I J C T A, 9(26) 2016, pp. 315-323

A Novel Algorithm Using Birch with *K*-Means Clustering Using Dataset in R Data Mining Tool

D. Osmond Niranjan Giftson[1], R. Jemina Priyadarshini[2] and L. Arockiam[3]

ABSTRACT

Clustering will be in the form of classification, in that it creates a classification of items with class labels. Clustering is a way of grouping or combining data objects into disjoint clusters. The clustering results directly depend upon the method of clustering algorithm that is applied to the dataset. This research paper proposes a novel hybrid method which combines the features of *K*-means clustering algorithm with BIRCH (a hierarchical clustering algorithm) in R Data Mining Tool. The proposed algorithm firstly generates a tree using hierarchical clustering algorithm which gives a large number of clusters when applied to a dataset and then clustering has been performed using *K*-means partitioning algorithm which reduces the number of clusters and with more accuracy, and less sum of square error. The proposed algorithm is applied on car dataset which is then compared with *K*-means and *K*-Medoid clustering algorithm in *R* Data Mining Tool. The comparison is done on the basis of number of iterations and sum of square error (Intra cluster similarity), in which the new algorithm performs better as compare to *K*-Means and *K*-medoid clustering algorithms using dataset.

Keywords: K-means clustering algorithm, BIRCH algorithm, sum square error, Hierarchical, Partitioning, birch Object.

1. INTRODUCTION

Data mining refers to

> *"using a variety of techniques to identify nuggets of information or decision-making knowledge in bodies of data, and extracting these in such a way that they can be put to use in the areas such as decision support, prediction, forecasting and estimation. The data is often voluminous, but as it stands of low value as no direct use can be made of it; it is the hidden information in the data that is useful"*

Data mining is concerned with the analysis of data and the use of software techniques for finding patterns and regularities in sets of data. It is the computer which is responsible for finding the patterns by identifying the underlying rules and features in the data.

2. LITERATURE SURVEY

Now a day's short-tempered growth of data collection, so data is stored in data warehouses and that data is retrieved by intranet or internet. Data mining is a process of excerpt or saves use full info from the large set of data. Today number of user is cumulative so use of web is rising exponentially; World Wide Web is a wide source of data. Web mining is a data mining technique that is used to routinely excerpt information from web. web mining is separated in three type such as data content mining (content of pages), structure mining

[1] Research Scholar, Department of Computer Science, Bishop Heber College, Tiruchirappalli. *E-mail: niranjand26@gmail.com*

[2] Associate Professor, Department of Computer Science, Bishop Heber College, Tiruchirappalli. *E-mail: jemititis@gmail.com*

[3] Associate Professor, Department of Computer Science, St. Joseph's College, Tiruchirappalli. *E-mail: larockiam@yahoo.co.in*

(structure of pages), usage mining (access or use of pages). Web usage mining is to discover browsing design from user's performances [1]. Web usage mining is helps to deal with sure web scaling problem such as user trend analysis through surfing, traffic flow analysis, spread control and behavior, web traffic management and many more [2].

Clustering can be in the form of classification, in that it creates a classification of items with class (cluster) labels. Classification means a controlled classification; *i.e.*, new, unlabeled objects are allocated a class label using established objects with known class labels. The term division and dividing are sometimes used as substitutes for clustering. The dividing is often used in connection with techniques that division graphs into sub graphs and that are not strongly related to clustering. Division often refers to the separation of data into group using simple techniques; *e.g.*, an image can be split into sections based only on pixel strength and color, or people can be divided into groups based on their income. Clustering is a process of identifying the similar data objects in large data sources [3]. The association between objects is signified in an immediacy matrix, in which rows and columns agree to items. The proximity matrix is the one and only input to a clustering algorithm.

3. EXISTING ALGORITHMS

3.1 *K*-Means Clustering Algorithm

K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\|x_i - v_j\|)^2$$

where,

'$\|x_i - v_j\|$' is the Euclidean distance between x_i and v_j.

'c_i' is the number of data points in i^{th} cluster.

'c' is the number of cluster centers.

3.2 BIRCH Clustering Algorithm

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) algorithm is an integrated hierarchical clustering algorithm. It uses the clustering features (Clustering Feature, CF) and cluster feature tree (CF Tree) two concepts for the general cluster description. Clustering feature tree outlines the clustering of useful information, and space is much smaller than the meta-data collection can be stored in memory, which

can improve the algorithm in clustering large data sets on the speed and scalability and is very suitable for handling discrete and continuous attribute data clustering problem. In the BIRCH tree a node is called a Clustering Feature. It is a small representation of an underlying cluster of one or many points. BIRCH builds on the idea that points that are close enough should always be considered as a group. Clustering Features provide this level of abstraction. Clustering Features are stored as a vector of three values: CF = (N;LS; SS). The linear sum (LS), the square sum (SS), and the number of points it encloses (N). All of these metrics can be calculated using only basic math:

$$LS = \sum_{P_i \in N} \vec{P}_i$$

$$SS = \sum_{P_i \in N} \left|\vec{P}_i\right|^2$$

If divided by the number of points in the cluster the linear sum marks the centroid of the cluster. As the formulas suggest both of these values can be computed iteratively. Any Clustering Feature in the tree can be calculated by adding its child Clustering Features:

$$CF_1 + CF_2 = (N_1 + N_2, LS_1 + LS_2, SS_1 + SS_2).$$

A CF tree is a height balanced tree that has two parameters namely, a branching factor, *B*, and threshold, *T*. The representation of a non-leaf node can be stated as {*CFi*, childi}, where, *i* = 1, 2, ..., *B*, Childi : A pointer to its child node. *CFi* : *CF* of the sub cluster represented by the ith child.

4. THE PROPOSED ALGORITHM

The Novel Algorithm

The Novel algorithm combines the features of BIRCH clustering algorithm which is a hierarchical clustering algorithm and Partitioning clustering algorithm *K*-Means algorithm. The algorithm is applied on Cars dataset. The Novel algorithm first make call to an algorithm that build a dendrogram containing large number of clusters on cars dataset.

These large numbers of clusters are difficult to predict and understand. After that the algorithm make call to *K*-Means clustering algorithm. In *K*-Means algorithm the numbers of clusters have to define in advance. In this paper the reading of within sum square error is taken for both *K*-Means and the proposed algorithm by changing the number of cluster values. The Basic steps of Novel algorithm are:

The algorithm starts with an empty tree. Each node of the tree is a cluster and the node stores the node or leaf label, the cluster number and the number of instances in that cluster.

1. Insert an instance to the tree. Here instance is the data point from the dataset.
2. Find the appropriate leaf *i.e.* find the best host where to add the data point based on the minimum distance.
3. Modify the node with new insertion.
4. Check if it can absorb the new data point on the basis of cutoff which is the branching factor.
5. If the node is full, split into two leaf node and add one more entry in the parent node.
6. If the new entry is in the non-leaf node then merge the two children which are similar under new entry.

7. Repeat above steps until the complete tree is generated.
8. Now consider the clusters generated in the above steps.
9. Perform traditional clustering steps of *K*-Means as described earlier in the paper. Use clusters found in the tree for random selection of seeds.

5. RESULTS AND DISCUSSION

5.1 *K*-Means Clustering

Description

Perform *K*-Means clustering on a data matrix.

Usage

*k*means(*x*, centers, iter.max = 10, nstart = 1,
algorithm = *c*(“Hartigan-Wong”, “Lloyd”, “Forgy”,
“MacQueen”), trace=FALSE)
S_3 method for class ‘kmeans’
fitted(object, method = *c*(“centers”, “classes”), ...)

Arguments

X numeric matrix of data, or an object that can be coerced to such a matrix (such as a numeric vector or a data frame with all numeric columns).

Centers either the number of clusters, say *k*, or a set of initial (distinct) cluster centres. If a number, a random set of (distinct) rows in *x* is chosen as the initial centres.

iter.max the maximum number of iterations allowed.

Nstart if centers is a number, how many random sets should be chosen?

algorithm character: may be abbreviated. Note that ”Lloyd” and ”Forgy” are alternative names for one algorithm.

Object an *R* object of class ”kmeans”, typically the result ob of ob <- kmeans(..).

method character: may be abbreviated. ”centers” causes fitted to return cluster centers (one for each input point) and ”classes” causes fitted to return a vector of class assignments.

Trace logical or integer number, currently only used in the default method (“Hartigan-Wong”): if positive (or true), tracing information on the progress of the algorithm is produced. Higher values may produce more tracing information.

... not used.

Details

The data given by x are clustered by the *k*-means method, which aims to partition the points into *k* groups such that the sum of squares from points to the assigned cluster centres is minimized. At the minimum, all cluster centres are at the mean of their Voronoi sets (the set of data points which are nearest to the cluster centre).

Value

kmeans returns an object of class ”kmeans” which has a print and a fitted method. It is a list with at least the following components:

Cluster A vector of integers (from 1:k) indicating the cluster to which each point is allocated.

centers A matrix of cluster centres.

Totss The total sum of squares.

withinss Vector of within-cluster sum of squares, one component per cluster.

tot.withinss Total within-cluster sum of squares, *i.e.* sum(withinss).

Create a data set

```
> library(MASS)
> set.seed(1234)
> x <- mvrnorm(1e5, mu=rep(0,5), Sigma=diag(1,5))
> x <- rbind(x, mvrnorm(1e5, mu=rep(10,5), Sigma=diag(0.1,5)+0.9))
> cl <- kmeans(x, 5, nstart = 10)
> plot(x, col = cl$cluster)
```

5.2 Novel Algorithm Birch with Kmeans Clustering

kmeans.birch(birchObject, centers, nstart = 1)

Arguments

birch Object

An object created by the function birch.

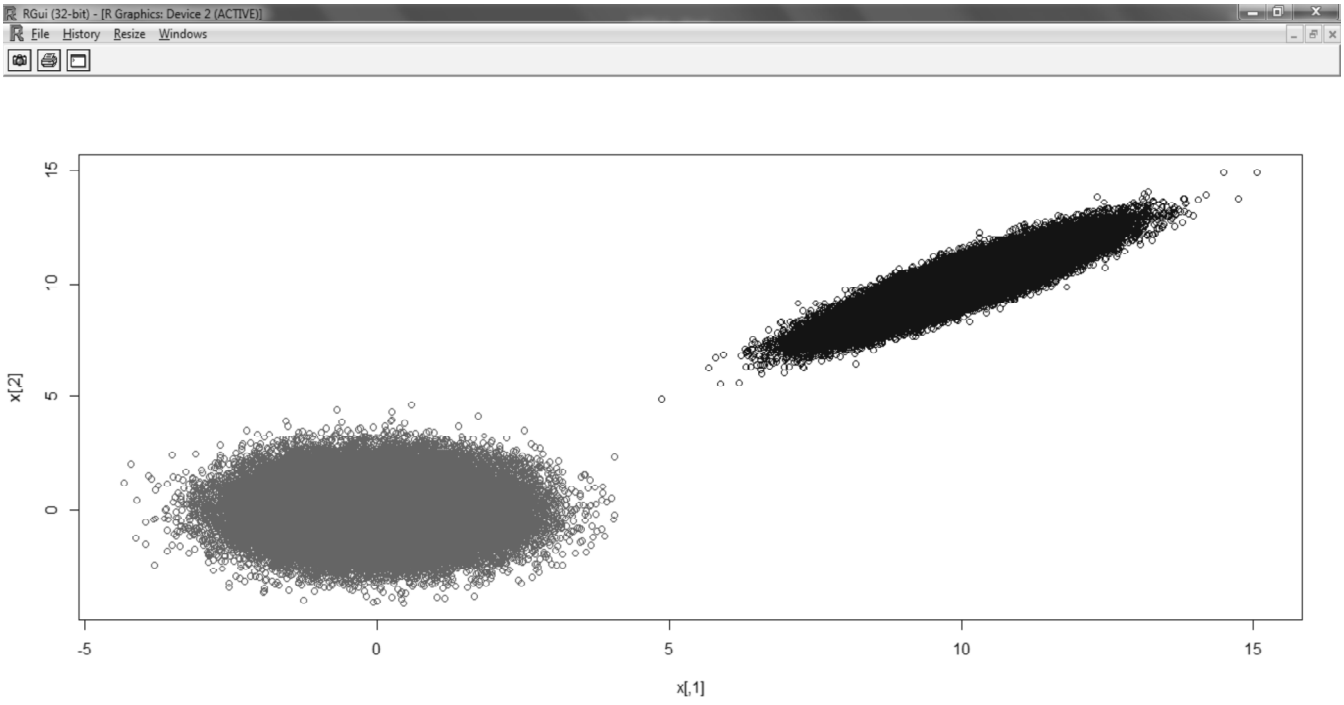

Figure 1: Clusters formed after applying only Hierarchical clustering algorithm

centers

Either the number of clusters or a set of initial (distinct) cluster centres. If a number, a random set of (distinct) subclusters in 'birchObject' is chosen as the initial centres.

nstart

If 'centers' is a number, how many random sets should be chosen?

Details

The birch object given by 'birchObject' is clustered by the *K*-Means method, adjusted for dealing with birch objects. The aim is to partition the subclusters into k groups such that the sum of squares of all the points in each subcluster to the assigned cluster centers is minimized. The result should be approximately similar to that found by performing *K*-Means on the original data set. However, this approximation depends on the "coarseness" of the underlying tree, and the size of the combinatorial problem. These aspect as discussed in detail in the references.

Values

Returns a list with components:

RSS

The total residual sum-of-squares of the clustering.

clust

A list containing a vector of which sub clusters make up the clustering (sub) and a vector with the underlying observations that make up the clusters (obs)

Create a data set

```
> library(MASS)
> set.seed(1234)
x <- mvrnorm(1e5, mu=rep(0,5), Sigma=diag(1,5))
x <- rbind(x, mvrnorm(1e5, mu=rep(10,5), Sigma=diag(0.1,5)+0.9))
```

Create birch object

```
birchObj <- birch(x, 5, keeptree = TRUE)
```

Leaving a tree in memory

```
birchObj <- birch(x, 5, keeptree=TRUE)
    birch.addToTree(x, birchObj)
    birchObj <- birch.getTree(birchObj)
    data(birchObj)
```

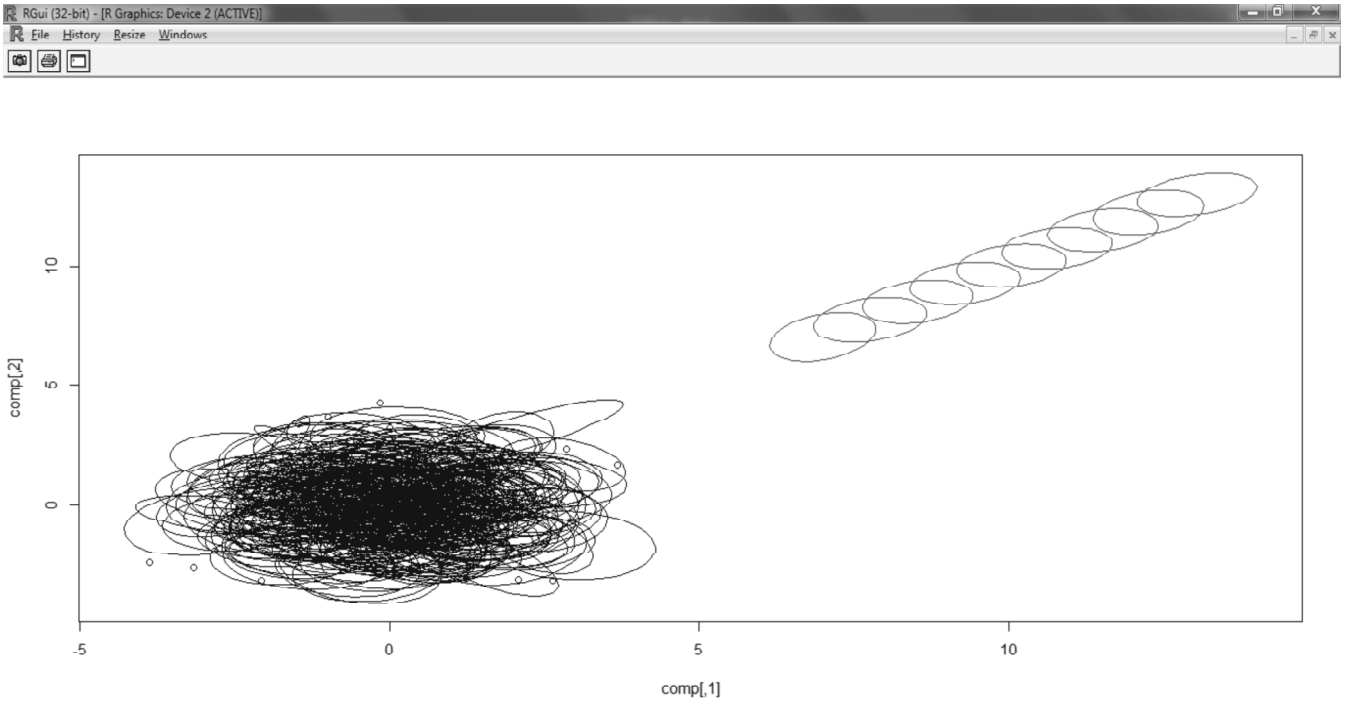

Figure 2: Clusters formed after applying *K*-Means on the above clusters generated.

Perform K-Means, specifying the number of clusters

```
kOut <- kmeans.birch(birchObj, 2, nstart=10)
 ## Perform K-Means, specifying the initial cluster centers
## See dist.clust for one method of initial cluster centers
kOut <- kmeans.birch(birchObj, matrix(c(0,10), ncol=5, nrow=2))
 ## To plot using the birch object
plot(birchObj, col=kOut$clust$sub)
```

The given 'birchObject' is clustered by the *K*-Means method, adjusted for dealing with birch objects. The aim is to partition the subclusters into k groups such that the sum of squares of all the points in each sub cluster to the assigned cluster centers is minimized.

The result is approximately similar to that found by performing *K*-Means on the original data set. However, this approximation depends on the "coarseness" of the underlying tree, and the size of the combinatorial problem.

The objective combining birch with *K*-Means is to minimize running time and data scans, thus formulating the problem for large databases. The clustering decisions on birch are made without scanning the whole data. Moreover, the algorithm exploit the non-uniformity of data – treat dense areas as one, and remove outliers (noise).

Sum of Square Error: For each data point in the cluster the distance from the data point to its cluster center is squared and the distances are summed.

$$SSE = \sum_{i=1}^{K} \sum_{xCi} \left| x - V_i \right|^2$$

Here $|x - V_i|$ is the *dist* defined above in the algorithm, x is the data point in the ith cluster, V_i is the centroid of the cluster The algorithm with small value of sum of square error is considered to be more accurate.

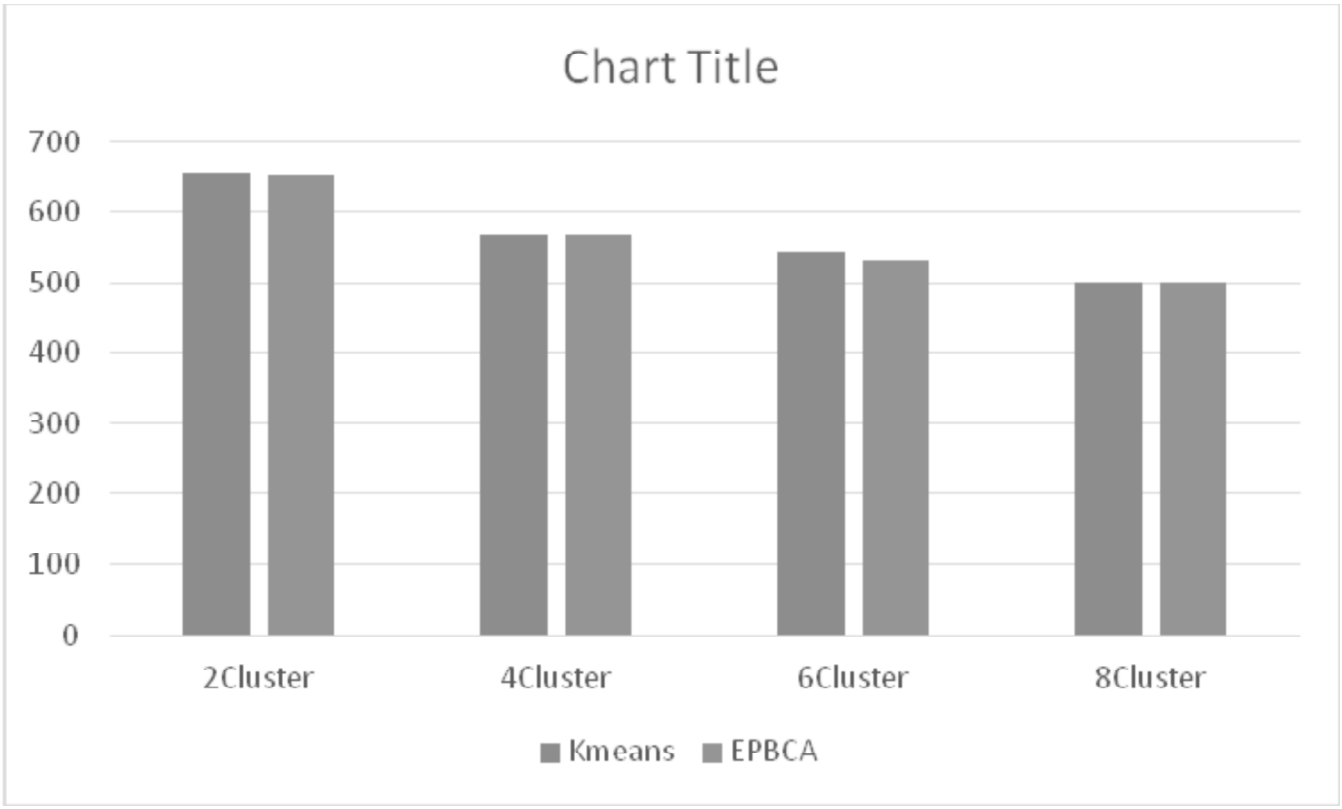

Figure 3: Graphical representation of within sum square error of *K*-Means and the Proposed Algorithm

Table 1
Comparison Between *K*-Means and Novel Algorithm with Number of Clusters And Sum of Square Error On Cars Dataset

No.of Cluster	*Kmeans*	*Novel*
2	654.74	653.12
4	570.21	569.153
6	543.143	532.329
8	501.14	500.411

5. CONCLUSION

In this research paper the features of Birch tree and K Means clustering algorithms are combined and a new Novel algorithm is proposed. The comparison of Novel is done with *K*-Means using R data mining tool. The comparison results are given in the table by changing the number of clusters value. The results show that as the number of clusters values increases the sum of square error value of Novel is always less than *K*-Means algorithm which means the proposes algorithm gives more accuracy with less sum of square error than *K*-Means algorithm.

REFERENCES

[1] Shalove Agarwal, Shashank Yadav, Kanchan Singh, "*K*-Means versus *K*-Means ++ Clustering Technique", *IEEE Second International Workshop on Education Technology and Computer Science*, 2012.

[2] NidalIsmael, Mahmoud Alzaalan, WesamAshour, "Improved Multi Threshold Birch Clustering Algorithm*" International Journal of Artificial Intelligence and Applications for Smart Devices*, **2**, 1-10, 2014.

[3] Saurabh Shah, Manmohan Singh, "Comparison of A Time Efficient Modified *K*-mean Algorithm with *K*-Mean and *K*-Medoid algorithm", *IEEE International Conference on Communication Systems and Network Technologies*, 2012.

[4] Shaina Dhingra, RimpleGilhotra, Ravishanker, "Comparative Analysis of Kohonen-SOM and *K*-Means data mining algorithms based on Academic Activities", *International Journal of Computer Applications*, 2013.

[5] Dr. S. Vijayarani, Ms. P. Jothi "An Efficient Clustering Algorithm for Outlier Detection in Data Streams" *International Journal of Advanced Research in Computer and Communication Engineering*, **2(9)**, 2013.

[6] Shi Na, Liu Xumin, Guan yong, "Research on kmeans Clustering Algorithm an Improved *K*-Means Clustering Algorithm*"*, *IEEE Third International Symposium on Intelligent Information Technology and Security Informatics*, 2010.

[7] Y. Ramamohan, K. Vasantharao, C. KalyanaChakravarti, A.S.K. Ratnam, "A Study of Data Mining Tools in Knowledge Discovery Process*"*, *International Journal of Soft Computing and Engineering (IJSCE)*, **2(3)**, 2012.

[8] J. Han and M. Kamber, "Data Mining: concepts and techniques", Beijing: China Machine Press, 2012.

[9] Yogita Rani, Manju, Harish Rohil, "Comparative Analysis of BIRCH and CURE Hierarchical Clustering Algorithm using WEKA 3.6.9", *The SIJ Transactions on Computer Science Engineering & its Applications, (CSEA)*, **3**, 1115-1122, 2013.

[10] RichaDhiman, ShevetaVashisht, "A Cluster analysis and Decision Tree Hybrid Approach in Data Mining to Describe Tax Audit*"*, *International Journal of Computers & Technology*, **4**, 2013.

[11] Wei-Lun Chang, Tzu-Hsiang Lin, "A Cluster-Based Approach for Automatic Social Network Construction", *IEEE International Conference on Social Computing / IEEE International Conference on Privacy, Security, Risk and Trust*, 2013.

[12] Archana Singh, Avantika Yadav, Ajay Rana, "Kmeans with Three different Distance Metrics", *International journal of computer Applications.*, **67**, 0975-8887, 2013.

[13] K. Kameshwaran, K. Malarvizhi, "Survey on Clustering Techniques in Data Mining", *International Journal of Computer Science and Information Technologies*, **5(2)**, 2272-2276, 2014.

[14] Ji Dan, Qiu Jianlin, Gu Xiang, Chen Li, He Peng "A Synthesized Data Mining Algorithm Based on Clustering and Decision Tree", *IEEE International Conference on Computer and Information Technology*, 2010.