Scientific Journal of Impact Factor(SJIF): 3.134
e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 1,Issue 11, November -2014

ANALYSIS OF HIERARCHICAL CLUSTERING ALGORITHM TO HANDLE LARGE DATASET

Mandani Kashmira[1], Prof.Hemani Shah [2]

[1] *Department of Computer Engineering, Student of PG studies-MEF Group of Institutions, onlykashu007@gmail.com*
[2] *Department of Computer Engineering, Faculty of PG studies-MEF Group of Institutions, hemani.shah@marwadieducation.edu.in*

Abstract — *Clustering, in Data Mining is useful for discovering groups and identifying interesting distributions in underlying data. Traditional data clustering algorithms either favor clusters with special shapes and similar sizes, or are very delicate in the presence of outliers. Nowadays most widely studied problem is identification of clusters in a large dataset. Hierarchical Clustering is the process of forming a maximal collection of subsets of objects (called clusters), with the property that any two clusters are either disjoint or nested. Hierarchical clustering combine data objects into clusters, those clusters into larger clusters, and so forth, creates a hierarchy of clusters, which may represent a tree structure called a dendrogram. Agglomerative clustering is a most flexible method and it is also used for clustering the large dataset, there is no need of the number of clusters as an input. In this paper we have introduced solution for decreasing time complexity of clustering algorithms by combining approaches of two different algorithms from which one is good in accuracy and other is fast that is helpful for information retrieval from large data.*

Keywords- Clustering, Agglomerative Clustering, Chameleon, Cure, Frequent Pattern Mining

I. INTRODUCTION

In data mining, hierarchical clustering works by grouping data objects into a tree of cluster. Hierarchical clustering methods can be further classified into agglomerative and divisive hierarchical clustering. This classification depends on whether the hierarchical decomposition is formed in a bottom-up or top-down fashion. Hierarchical techniques produce a nested sequence of partitions, with a single, all inclusive cluster at the top and singleton clusters of individual objects at the bottom [2]. Each intermediate level can be viewed as combining two clusters from the next lower level or splitting a cluster from the next higher level. The result of a hierarchical clustering algorithm can be graphically displayed as tree, called a dendrogram [4]. This tree graphically displays the merging process and the intermediate clusters. This graphical structure shows how points can be merged into a single cluster. Hierarchical methods suffer from the fact that once we have performed either merge or split step, it can never be undone [3]. This inflexibility is useful in that it leads to smaller computation costs by not having to worry about a combinatorial number of different choices. However, such techniques cannot correct mistaken decisions that once have taken. There are two approaches that can help in improving the quality of hierarchical clustering: (1) Firstly to perform careful analysis of object linkages at each hierarchical partitioning or (2) By integrating hierarchical agglomeration and other approaches by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro-clusters using another clustering method such as iterative relocation. One promising direction for improving the clustering quality of hierarchical methods is to integrate hierarchical clustering with other clustering techniques for multiple phase clustering. So in order to make improvement in hierarchical clustering we merge some other techniques or method in to it [1]. Every day, 2.5 quintillion bytes of data are created and 90 percent of the data in the world today were produced within the past two years. Our capability for data generation has never been so powerful and enormous ever since the invention of the information technology in the early 19th century [6]. So the growth of the data and its synchronization formats becomes very complex to get some fruitful knowledge for them.

II. RELATED WORK

2.1 Type of Clustering

A collection of similar data objects is known as clustering and clustering have two types of similarities [11].

2.1.1 **Intra-class similarity**- Objects are similar to objects in same cluster. Intra-class similarity is based on the near data objects in a cluster [15]. In a similarity measured in clustering is an important role in doing good clustering so intra-class similarity gives the large value for good clustering. it is measured by using this equation:

$$ICS(\pi(D)) = \sum_{i=1}^{|\pi(D)|} \frac{1}{|C_i|^2} \sum_{d_{id}' \epsilon C_i} Similarity(d, d')$$

- Interclass dissimilarity- Objects are dissimilar to objects in other clusters. Interclass similarity gives less value for good clustering [15]. So similarity between different cluster object is less. This formula is used to measure similarity of cluster objects in this method.

$$ECS(\pi(D)) = \sum_{i=1}^{|\pi(D)|} \sum_{j=i+1}^{|\pi(D)|} \frac{1}{|C_i|\,|C_j|} \sum_{d \epsilon C_i, d' \epsilon C_i} Similarity(d, d')$$

2.2 Various Types of Clustering Method

Clustering is a process of classify data into number of cluster. Clustering classifies data based on different types of methods that methods are as follows:

- Hierarchical clustering method
- partitioning clustering method
- Density based clustering method
- Grid based clustering method
- Model based clustering method

2.2.1 Partitioning clustering

In partitioning method have N object database and that database is partitioned in k groups using partitioning method [11]. All objects contain in one cluster and at least one object contain in each group. This method is suited for small to medium sized data set to find spherical-shaped clusters. It is used for complex data set and cluster very large data set. The representative partitioning clustering algorithms are K-MEDOID, K-MEANS [11]. K-means clustering algorithm is working in that manner such as first select the randomly one node from the number of objects. K is a center of cluster. Similar data form a cluster, similarity find based on the distance between object and center [7].

2.2.2 Model based clustering

Model based clustering method constructs the model for every clusters and find a data which is fit to that model and this method is automatically give the number of clusters. This method is robust. The representative model-based clustering algorithm is EM. Model based method is often based on probability distribution of data. Individual distribution is called component distribution. In this method, probability distribution is done by the mixture density model. EM method acquires statistics from traditional mixture model and depends on those statistics; it performs clustering in model based clustering method.

2.2.3 Density-based clustering

In density based method divide data in cluster based on the density of objects. So distance between cluster objects is less and number of objects is growing. So density of cluster is growing. And it has same advantages such as reduced effect of noise (outliers) and discovers cluster of arbitrary shape, input data scan only once, needs density parameters to be initialized [12]. Here Density-based clustering algorithms the data space contains dense regions of objects is, considered as a cluster and clusters are separated by regions of low density. Density based algorithms depends on each object with a density value defined by the number of its neighbor objects within a given radius. Density of objects is greater than a specified threshold is defined as a dense object and initially is formed a cluster itself [12]. Two clusters are merged if the y shares a common neighbor that is also dense. The DBSCAN, OPTICS, HOP, and DENCLUE algorithms are representative of density-based clustering algorithms concept [12].

2.2.4 Grid based clustering

In grid based clustering method, data is divided into data space in number of cell that forms grid structure. Grid structure depends on the number of cell rather than number of object. Perform clustering in a grid so complexity is reduced in grid based clustering method. Statistic attribute are gathered form grid cell. Performance depends on the size of the grid that is less than the number of objects contained in a cluster [13]. The representative grid based clustering algorithms are STING, WAVE CLUSTER, and CLINQUE [13].

2.2.5 Hierarchical clustering

Hierarchical clustering builds a cluster hierarchy such as a tree of clusters. Hierarchical clustering processes is a sequence of divide or merge clusters. In which each cluster have chilled and structure that is more informative than the unstructured set of clusters returned by flat clustering. No need to give number of cluster initially. Good result visualization and complexity is high. Hierarchical clustering is used in information retrieval [12]. In which distance of objects is measured and merge or divide cluster objects based on the distance. Unstructured data is divided or merged effectively in hierarchical clustering.

2.2.5.1 Divisive (Top-Down) Approach

Here starting with a one cluster of all objects and recursively splitting each cluster until the termination criterion is reached [12]. The most useful part of hierarchical clustering is that it can be applicable to any type of attributes, so it is easy to apply for any task related to web usage mining with respect to web data.

2.2.5.2 Agglomerative Technique (Bottom-Up)

Hierarchical and partition are clustering methods, in the partitioning method it is required the number of clusters as a input while in hierarchical clustering method there are no need to number of cluster as a input, so unknown data set can be given as a input. Hierarchical clustering contains two methods top-down and bottom-up. Agglomerative clustering is a bottom-up method [7]. That method is simple and very flexible method. Agglomerative clustering algorithm works by grouping the data one by one on the basis of the nearest distance measure of all the pair wise distance between the data point. Again distance between the data points is recalculated but which distance is to be considered is decided when the groups has been formed [7]. Single linkage, complete linkage, average linkage and centroid distance between two points, grouping the data until one cluster is remaining. Agglomerative clustering starting with one point clusters and recursively merging two or more most similar clusters to a single cluster (parent) until the termination criterion is reached [12]. (E.g. k – clusters have been built). It has advantages that No apriori information about the number of clusters required and Easy to implement and gives best result in some cases. Here Algorithm can never undo what was done previously. And sometimes it is difficult to identify the correct number of clusters by the dendogram. Major drawbacks are that initial error propagation, dimensionality, complexity and large data set size.

- **ALGORITHM**

Agglomerative clustering algorithm steps:

1) Let be the set of data points.
2) Then find distance between the clusters.
3) Merge pair of clusters that have smallest distance.
4) Update distance matrix.
5) If all the data points are in one cluster then stop, else repeat from step 2) [7].

2.3 Various technique of agglomerative clustering

- **single linkage**

Single linkage clustering is one of the methods of agglomerative hierarchical clustering. In single linkage clustering link between two clusters is made by single object pair, it is also known as a nearest neighbor clustering. Distance between two clusters is based on points of clusters that is small or nearest. Mathematically, the linkage function is D(X, Y) – the distance between clusters X and Y – is described by the equation

$$D(X,Y) = \min_{x \epsilon X, y \epsilon Y} d(x,y)$$

Where X and Y are any two sets of objects considered as clusters and D(x, y) function denotes the distance between the two elements x and y [10].

- **Complete linkage**

In complete-linkage clustering, the link between two clusters contains all element pairs, it is also known as a farthest neighbor clustering. Distance between two clusters depends on objects of clusters that is maximum or farthest. Mathematically, the complete linkage function—the distance D(x, y) between clusters X and Y— is described by the following expression [10]:

$$D(X,Y) = \max_{x \epsilon X, y \epsilon Y} d(x,y)$$

Where d(x,y) is the distance between elements and x and y are two sets of elements (clusters)

- **Average linkage**

In average linkage clustering the link between two clusters contains one point of cluster to all points of other cluster [10]. In which average distance between pairs of clusters data points denote the distance of two clusters.

$$Sim(C_i, C_j) = \frac{1}{|C_i \cup C_j|(|C_i \cup C_j - 1)} \sum_{\vec{x} \in (C_i \cup C_j)} \sum_{\vec{y} \in (C_i \cup C_j), \vec{y} \neq \vec{x}} Sim(\vec{x}, \vec{y})$$

- **Centroid linkage**

In centroid linkage clustering, the link between two clusters contains one point center of cluster to center points of other cluster. The distance between clusters is defined as the (squared) Euclidean Distance between cluster centroids [10].

2.4 Issues in agglomerative clustering

In general agglomerative clustering have issues such as dimensionality, initial error propagation, complexity and large data set size issues.

- **Large data set size**

Today's data is growing so storage of data is expanded. And data set size is increased day by day. Therefore clustering of data is needed because of all data or information is not important for doing any operation. And all data is classified in attributes. But in agglomerative clustering algorithm have issues of not classifying large data set [16]. In which number of data or information of attribute is large but it is not important that number of attribute is more. So it is independent about how many attribute contain in a data base.

Agglomerative clustering algorithm is performing on small or medium size data set. This method has limitation that is not work on large data set size, so large data set size issue is a major issue in agglomerative clustering algorithm. Large data set size problem can be depend upon three reasons such as 1) Data set can have large number of elements 2) In data set contain each element can have number of features. 3) Many clusters can be contain in data set for discover all the data [14].

- **High Dimensionality**

Agglomerative clustering algorithms are designed for two or three dimension data. So high dimension data is a major challenge for clustering because of dimensionality is increased [17]. Small number of dimension is relevant to the exiting clusters. And many dimensions are irrelevant. Therefore noise in data is increased. When dimensionality is increased data becomes parsed because data points are located on different dimension subspaces [17].

Dimensionality depends on the number of attribute contain in data set instead of large number of features of each attribute [17].

- **Complexity**

Complexity is an amount of time or space required by an algorithm for given input size. Agglomerative clustering algorithm complexity is O (n3) because the similarities for N x N metrics scan every time and find similarity N-1 iterations and given N number of clusters as an input [9]. So complexity of agglomerative clustering algorithm is high [9].

And in which storage space is needed for similarity metrics so complexity is high in that method.

- **Initial error propagation**

In an agglomerative clustering method error is contained in an initial step of the clustering and that is propagated in last step of the process of clustering [2]. In agglomerative clustering algorithm, to merge number of clusters in one cluster at that time one cluster contain error and that error is not solved that step and that cluster is merged with other cluster have data is a without error or true data, so data of that two cluster may be clash, So error can be occurred in that data is more [2].

In agglomerative clustering algorithm all that types of cluster is merge and error propagation can be more so result of that method is a not accurate at the final step of algorithm [2].So initial error propagation is a problem of agglomerative clustering algorithm.

III. BRIEF ON LARGE DATA

Large Data is notable not because of its size, but because of its relationship to other data. Due to efforts to mine and aggregate data, Large Data is fundamentally networked (threaded with connections). Its value comes from the patterns that can be derived by making connections between pieces of data, about an individual, about individuals in relation to others, about groups of people, or simply about the structure of information itself [7]. Besides this, large dataset has enormous volume, high velocity, much variety and variation. These features of Large Dataset present the main challenges in analyzing Large dataset which are: (1)Efficient and effective handling of large data, (2)Processing time and accuracy of results trade –off; and(3) Filtering important and relevant data from all the data collected[4] [3].

3.1 The main challenges identified for the IT Professionals in handling large dataset are:

- The designing of such systems which would be able to handle such large amount of data efficiently and effectively.
- The second challenge is to filter the most important data from all the data collected by the organization. In other words we can say adding value to the business [3].

3.2 Clustering Challenges of Large Dataset [3]

Clustering in Large data is required to identify the existing patterns which are not clear in first glance.

The properties of large data pose some challenge against adopting traditional clustering methods:

- **Type of dataset**:

The collected data in the real world often contain both numeric and categorical attributes. Clustering algorithms work effectively either on purely numeric data or on purely categorical data; most of them perform poorly on mixed categorical and numerical data types.

- **Size of dataset**:

The size of the dataset has a major effect on the clustering quality. Some clustering methods are more efficient clustering methods than others when the data size is small, and vice versa.

- **Handling outliers/ noisy data:**

A successful algorithm will often be able to handle outlier/noisy data because of the fact that the data in most of the real applications are not pure. Also, noise makes it difficult for an algorithm to cluster an object into a suitable cluster. This therefore affects the results provided by the algorithm.

- **Time Complexity:**

Most of the clustering methods must be repeated several times to improve the clustering quality. Therefore if the process takes too long, then it can become impractical for applications that handle large dataset.

- **Stability:**

One of the important features for any clustering algorithm is the ability to generate the same partition of the data irrespective of the order in which the patterns are presented to the algorithm.

- **High dimensionality:**

As the number of dimensions increases, the data become increasingly sparse, so the distance measurement between pairs of points becomes meaningless and the average density of points anywhere in the data is likely to be low. Mathematically, nearest neighbor query becomes unstable.

- **Cluster shape**: A good clustering algorithm should be able to handle real data and their wide variety of data types, which will produce clusters of arbitrary shape.

IV. SUGGESTED SOLUTION

Here before suggesting one solution for the above mention problems, let us define two algorithms that are birch and cure. Before defining birch and CURE, let us go through chameleon and rock briefly.

4.1 Chameleon Algorithm:

CHAMELEON operates on a sparse graph in which nodes represent data items, and weighted edges represent similarities among the data items. This sparse graph representation of the data set allows CHAMELEON to scale to large data sets and to operate successfully on data sets that are available only in similarity space and not in metric spaces . CHAMELEON finds the clusters in the data set by using a two phase algorithm. During the first phase, CHAMELEON uses a graph partitioning algorithm to cluster the data items into a large number of relatively small sub-clusters. During the second phase, it uses an agglomerative hierarchical clustering algorithm to find the genuine clusters by repeatedly combining together these sub-clusters. The key feature of CHAMELEON's agglomerative hierarchical clustering algorithm is that it determines the pair of most similar sub-clusters by taking into account both the inter-connectivity as well as the closeness of the clusters; and thus it overcomes the limitations discussed in Section 3 that result from using only one of them. Furthermore, CHAMELEON uses a novel approach to model the degree of inter-connectivity and closeness between each pair of clusters that takes into account the internal characteristics of the clusters themselves. Thus, it does not depend on a static user supplied model, and can automatically adapt to the internal characteristics of the clusters being merged.

- **Steps for Chameleon algorithm:**

Step 1: Take any data set as an input
Step 2: Construct a sparse graph
Step 3: Use K-nearest neighbor graph for the sparse graph
Step 4: Now partition that sparse graph
Step 5: After portioning merge the partitions
Step 6: consider merged sections as final clusters

- **Advantages:**

It does not depend on user supplied information but it automatically adapts to the internal characteristics of the clusters being merged.
It is more accurate than rock.
It is capable to handle large data.
Cluster shape can be arbitrary.

- **Disadvantages:**

It cannot handle noisy data.

4.2 Rock Algorithm:

ROCK (RObust Clustering using linKs) is a clustering algorithm for data with categorical and Boolean attributes. It redefines the distances between points to be the number of shared neighbors whose strength is greater than a given threshold and then uses a hierarchical clustering scheme to cluster the data. ROCK deals primarily with market basket data. Traditional Euclidean distance measures are not appropriate for such data and instead, ROCK uses the Jaccard coefficient (section 2.2.3.3) to measure similarity. This rules out clustering approaches such as K-means or Centroid based hierarchical clustering.

- **Steps for ROCK:**

Step 1: Obtain a sample of points from the data set.
Step 2: Compute the link value for each set of points, i.e., transform the original similarities (computed by the Jaccard coefficient) into similarities that reflect the number of shared neighbors between points.
Step 3: Perform an agglomerative hierarchical clustering on the data using the "number of shared neighbors" similarities and the "maximize the shared neighbors" objective function defined above.
Step 4: Assign the remaining points to the clusters that have been found.

- **Advantages of ROCK**:

The shape of clustering can be arbitrary.

- **Disadvantages:**

It cannot handle high dimensional data.
It cannot handle noisy data.
Dataset must be categorical or numerical only.
It takes more time compared to other hierarchical algorithms.

4.3 Birch Algorithm:

BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies) is an integrated agglomerative hierarchical clustering method. It is mainly designed for clustering large amount of metric data. A CF tree is a height-balanced tree that stores the clustering features for a hierarchical clustering. It is similar to B+-Tree or R-Tree. CF tree is balanced tree with a branching factor (maximum number of children per non leaf node) B and threshold T. Each internal node contains a CF triple for each of its children. Each leaf node also represents a cluster and contains a CF entry for each sub cluster in it. A sub cluster in a leaf node must have a diameter no greater than a given threshold value. An object is inserted to the closest leaf entry (sub cluster). A leaf node represents a cluster made up of all sub clusters represented by its entries. All the entries in a leaf node must satisfy the threshold requirements with respect to the threshold value T, that is, the diameter of the sub cluster must be less than T. If the diameter of the sub cluster stored in the leaf node after insertion is larger than the threshold value, then the leaf node and other nodes are split. After the insertion of the new object, information about it is passed toward the root of the tree. The size of the CF tree can be changed by modifying the threshold. These structures help the clustering method achieve good speed and scalability in large databases. Birch is also effective for incremental and dynamic clustering of incoming objects.

- **Steps for Birch Algorithm:**

Data phase:
Step 1: take data in the form of CF tree

Work with CF tree:
Step 2: specify range by constructing a smaller CF tree

Idea of Smaller CF Tree:
Step 3: Apply global clustering

Cluster tuning
Step 4: cluster refining again

Step5: better results by repeating process on whole data set

- **Merits & Demerits of BIRCH:**

It is local in that each clustering decision is made without scanning all data points and currently existing clusters.
It exploits the observation that data space is not usually uniformly occupied and not every data point is equally important.
It makes full use of available memory to derive the finest possible sub-clusters while minimizing I/O costs.
It is also an incremental method that does not require the whole dataset in advance.

- **Demerits:**

Data order sensitivity i.e. data must be specify in a specific order.
It is unable to deal with non-spherical clusters of varying size because it uses the concept of diameter to control the boundary of a cluster.

4.4 CURE Algorithm:

Cure employs a novel hierarchical approach that adopts a middle ground between the centroid based and all-point extremes. In CURE a constant no of c of well scattered points in a cluster are chosen first. The scattered points capture the shape and extent of the cluster. The chosen scattered points are next shrunk towards the centroid of the cluster by fraction alpha. These scattered points after shrinking are used as representative of a cluster. The cluster with a closest pair of representative points are the clusters that merged at all steps of CURE. CURE utilizes multiple reprehensive points for each cluster that are generated by selecting well scattered Points from the cluster and shrinking them toward the center of the cluster by a specified fraction.

- **Steps for CURE Algorithm:**

Step 1: Get random data from the input
Step 2: Partition that sample according to dimensions
Step 3: partially cluster partition
Step 4: Eliminate Outliers
Step 5: cluster partial cluster
Step 6: cluster labeling

- **Merits & Demerits of CURE:**

CURE can detect cluster with non spherical shape and wide variance in size using a set of representative points of each cluster.
CURE can also be good in execution time in presence of large database using random sampling and partitioning methods.
CURE works well when dataset contain outliers. They are detected and eliminated.

- **Demerits:**

Consider only one point as representative of a cluster
Good only for convex shaped, similar size and density, and if *k* can be reasonably estimated.

V. PROPOSED WORK

Before planning to a new approach let us compare existing approaches. The simulation has been made on the platform of the weka APIs and Eclipse IDE. And the results are as below. The results are derived as time taken and output clusters and number of data points.

Algorithm	Selling	Ionosphere	Voting	Medical
Data point	210	350	500	750
Chameleon	0.03	0.19	0.18	0.66
Rock	0.12	1.69	0.87	2.83
Birch	0.06	0.48	0.63	2.19
CURE	0.06	0.37	0.39	1.84

Table .1 comparison based on time and data points on various datasets

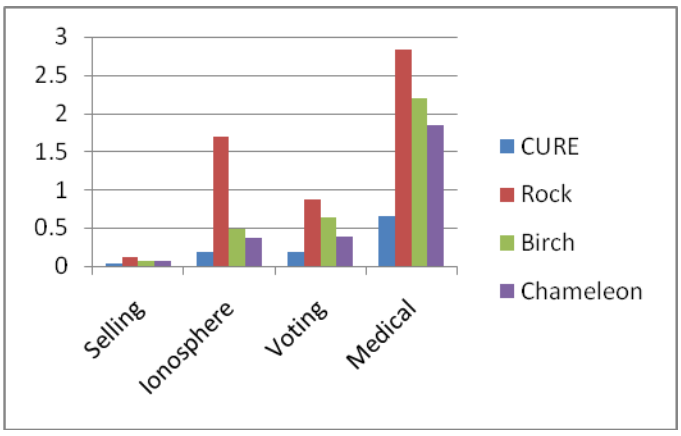

Fig. 1 Graph based on various datasets and various Algorithms vs. Time

By comparing from the graph, here is proposed a new approach that is combination of advantages of birch and CURE algorithm.

VI. CONCLUSION

After considering a lot of research underneath, the conclusion is that Hierarchical Clustering algorithm is the most flexible method to handle large data sets. Chameleon algorithm is best among the above because it work on the graph structure which required less memory to store the data. The rest of the algorithm is work on the tree based structure which requires more space. Chameleon algorithm and sparse graph construction makes the clustering process accurate

and CURE algorithm and partition approach makes the clustering process easy and fast as small parts are easy to handle with. So, over all it will be good algorithm for clustering with aspect of accuracy and time parameters.

REFERENCES

[1] A. Fahad, N. Alshatri, Z. Tari, Member, IEEE , A. Alamri, I. Khalil A. Zomaya, Fellow, IEEE, S. Foufou, and A. Bouras," A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis", IEEE Transactions on Emerging Topics in Computing,Volume:PP 1-12,12 June 2014,ISSN :1268-6750

[2] Tian Zhang, Raghu Ramakrishnan, MironLivny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases," SIGMOD '96, pp.103-114, 1996

[3] Yogita Rani, Manju, Harish Rohil, " Comparative Analysis of BIRCH and CURE Hierarchical Clustering Algorithm using WEKA 3.6.9", The Standard International Journals (The SIJ), 2014, ISSN: 2312 – 2381

[4] D.Pramodh Krishna, A.Senguttuvan&T.SwarnaLatha, "Clustering on Large Numeric Data Sets Using Hierarchical Approach: Birch", Volume 12 Issue 12 Version 1.0 Year 2012, Global Journals Inc. (USA), Online ISSN: 0975-4172 & Print ISSN: 0975-4350

[5] Sadaf Khan, Rohit Singh, "Cup – Clustering Using Priority: An Approximate Algorithm For Clustering Big Data", International Conference on Computer Science and Mechanical Engineering, 10th August 2014, Jaipur, India, ISBN: 978-93-84209-42-1

[6] SudiptoGuha, Rajeev Rastogi, Kyuseok Shim, "CURE: An efficient clustering algorithm for large dataset", Stanford university publication, 2011.

[7]Xindong Wu, XingquanZhu,Gong-Qing Wu And Wei Ding, " Data Mining With Big Data", Ieee Transactions On Knowledge And Data Engineering, Vol. 26, No. 1, January 2014

[8] DR. A. N. Nandakumar, NanditaYambem,"A Survey on Data Mining Algorithms on Apache HadoopPlatform"",International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 1, January 2014

[9]Aditya B. Patel, Manashvi Birla, Ushma Nair, "Addressing Big Data Problem Using Hadoop and Map Reduce", NIRMA UNIVERSITY INTERNATIONAL CONFERENCE ON ENGINEERING, NUiCONE-2012,.

[10] JamesManyika, Brad Brown et.al, "Big Data: The next frontier for Innovation, Competition, and Productivity," McKinsey Global Institute, June 2011.

[11] Murat Ali , Ismail HakkiToroslu, "Smart Miner: A New Framework for mining Large Scale Web Usage Data" WWW 2009, April 20-24. 2009 Madrid, Spain.ACM 978-1-60558-487-4/09/04.

[12]B. Huberman, "Sociology of Science: Big Data Deserve a Bigger Audience" Nature, vol. 482, Pp. 308-311, 2012.

[13] A. Labrinidis and H. Jagadish, "Challenges and Opportunities with Big Data" Proc. VLDB Endowment, vol. 5, no. 12, Pp. 693-699,2012.

[14] J.A.S. Almeida, L.M.S. Barbosa, A.A.C.C. Pais& S.J. Formosinho, "Improving Hierarchical Cluster Analysis: A New Method with Outlier Detection and Automatic Clustering", *Chemometrics and Intelligent Laboratory Systems*, Vol. 87, Pp. 208–127.

[15] L. V. Bijuraj," Clustering and its Applications" Conference on New Horizons in IT-NCNHIT 2013

[16] Aasthajoshi, RajneetKaur," A Review: Comparative Study of Various Clustering Techniques in Data Mining", Volume 3, Issue 3, March 2013

[17] M.Vijayalakshmi, M.Renuka Devi, "A Survey o f Different Issue of Different clustering Algorithms Used in Large Datasets" , Volume 2, Issue 3, March 2012

