International Journal of Innovative Research in Technology & Science

ISSN: 2321-1156 Volume VI Issue VI, November 2018

IMPROVING ACCURACY OF BIG DATASET BASED SBWG METHOD AND KMCM

Adish Jain, M. Tech. Scholar, Department of EC, NRI, RGPV, Bhopal, India; adijain08@gmail.com;
Prof. Puran Gour, Asst. Professor, Department of EC, NRI RGPV, Bhopal, India; Purangour@rediffmail.com

Abstract

The main goal of the data mining process is to extract useful information from big data set and transform it into an understandable form for further use. It was not possible to extract useful information from the large datasets or data streams. Now this can be achieved by the capability of big data mining. The overlay based parallel data mining architecture executes processing by utilize the overlay network and fully distributed data management, which can achieve high scalability and service availability. Clustering is the method of grouping the information into different categories so that objects or information in one clusters are highly similar and dissimilar with object or information in other clusters. In the research different approaches separated data in form of the clustering and clustering are divided into different categories. Separated and mix information of objects into group those forms larger clusters and so on. In divided clustering different portion are created based on some criteria that are compared with well-known K-Mean algorithm given better accuracy. In this research in order to localization of point values to data sets are taken from well-known UCI machine learning repository. Experiments supported the quality information UCI show that the projected technique will turn out a high purity cluster results and eliminate the sensitivity. It is existing thing human being and totally different point values assign and minimize error values. During the implementation both existing scheme and k-means clustering method (KMCM) are applied on the datasets and try to find which algorithm provides good accuracy as compare existing method (SBWG). To provide the ability to make sense and maximize utilization of such vast amounts of web data for knowledge discovery .KMCM has proved to be more efficient in terms of quality and optimal result. In this experiment, successfully gets highest accuracy result to train dataset.

Keywords: Data Mining Technique, Cluster Technique, SOM, K-Mean Cluster Method, Accuracy Analysis, Dataset, Big Dataset, Redundancy Analysis.

I. INTRODUCTION

Big data technologies defines a new generation of technologies and architectures, designed solely to economically extract useful information's from very large volumes of a wide variety of data, by permitting high velocity capture, discovery, and analysis [1].Data mining permits us to extract information from our historical information and predict outcomes of our future things. Cluster is a vital data processing task. It will be described because the method of organizing objects into teams whose members are similar in a way. cluster also can be outline because the method of grouping the info into categories or clusters, so objects inside a cluster have high similarity compared to 1 another however are terribly dissimilar to things in alternative clusters. Principally cluster will be done by two strategies, hierarchical and Partitioning methodology. The field of information mining and knowledge discovery is rising as a replacement, elementary analysis space with necessary applications to Science, engineering, medicine, business, and education. Data processing makes an attempt to formulate analyze and implement basic induction processes that facilitate the extraction of significant data and information from unstructured knowledge. Size of databases in scientific and business application is large wherever the amount of records in an exceedingly dataset will vary from thousands of millions. Clustering could also be outlined as an information reduction tool i.e. used to produce subgroups that are a lot of and a lot of manageable than individual data point. Basically, agglomeration is justified as a method used for grouping a large variety of knowledge into significant teams or clusters supported similarity types of objects data. Clusters are the teams that have knowledge similar on basis of common options and dissimilar to knowledge in different clusters. Data mining is that the process of extracting hidden, previously unknown and helpful data from giant knowledge bases and data warehouses. Data mining process involve steps like knowledge cleansing, integration, selection, transformation, data processing technique, and pattern analysis and information illustration. various data processing techniques are used like classification, clustering, association rules, successive patterns, Prediction, call trees, etc. are used in various applications. Here we tend to discuss regarding agglomeration algorithms like fuzzy c-means, k-means. Agglomeration is that the method of organizing knowledge objects into a collection of disjoint categories known as clusters. Clustering is an example of classification. Classification refers to a procedure that assigns knowledge objects to a collection of categories. Unattended means that clustering doesn't depend on predefined categories and training examples whereas classifying the information objects. Cluster analysis seeks to partition a given knowledge set into teams supported such that options so the information points among a bunch are a lot of kind of like one another than the points in numerous teams. Therefore, a cluster could be a collection of objects that are similar among themselves and dissimilar to the objects belonging to different clusters. Clustering is a crucial space of analysis that finds applications in several fields together with bioinformatics, pattern recognition, image process, marketing, data processing, economics, etc. Cluster analysis could be a one in every of the first knowledge analysis tool within the data processing. Clustering algorithms are primarily divided into two categories, hierarchical algorithms and Partition algorithms. A hierarchical clustering algorithm divides the given knowledge set into smaller subsets in hierarchical fashion. A partition clustering algorithm partition the information set into desired variety of sets in a very single step. Cluster analysis is one of the most important problems for data mining and machine learning. It is based on the idea of like attracts like for the samples that to be classified. Clustering has widely used in many fields such as data mining and knowledge discovery, pattern recognition, decision support, machine learning and image segmentation [2]

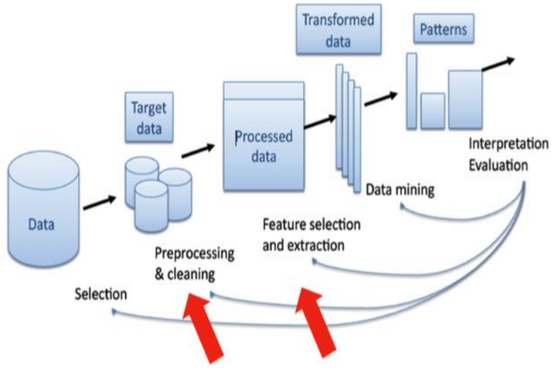

Figure 1 Knowledge discovery Process

Data Mining: Data Mining (or information Discovery in Databases) describes the big idea of finding "interesting" patterns in large collections of information. There's a large quantity of data available within the data business. This knowledge is of no use till its regenerate into helpful information. It's necessary to research this large quantity of data and extract helpful information from it. Extraction of knowledge isn't the only method we need to perform; data mining so describes the abstract goals of what must be done, and depends upon a large range of various techniques to achieve them, like artificial neural networks, cluster analysis different processes such as data cleaning, data Integration, information Transformation, data mining, Pattern analysis and knowledge Presentation. In this processes are overcome problem data error, they might be able to use this data in several applications like error detection in large data set, market dataset analysis research, Production control, Science Exploration, etc. Clustering could also be defined as an information reduction tool i.e. used to produce subgroups that are a lot of and a lot of manageable than individual data point. Basically, clustering is justified as a method used for grouping a large variety of knowledge into significant teams or clusters supported similarity types of objects data. Clusters area unit the teams that have knowledge similar on basis of common options and dissimilar to knowledge in different clusters Cluster analysis teams knowledge objects primarily based solely on data found in knowledge that describes the objects and their relationships. The objects among a bunch are kind of like one another each different and totally different from the objects in other teams. Cluster analysis is one in every of the key data processing techniques, wide used for several sensible applications in varied rising areas like Bioinformatics. clustering is an unsupervised methodology that subdivides associate input file set into a desired variety of subgroups so the objects of a similar subgroup are going to be similar (or related) to at least one another and totally different from (or unrelated to) the objects in different teams [3].

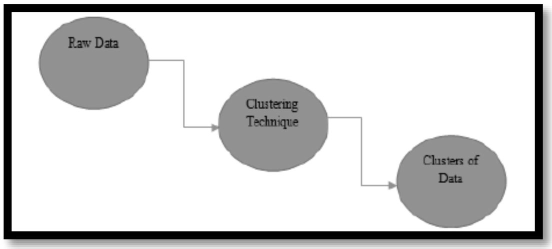

Figure2 Stages of Clustering

II. Literature Survey

The section describe about previous related work under the data mining.

R. Amutha et al. [8] discuss that once two or a lot of algorithms of same class of cluster technique is used then best results are going to be no heritable. During this paper, varied cluster algorithms are mentioned. 2 k-means algorithms discuss here are: Parallel k/h-Means cluster for giant knowledge Sets and a unique K-Means primarily based cluster formula for top Dimensional data Sets. Parallel k/h-Means formula is meant to deal with terribly giant knowledge sets. The applying results of this formula has been proved with ninetieth potency on a distributed computing setting. These results show that this formula is scalable. Novel K-Means primarily based Clustering provides the benefits of exploitation each HC and K-Means. Using these two algorithms, house and similarity between the info sets present every nodes is extended.

Sudipto Guha et al. [9] planned a replacement hierarchical cluster algorithmic rule referred to as CURE that's stronger to outliers, and identifies clusters having non-spherical shapes and wide variances in size. This can be achieved in CURE method by representing every cluster by a particular mounted variety of points that are generated by choosing well scattered points from the cluster so shrinking them toward the middle of the cluster by a specified fraction. To handle giant databases, CURE employs a mixture of sampling and partitioning. Beside the outline of CURE algorithmic rule, the author also delineated, form of options it uses, and why it uses totally different techniques.

Twinkle Garg et al. [10]. "Survey on varied increased K-Means Algorithms". Data mining is defined as a way used to extract and mine the invisible, meaningful info from mountain of knowledge. Clustering is a very important technique that has been introduced within the space of knowledge mining. Clustering is defined as a technique wont to cluster similar information into a group of clusters supported some common characteristics', K-means is one amongst the popular partitioned based mostly clustering algorithms within the space of analysis. The impact issue of k-means is its simplicity, high potency and scalability. However, is additionally includes of range of limitations: random choice of initial centroids, range of cluster K need to be initialized and influence by outliers. Visible of those deficiencies, this paper presents a survey of enhancements done to ancient k-means to handle such limitations.

Shafeeq et al. [11] present a changed K-means algorithmic program to enhance the cluster quality and to repair the best variety of cluster. As input variety of clusters (K) given to the K-means algorithmic program by the user. However within the sensible state of affairs, it's terribly difficult to repair the quantity of clusters before. The strategy planned during this paper works for each the cases i.e. for famous variety of clusters before additionally as unknown variety of clusters. The user has the pliability either to repair the variety| of clusters or input the minimum number of clusters needed. The new cluster centers are computed by the algorithmic program by incrementing the cluster counter by one in every iteration till it satisfies the validity of cluster quality. This algorithmic program can overcome this downside by finding the best variety of clusters on the run. The planned approach takes additional machine time than the K-means for larger knowledge sets. It's the foremost disadvantage of this approach.

Kalpana D. Joshi et al. [12]. "Modified K-Means for higher Initial Cluster Centres". The k-means cluster algorithmic program is most popularly used in processing for planet applications. The potency and performance of the k-means algorithmic program is greatly affected by initial cluster centers as utterly completely different initial cluster centers usually cause different cluster. Throughout this paper, we have a tendency to tend to propose a modified k-means algorithmic program that has additional steps for selecting higher cluster centers. We tend to tend to cipher Min and easy lay distance for every cluster and notice high density objects for selection of upper k.

Libao ZHANGet al. [13]propose a simple and qualitative methodology using k means clustering algorithm to classify NBA guards and used the Euclidean distance as a measure of similarity distance. This work display by using k-Means clustering algorithm and120 NBA guardsí data. Manual classification of traditional methods is improved using this model. According to the existing statistical data, the NBA players are classified to make the classification and evaluation objectively and scientifically. This work show that this is very effective and reasonable methodology. Therefore, based on classification result the guardsí type can be defined properly. Meanwhile, the guardsí function in the team can be evaluated in a fair and objective manner.

Xiao-Feng Xie et al. [14]. "Adaptive Particle Swarm improvement on Individual Level". Found out an adaptive particle swarm optimization (PSO) on individual level. By analyzing the social model of PSO, an exchange criterion supported the variety of fitness between current particles and also the best historical expertise is introduced to keep up the social attribution of swarm adaptively by removing inactive particles. Functions were tested that indicates its improvement within the average performance. An adaptive particle swarm optimization (PSO) on individual level is conferred. By analyzing the social model of PSO, an exchange criterion supported the variety of fitness between current particles and also the best historical expertise is introduced to keep up the social attribution of swarm adaptively by coming out inactive particles. The testing of 3 benchmark functions indicates it improves the typical performance effectively.

Ran Vijay Singh et al. [15] present a changed k-means algorithmic rule supported the sensitivity of initial center of clusters. During this algorithmic rule whole area is partitioned off into completely different segments. Subsequently frequency of information points in every section is calculated. The most likelihood of information points to contain the center of mass of cluster is within the section that shows the most frequency. If information points {of completely different of various} sections have same highest frequency and therefore the boundary of section crosses the brink ëkí it's necessary to merge different segments so take the very best k segment for calculative the initial center of mass of clusters. A threshold distance is additionally outlined for every cluster is center of mass to match the gap between information points and clusters center of mass. This work shows that changed k-means algorithmic rule can decrease the quality and energy of numerical calculation, maintaining the easiness of implementing the k-means algorithmic rule.

D. Virmani et al. [16]. Normalization based K means bunch Algorithm. K-means is an economical cluster technique used to separate similar info into groups supported initial centroids of clusters. Throughout this paper, group action based K-means bunch algorithmic rule (N-K means) is planned. Planned N-K means bunch algorithmic rule applies group action before bunch on the accessible info what is more as a result of the planned approach calculates initial centroids supported weights. Experimental results prove the betterment of planned N-K means bunch algorithmic rule over existing K-means bunch algorithmic rule in terms of complexity and overall performance.

Marjan Kuchakist et al. [17] offers a summary of some specific ranked cluster algorithmic rule. Firstly, author classified bunch algorithms, so the most centered was on ranked bunch algorithms. One in every of the most functions of describing these algorithms was to reduce disk I/O operations, consequently reducing time quality. They need conjointly declared attributes, disadvantages and benefits of all the thought-about algorithms. Finally, comparison between all of them was done in keeping with their similarity and distinction.

J. James Manoharan et al. [18]. “Outlier Detection using increased K-Means clustering algorithm and Weight based mostly Center Approach”. In data processing there square measure innumerable ways are used to find the outlier by creating the clusters of knowledge then sight the outlier from them. Generally bunch technique plays a really necessary role in data processing. Clustering means that grouping the similar information objects along supported the characteristic they possess. Outlier Detection is a very important issue in information mining; significantly it's been wont to determine and eliminate abnormal information objects from given information set wherever outlier is that the information item whose price falls outside the bounds within the sample information could indicate abnormal information.

During this work we've urged a clustering primarily based mostly outlier detection algorithm for effective data processing that uses increased k-means clustering algorithm to cluster the info sets and weight based center approach. In planned approach, 2 techniques are combined to expeditiously notice the outlier from the info set. Threshold price are often calculated programmatically by taking absolute quantity of minimum and most value of a specific cluster. The experimental results demonstrate that increased technique takes least computational time and concentrates on reducing the outlier that would improve efficiency of k-means bunch for achieving the higher quality clusters.

III SIMULATION AND RESULT ANALYSIS

MALAB TOOL: MAT-LAB is a software package for high performance numerical computation and visualization. It provides an interactive environment with hundreds of built-in function for technical computation, graphics and animations. The name MAT-LAB stands for Matrix Laboratory. One of most feature of MAT-LAB is its platform independence. Once you are in MATLAB, for the most part, it does not matter which computer you are on. In MAT-LAB the M-files are the standard ASCII text files, with an .m extension to the file name. There are two files of this file: script file and function file. All most programs in write in MAT-LAB are saved in M-files. Fig-files are binary files with a .fig extension that can be opened again in MAT-LAB as figures. Such files are created by saving a figure in this format using save or save as option from File menu or using the save as command in command window-files are compiled M-files with a .p extension that can be executed in MAT-LAB directly. There are several optional toolboxes are available from developers of MAT-LAB.

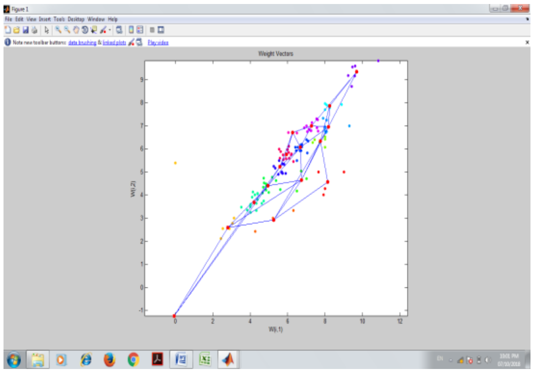
Figure 1
File Edit View Insert Tools Desktop Window Help
Note new toolbar buttons: data brushing & linked plots Play video
Weight Vectors
W(i,2)
W(i,1)
-1
0
1
2
3
4
5
6
7
8
9
0
2
4
6
8
10
12

Figure 3 SOM based weight graph

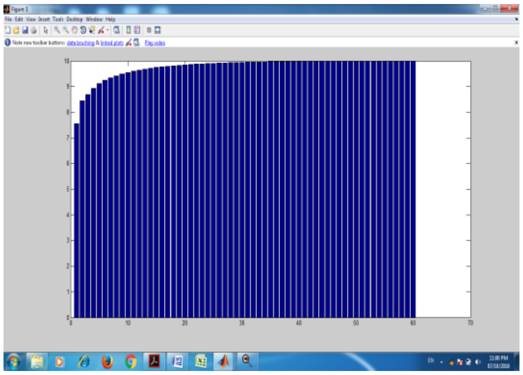
Figure 3
File Edit View Insert Tools Desktop Window Help
Note new toolbar buttons: data brushing & linked plots Play video
0
1
2
3
4
5
6
7
8
9
10
0
10
20
30
40
50
60
70

Figure 4 SOM based weight graph values analysis graph

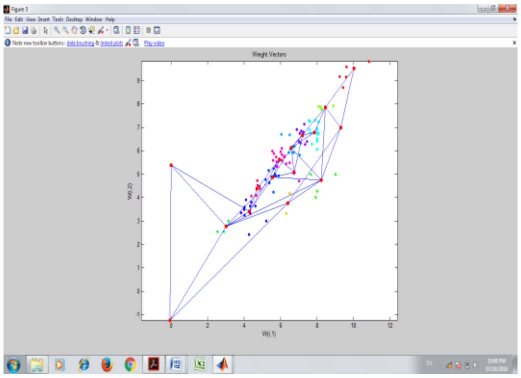

Figure 5 KMCM based weight graph

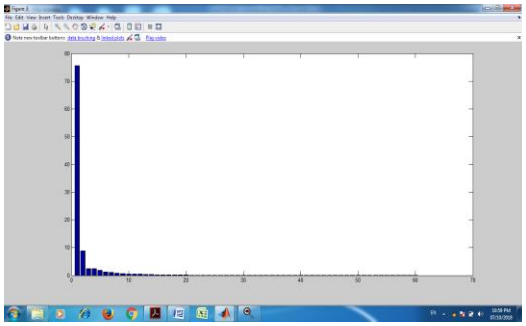

Figure 6 KMCM based weight graph values analysis graph

Proposed Approach: In recent research biomedical and health informatics system are based on online web services is very helpful to web users. Because system provide intelligently knowledge extract from social media and provide improve healthcare to the users with cost effective manner. In our approach, we will take the exemplary data of lung cancer from different web sites. The data contain post word, comments, user rank, treatment and side effect of treatment. That input data value converts into token, after the token creation we inference sentiments of token i.e. positive and negative sentiments, the sentiments passed into the SBWG and KMCM analyze word frequency data that derived from user forum posts, and inference the treatments accuracy and side effect percentage. After the result inference from the SOM modelling, we tune the related post and treatments as well as side effect of treatments for better care of health against lung cancer. Our new proposed algorithm is better as compare SBWG .because PA (KMCM) is minimize flat in dataset and improve accuracy of retrieval data. Large dataset using healthcare.

IV. RESULT ANALYSIS

SBWG: Cancer data analysis using old method SOM based weight graph (SBWG) method to represent weight graph in show below and more error and accuracy less. Different from homogeneous information in database. Analysis and generate a set of weight graph P = p1, p2 …..p_n with corresponding weight vector w = w1,w2…. w_m. SOM based weight graph (SBWG) method as described and show figure 3. Time analysis finds using our proposed method (KMCM) minimum time but old method (SBWG) time more and accuracy analysis finds using our proposed method (KMCM) accuracy more but old method (SBWG) accuracy minimum.

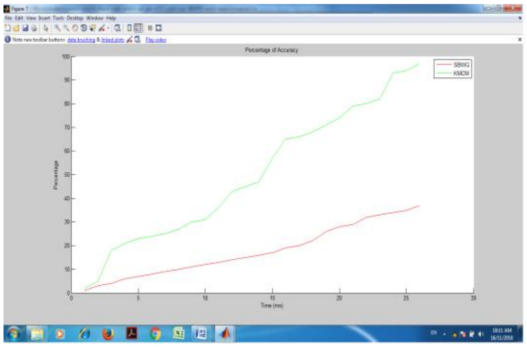

Figure 7 Accuracy analysis between KMCM and SBWG

VII. Conclusion

In real-world applications managing and mining huge dataset is difficult task, because the information concern large in an exceedingly volume, distributed and decentralized management and complex. There are many challenges at information, model and system level. They have computing platform to handle this huge information. Framework is one in every of the foremost necessary elements of huge processing, typical parallel data processing design won't offer data processing services just in case of network disruption. Because of Increase within the quantity of information within the field of genetics, meteorology, biology, environmental analysis, it becomes tough to handle the information, to search out Associations, patterns and to research the big data set. The main goal of data mining method is to extract helpful information from huge dataset .To simulate an improved innovative and best performance analysis supported medical dataset used data processing techniques to k-means cluster method for increase the performance utilization medical dataset. Proposed techniques are performance analysis dataset and find optimal result. Medical dataset analysis using old method SOM based weight graph method and KMCM. In organize that proposed clustering algorithm with higher accuracy is optimal performance analysis with totally different groups. Here proposed clustering (KMCM) is better as compare to SBWG. KMCM is more accuracy supported base on minimizing redundancy in dataset and minimize fault. The results produced were satisfactory in terms of good accuracy. The performance analysis based on most of clustering algorithm has been compared and analyzed with a number of the existing evolutionary clustering algorithms however this has proved to be more efficient in terms of quality and optimal result. In this experiment, successfully gets highest accuracy show in result. It can be found better result and validation data, so much satisfactory result.

REFERENCES

[1] Xindong Wu, Fellow, IEEE, Xingquan Zhu, Gong-Qing Wu, and Wei Ding" Data Mining with Big Data" IEEE Transactions on Knowledge and Data Engineering, VOL. 26, NO. 1, JANUARY 2014.

[2] S Ray and R H Turi: Determination of number of clusters in K-means clustering and application in color image segmentation, (invited paper) in N R Pal, A K De and J Das (eds), Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques(ICAPRDT'99), Calcutta, India, 27-29 December, 1999.

[3] Kaufman, L., Rousseeuw, P.J., Finding Groups in Data. An Introduction to Cluster Analysis. Wiley, Canada. 1990.

[4] Redmond.S.J, Heneghan. C, A method for initializing the k-means clustering algorithm using kd-trees. Pattern Recognition Lett. 28, 2007, 965–973.

[5] J.A.S. Almeida, L.M.S. Barbosa, A.A.C.C.Pais and S.J.Formosinho (2007), Improving Hierarchical Cluster Analysis:A new method with outlier detection and automatic clustering, Chemo metrics and Intelligent Laboratory Systems, 87, pp. 208-

[6] Tian Zhang, Raghu Ramakrishnan, MironLinvy (1996), BIRCH: an efficient data clustering method for large databases, International Conference on Management of Data, In Proc. of 1996 ACM-SIGMOD Montreal, Quebec.

[7] Kapil Joshi, Himanshu Gupta, Prashant Chaudhary, Punit Sharma, "Survey on Different Enhanced K-Means Clustering Algorithm", International Journal of Engineering Trends and Technology, Volume 27 Number 4 - September 2015.

[8] R. Amutha, Renuka. K,îDifferent Data Mining Techniques And Clustering Algorithmsî, International Journal Of Technology Enhancements And Emerging Engineering Research, VOL 3, ISSUE 11, ISSN 2347-4289.

[9] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim (1998), CURE: An Efficient Clustering Algorithm For Large

[10] Twinkle Garg, Arun Malik, "Survey on Various Enhanced K-Means Algorithms", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 11, November 2014.

[11] Shafeeq,A., Hareesha,K.,ìDynamic Clustering of Data with Modified K-Means Algorithmî International Conference on Information and Computer Networks, vol. 27 ,2012

[12] Kalpana D. Joshi, P.S. Nalwade, "Modified K-Means for Better Initial Cluster Centres", IJCSMC, Vol. 2, Issue. 7, pg.219 – 223, July 2013.

[13] Libao ZHANG, Faming LU, An LIU, Pingping GUO, Cong LIU,îApplication of K-Means Clustering Algorithm for Classification of NBA Guardsî, International Journal of Science and Engineering Applications Volume 5 Issue 1, 2016, ISSN-2319-7560 (Online).

[14] Xiao-Feng Xie, Wen-Jun Zhang, and Zhi-Lian Yang, "Adaptive Particle Swarm Optimization on Individual Level," IEEE, International Conference on Signal Processing (ICSP), Beijing, China, pp. 1215-1218, 2002.

[15] Ran Vijay Singh, M.P.S Bhatia, ìData Clustering with Modified K-means Algorithmî, Recent Trends in Information Technology, 2011 IEEE International Conference on 3-5 June 2011(pp. 717- 721.

[16] Deepali Virmani,Shweta Taneja,Geetika Malhotra, "Normalization based K means Clustering Algorithm", International Journal of Advanced Engineering Research and Science, Vol-2, Issue-2, ISSN: 2349-6495 ,Feb.- 2015.

[17] MarjanKuchaki Rafsanjani, Zahra Asghari Varzaneh, Nasibeh Emami Chukanlo (2012), A survey of hierarchical clustering algorithms, The Journal of Mathematics and Computer Science, 5,.3, pp.229- 240.

[18] J. James Manoharan, Dr. S. Hari Ganesh, Ph.D., Dr. J.G.R. Sathiaseelan, "Outlier Detection Using Enhanced K-Means Clustering Algorithm And Weight Based Center Approach", IJCSMC, Vol. 5, Issue. 4, pg.453 – 464, April 2016.

[19] Deepali Virmani, Shweta Taneja, Geetika Malhotra, "Normalization based K means Clustering Algorithm", International Journal of Advanced Engineering Research and Science, Vol-2, Issue-2, ISSN: 2349-6495, Feb. - 2015

[20]. Marjan Kuchaki Rafsanjani, Zahra Asghari Varzaneh, Nasibeh Emami Chukanlo (2012), A survey of hierarchical clustering algorithms, The Journal of Mathematics and Computer Science, 5,.3, pp.229- 240.

[21]. J. James Manoharan, Dr. S. Hari Ganesh, Ph.D., Dr. J.G.R. Sathiaseelan, "Outlier Detection Using Enhanced K-Means Clustering Algorithm And Weight Based Center Approach", IJCSMC, Vol. 5, Issue. 4, pg.453 – 464, April 2016.