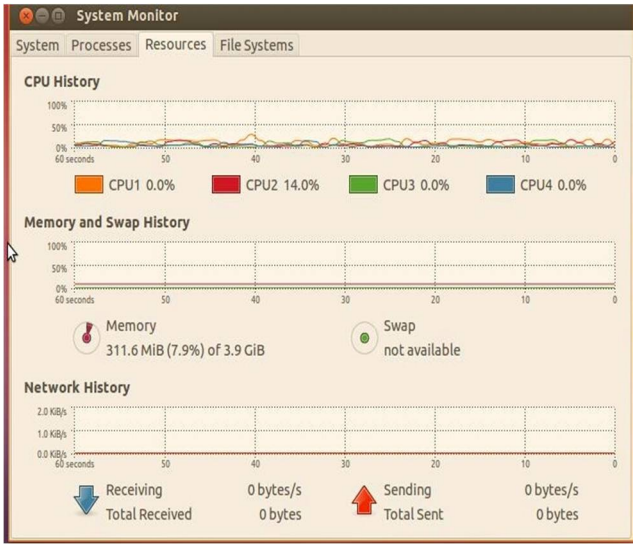


Fig 2: Four core processor system

IV. CONCLUSIONS

In this paper we have presented the previous work done in the field of k-means clustering algorithm and found that serial K means clustering algorithm is simple but it take much time to compute the distances between data points and cluster central points. We have found that still the performance of K means clustering for very large dataset is not so good. Fortunately, the distance computation of one data point is independent with the distance computation of other data points. Therefore, k-means clustering is a good candidate for parallelism. The proposed K means data mining algorithm will design to be used on large distributed data sets by using parallel processing. Hence we can improve the efficiency of conventional K-means clustering algorithm by applying multi-core computation over input datasets.

V. FUTURE SCOPE

On the basis of research review, we observed that in k- mean clustering algorithm requires desired number of cluster should be mention prior to the execution of algorithm. Conventional k-means algorithm is good enough for small dataset but computational time for large dataset is not good. Future possibility is to work in parallel k-means clustering algorithm to reduce the computational time for large dataset

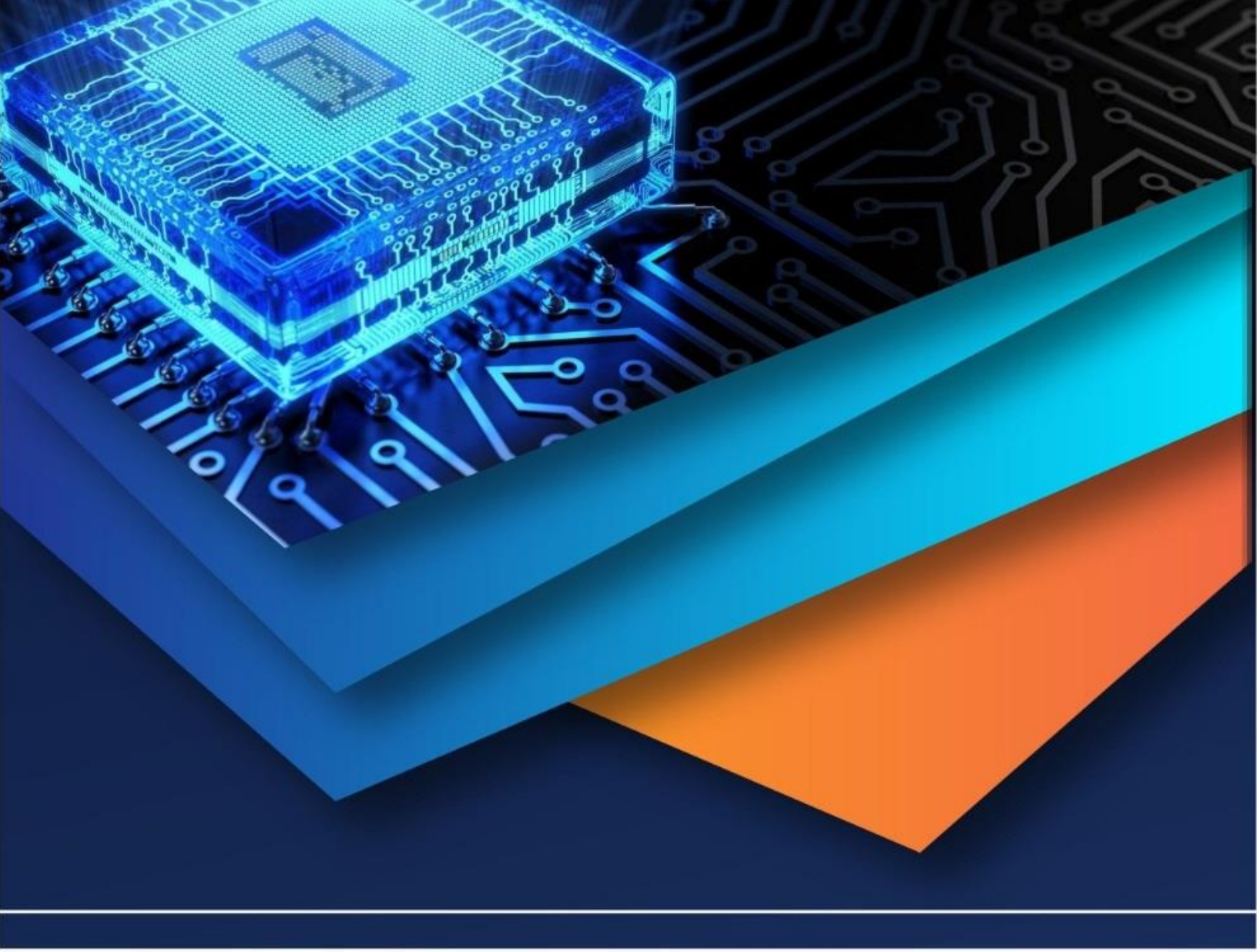
VI. ACKNOWLEDGMENT

I would like to thank my research guide MR. S R Tandan sir for his valuable support throughout the research work.

REFERENCES

[1] Amresh Kumar, Kiran M, Saikat Mukherjee " Verification and Validation of MapReduce Program model for Parallel K-Means algorithm on Hadoop Cluster"Volume 72– No.8, May 2013
[2] Dr.Urmila R. Pol Department Of Computer Science, Shivaji University, Kolhapur. "Enhancing K-means Clustering Algorithm and Proposed Parallel K-means Clustering for Large Data Sets" Volume 4, Issue 5, May 2014.
[3] Fazilah Othman,Rosni Abdullah,Nur Aini ,Abdul Rashid "Parallel K means clustering algorithm on DNA dataset" Springler-Verlag berlin Heideberg 2004
[4] Gopi Gandhi,rohit shrivastav "A comparative study on partitioning techniques on clustering algorithm" ,international journal of computer applications vol 87 feb 2014
[5] Jyoti yadav, Monika Sharma "A review of k means algorithm" International Journal of Engineering Trends and Technology (IJETT) – Volume 4 Issue 7- July 2013Technology Vol. 1, Issue 9, November 2013
[6] Kahkashan Kouser, sunita "comparative study of k means algorithm by different measure '' International journal of innovative research in computing and communication engineering volume 1 issue 9 nov 2013
[7] Malay K. Pakhira"A Modified k-means Algorithm to Avoid Empty Clusters" Vol 1, No. 1, May 2009

[8] Sanpawat Kantabutra and Alva L. Couch Department of Computer Science "Parallel K-means Clustering Algorithm on NOWs"Volume1, Issue 6,Jan-Feb 2000
[9] Manpreet kaur,usuir kaur "A survey on clustering principles with k means clustering algorithm using different methods in detail" International Journal of Computer Science and Mobile Computing Vol. 2, Issue. 5, May 2013
[10] Megha mandloi "A survey on clustering algorithm and k means" International Journal of Research in Engineering Technology and Management Volume: 02 Issue: 04 July 2014
[11] Megha gupta ,vishal shrivastav "Review of various technique in clustering "International journal of advanced computer research June 2013
[12] Raed T. Aldahdooh, Wesam Ashour "DIMK-means ―Distance-based Initialization Method for K-means Clustering Algorithm" January 2013
[13] T. Hitendra Sarma , P. Viswanath ,B. Eswara Reddy "A hybrid approach to speed-up the k-means clustering method" April 2013. International Journal of Machine Learning and Cybernetics (IJMLC)
[14] T. Velmurugan"Performance based analysis between k-Means and Fuzzy C-Means clustering algorithms for connection oriented telecommunication data"June 2014
[15] Zhiyi Fang, Chunfei Zhang "An Improved K-means Clustering Algorithm" 2013
[16] Jiawei Han and Micheline kamber, "Data mining concept and Techniques"Second edition
[17] Sudhir Singh and Nasib Singh Gill "Analysis and Study Of K-Means Clustering Algorithm" Vol. 2 Issue 7, July – 2013
[18] Sonia Guliani,Alpana Vijay Rajoriya "An Enhanced Clustering Algorithm by Comparative Study on K-Means Algorithm"Vol. 4 Issue 06, June-2015
[19] K. A. Abdul Nazeer, M. P. Sebastian "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm", Proceedings of the World Congress on Engineering 2009 ,Vol. WCE 2009, July 1 - 3, 2009, London, U.K.
[20] Data mining techniques by A K PUJARI
[21] Juby Mathew, Dr. R Vijayakumar "comparative analysis of parallel k means and parallel fuzzy c means cluster algorithm" International Journal of Science, Engineering and Technology Research September 2014
[22] Jageshwer Shriwas, Shagufta Farzana," Using Text Mining and Ruled Based Technique For Prediction of Stock Market Price" International Journal of Emerging Technology and Advanced Engineering (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 1, January 2014
[23] Jageshwer Shriwas, Shagufta Farzana," Prediction of Stock Market Price Using Classification Rules and Graph Based Analysis", CIIT Data Mining and Knowledge Engineering, Vol 5, No 7 (2013)
[24] Jageshwer Shriwas, Dr Samidha Dwivedi Sharma," Stock Price Prediction Using Hybrid Approach of Rule Based Algorithm and Financial News" Int.J.Computer Technology & Applications,Vol 5 (1),205-211,jab-feb 2104.
[25] Jageshwer Shriwas1, Dr. Samidha Dwivedi Sharma" New Trends for Stock Market Prediction using NIPS", International Journal of Emerging Technology and Advanced Engineering (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 4, April 2014)
[26] S Pandey, R Miri, S R Tandan "Diagnosis and Classification of Hypothyroid Disease Using Data Mining Techniques" International Journal of Engineering Research and Technology, (2013)
[27] P Gupta, S R Tandan, R Miri "Decision Tree Applied For Detecting Intrusion "International Journal of Engineering Research and Technology (IJERT) (2013)
[28] Khushboo Sharma, S R Tandan "An Optimized Parallel Confidence Measures Algorithm on Web Log Data" International Journal of Engineering Research and Technology (IJERT) (2013)
[29] Asha Miri, S.R.Tandan, Rohit Miri "Pseudo Code to Eliminate Unwanted Data Sets for Fuzzy Mining Association Rule" International Journal for Research in Applied Science & Engineering Technology (IJRASET) (2015).
[30] Rohit Miri, Priyanka Tripathi, S R Tandan" Exploration of Novel Algorithm for Reduced Computational Time by Using Fuzzy Classification Technique in Data Mining" International Journal for Research in Applied Science & Engineering Technology (IJRASET) (2015
[31] Rohit Miri, Priyanka Tripathi, S R Tandan "Novel Algorithm For Finding The Range Of Fuzzy Values For Quantitative Data Sets By Data Mining And Fuzzy Technique" Journal Of Advanced Database Management & Systems Journal Of Advanced Database Management & Systems, (2015)
[32] S R Tandan, Rohit Miri, ,Dr Priyanka Tripathi "A Bird Eye Review on Reduced Time Complexity by Using Data Mining and Fuzzy Techniques" international journal for research in applied science and engineering technology (ijraset), (2014)
[33] S R Tandan Rohit Miri, Priyanka Tripathi "TRApriori Classification Based Algorithm by Fuzzy Techniques to Reduced Time Complexity" International Journal of Computer Science & Information Technology, International Journal of Computer Science & Information Technology, (2014)
[34] Rohit Miri, Priyanka Tripathi, Keshri Verma, S.R. Tandan "Novel Algorithm to Reduced Computational Data Sets for Fuzzy Association Rule" International Journal For Research In Applied Science And Engineering Technology (IJRASET), (2015).
[35] R Miri, P Tripathi, S R Tandan "Novel Algorithm for Reduced Computational Data by Using Fuzzy Classification and Data Mining Techniques" Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies" ACM, (2014)
[36] R Miri, P Tripathi, S R Tandan " Neuro-Fuzzy Based Integrated and Optimized Search Engine for Effective and Reliable Information Retrieval System" Proceedings of the 2014 International Conference on Information and Communication Technology for Competitive Strategies" ACM, (2014)



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY