Comparison of Parameter Free MST Clustering Algorithm with Hierarchical Agglomerative Clustering Algorithms

Ramakrishnam Raju. BHVS
Department of information Technology,
SRKR Engineering College,
Bhimavaram, Andhra Pradesh, India, 534204.

Valli Kumari.V
Dept of Computer Science & Systems Engineering,
Andhra University, Visakhapatnam,
Andra Pradesh, India, 530003

ABSTRACT

Clustering is a splitting up of data into groups of similar objects called clusters. The objects in a cluster are similar between themselves and dissimilar compared to objects of other clusters. This paper is intended to study and compare different data clustering algorithms. The algorithms in investigation are: hierarchical agglomerative clustering algorithms: Parameter Free Minimum Spanning Tree (MST) clustering algorithm and single link, complete link and average link clustering algorithms. K-means partitional clustering algorithm is used in the results as a reference. Our experimental evaluation shows that Parameter Free Minimum Spanning Tree algorithms are lead to better clustering results than hierarchical agglomerative algorithms, which suggests that Parameter Free Minimum Spanning Tree clustering algorithms are well-suited for clustering.

General Terms

Data Mining, Clustering.

Keywords

Clustering; hierarchical, Partitional, Minimum Spanning Tree.

1. INTRODUCTION

Clustering is a division of data into groups of similar objects, called clusters. The objects in a cluster are similar between themselves and dissimilar compared to objects of other clusters. Clustering methods are broadly classified as hierarchical and partitioning clustering. Hierarchical clustering views each data point as a node and at each iterative step merges two neighboring nodes to form a new node. A tree is constructed in a bottom up fashion after $n - 1$ steps, where n is the number of data points. For merging two nodes different linkage methods can be used which decide the neighboring pair of nodes to merge. Single linkage computes shortest distance of pairwise points from the two nodes. Complete linkage computes largest distance of pair-wise points from the two nodes. Average linkage computes average distance of pair-wise points from the two nodes. The problem with hierarchical clustering is that it decides the nodes to be merged locally on the basis of some form of linkages without taking a global objective into consideration and once nodes are merged they cannot be separated in later steps [1].

In contrast to the hierarchical clustering algorithm, partitional clustering find a single partition of the patterns instead of a clustering structure. It usually generates clusters by evaluating a criterion function which is defined locally or globally and attempts to recover the natural clusters present in the patterns. The advantage of partitional clustering methods is that they are especially appropriate in the analysis of large data sets, wherever a dendrogram based hierarchical clustering method is computationally expensive and is impractical with more than a few hundreds patterns [2,3]. The problem with partitional algorithm is the setting of parameter for the number of desired output clusters. The graph-theoretic clustering is one of the partitional clustering techniques to partition the given dataset. The well-known graph-theoretic clustering algorithm is based on building of the minimal spanning tree (MST) of the data [4], and then deleting longer edges from the MST to generate clusters.

In cluster analysis, one of the most important issues is the measure used to evaluate the quality of the clustering results that are produced. This measure can then be used to compare the solutions from different algorithms. This paper is intended to study and compare agglomerative hierarchical clustering with three linkages and Partitional clustering algorithm based on Minimum Spanning Tree. The results are to be compared on the basis of validity ratio [5] as a measure of cluster analysis. As a reference K-Means Algorithm is also presented in the evaluation.

2. TERMINOLOGY

2.1 Hierarchical Agglomerative Clustering Algorithm

A hierarchical clustering is a nested sequence of partitions [6]. This technique works on both bottom-up and top-down approaches. Based on the approach hierarchical clustering is further subdivided into agglomerative and divisive [6]. The agglomerative hierarchical technique follows bottom up approach whereas divisive follows top-down approaches. Hierarchical clustering uses different linkage criteria, which specifies the dissimilarity in the sets as a function of the pair-wise distances of observations in that sets. The linkage criteria could be of single linkage, average linkage and complete linkage [6]. In this paper, agglomerative clustering algorithm with all the three linkages is implemented for comparison. The advantages of hierarchical clustering algorithms are the reason for selecting this category for discussion. The advantages are flexibility, popularity and these algorithms are more versatile [7].

For n samples, agglomerative algorithms [8] begin with n clusters and each cluster contains a single sample or a point. Then two clusters will merge so that the similarity between them is the closest until the number of clusters becomes one or as given by the user [9,10],

1. Start with n clusters, and a single sample indicates one cluster.
2. Find the most similar clusters C_i and C_j then merge them into one cluster.
3. Repeat step 2 until the number of cluster becomes one or as specified by the user.

The distances between each pair of clusters are computed to choose two clusters that have more opportunity to merge. There are different methods to calculate the distances between the clusters C_i and C_j.

Notation:

$X_1, X_2, \dots , X_k$ =Observations from cluster 1

$Y_1, Y_2, \dots , Y_k$ = Observations from cluster 2

d(x, y) = Distance between a subject with observation vector *x* and a subject with observation vector *y*.

The methods for calculating distance between clusters are called linkage methods [11] and are as shown below:

Single Linkage: The distance between the two closest members of two clusters is,

$$d_{12} = \min_{ij} d(X_i, Y_j) \tag{1}$$

Complete Linkage: The distance between the two farthest members of two clusters is,

$$d_{12} = \max_{ij} d(X_i, Y_j) \tag{2}$$

Average Linkage: This method involves looking at the distances between all pairs and averages all of these distances.

$$d_{12} = \frac{1}{kl} \sum_{i=1}^{k} \sum_{j=1}^{l} d(X_i, Y_j) \tag{3}$$

Several more complicated agglomerative clustering algorithms [11], including group average linkage, median linkage, centroid linkage, and Ward's method, can also be constructed by selecting appropriate coefficients in the formula. Single linkage, complete linkage and average linkage consider all points of a pair of clusters, while calculating their inter-cluster distance, and are also called graph methods. In recent years, with the requirement for handling large-scale data sets in data mining and other fields, many new Hierarchical Clustering techniques have emerged and greatly improved the clustering performance. Typical examples include CURE [12], ROCK [13], Chameleon [14], and BIRCH [15].

2.2 Partitional Graph Theoretic Clustering

In graph based clustering methods, objects or their representatives are considered as vertices of the graph, and edges represent the relationships between them. The graph may also be a weighted or a non-weighted graph. In the graph based clustering methods, if the graph is a weighted graph, the labels of the edges are mostly derived from the corresponding similarity or distance measures of the objects. The main aim of the graph based clustering process is to determine and to eliminate those edges, which connect less similar objects. Such edges are called inconsistent edges. The elimination of the inconsistent edges leads to a clustering result, where the disconnected subgraphs yield the resulted clusters.

Minimum Spanning Tree (MST) of a graph is a tree which has all the vertices of the graph as nodes such that the total edge weight of the tree is minimum. This approach utilizes graph theory to find clusters. It views data points as nodes and defines a distance measure (like Euclidean distance [4]) between nodes as weight of an edge between nodes that is connecting the two nodes. A minimum spanning tree is constructed, now in order to get k clusters the tree has to be k-partitioned by removing k−1 edges [16]. There can be various criteria for removing− k1 edges like picking the longest k− 1 edges or partition with a global objective of minimizing the total distance between the center of each cluster and its data points.

Using a minimal spanning tree for clustering was initially proposed by Zahn [4]. A minimal spanning tree can be efficiently computed in O(n^2) time using either Prim's [17] or Kruskal's [18] algorithm. Clustering by minimal spanning tree can be viewed as a hierarchical clustering algorithm which follows the divisive approach. Using this method first a linked structure of the objects is constructed, and then the clusters are recursively divided into subclusters. In this paper the Parameter Free Minimum Spanning Tree clustering algorithm that is used is based on two phase process [19] with minimum user intervention; splitting the initial MST to get rough clustering and then fine tuning is done through merging the neighboring clusters. The algorithm is as shown below.

The MST Algorithm:

```
Input : S the point set.
Output : number of clusters and validity index

Let $e_l$ be an edge in the MST constructed from S
Let W be the weight of $e_l$
Let σ be the standard deviation of the edge weights in MST
Let $S_T$ be the set of disjoint subtrees of MST
Let $n_c$ be the number of clusters

1. Construct an MST from S
2. Compute the average weight of $\hat{W}$ of all the Edges from
MST
3. Compute standard deviation σ of the edges MST
4. $S_T$ = ø; $n_c$ = 1; C = ø; validity ratio=Inf;
5. Repeat
6.   For each el Є MST
7.     If ($W_e > \hat{W} + \sigma$)
8.       Remove el from MST
9.       $S_T = S_T$ U { T' } // T'' is new disjoint
           Subtree (regions)
10.      $n_c = n_c+1$
11.      Compute the center Ci of Ti
```

```
12.   End
13.       Compute Validity Index
14. Until (all the edges whose length $W_e > \hat{W} + \sigma$ are   removed)
15. While (Validity index decreasing  )
16.     Compute the center $C_i$ of each $T_i$  $\in S_T$
17.     Merge the two clusters whose distance between   their
            centers  is minimum.
18.         $n_c = n_c-1$
19.   Compute Validity Index
20. End
21. $n_c = n_c+1$
22. Return $n_c$ , Validity Index values.
```

The above algorithm does not require the user to specify the parameters to terminate the algorithm. The splitting Phase of the algorithm (lines 6-13) terminates when the condition We >Ŵ + σ, is not satisfied. The second phase of the algorithm i.e. merging (lines 16-19) continues as long as the validity ratio is decreasing. In this algorithm the user intervention is minimized. The final number of clusters and validity index of clustering are returned at line 22. The block diagram of our method is as shown in the Fig. 1.

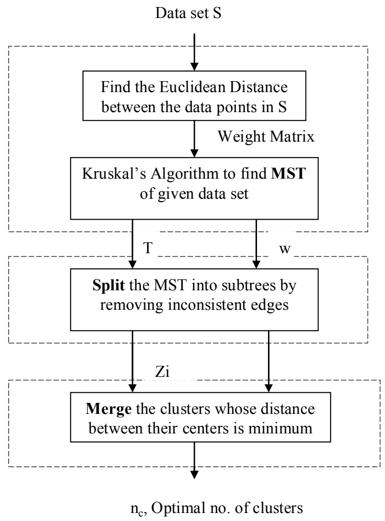

Data set S
Find the Euclidean Distance between the data points in S
Weight Matrix
Kruskal's Algorithm to find MST of given data set
T
w
Split the MST into subtrees by removing inconsistent edges
Zi
Merge the clusters whose distance between their centers is minimum
n_c, Optimal no. of clusters

Fig 1: Block Diagram showing the MST Algorithm

2.3 Hierarchical vs Graph Theoretic Algorithms

Clustering techniques based on hierarchical and graph theoretic approaches are shown to be related in certain ways [2]. For example, Single-link clusters are subgraphs of the minimum spanning tree of the data [20] and as well as are the connected components [21]. The Complete-link clusters are maximal complete subgraphs, and are linked to the node colorability of graphs [22]. The maximal complete subgraph was considered the stringent definition of a cluster in [23] [24]. A graph-oriented approach for non-hierarchical structures and overlapping clusters is presented in [25]. The Delaunay graph (DG) is obtained by connecting all the pairs of points that are Voronoi neighbors. The DG contains all the neighborhood information contained in the MST and the relative neighborhood graph (RNG) [26]. More inter-cluster distance measures, based on mean, were introduced by Yager [27], with additional discussion on their possible effect to control the hierarchical clustering process.

Graph theory can also be applied for nonhierarchical clusters. Zahn's clustering algorithm try to find connected components as clusters by detecting and discarding inconsistent edges in the minimum spanning tree [4]. Hartuv and Shamir treated clusters as highly connected subgraphs (HCS), where highly connected means the connectivity (the minimum number of edges needed to disconnect a graph) of the subgraph is at least half as great as the number of the vertices [28]. A minimum cut (mincut) procedure, which aims to separate a graph with a minimum number of edges, is used to find these HCSs recursively.

The difference between partitional and hierarchical approaches is that partitional method divides the data into pre-defined number of partions where as hierarchical method generates a tree structure of nested partitions. The hierarchical approach does not demand for the number of clusters in advance. The time and space complexities of the partitional algorithms are lower than that of the hierarchical algorithms [29]. The single-link clustering algorithm works well on data sets containing non-isotropic clusters whereas typical partitional algorithm works well only on data sets containing isotropic clusters [30]. The construction of hybrid algorithms that exploit the good features of both categories is shown in [31].

Table 1 lists the time and space complexities of several well-known algorithms, where, n is the number of patterns to be clustered, k is the number of clusters, and l is the number of iterations.

Table 1. Complexity of clustering algorithms

Clustering Algorithm	Time Complexicity	Space Complexicity
k-Means	O(nkl)	O(k)
Single-link	O(n^2log^n)	O(n^2)
Complete-Link	O(n^2log^n)	O(n^2)
MST	O(n^2)	O(n)

3. VALIDITY INDEX

Validity index is generally used to evaluate the clustering results quantitatively. In this paper the validity index, which is based on compactness and isolation is used. Compactness measures the internal cohesion among the data elements whereas isolation measures separation between the clusters. The compactness is measured by Intra-cluster distance and separation is measured by Inter-cluster distance [32], which is defined as follows.

Intra-cluster distance: This is the average distance of all the points within a cluster from the cluster centre. This measure is :

$$Intra = \frac{1}{N}\sum_{i=1}^{K}\sum_{x\in C_i} u_{ij}{}^{m}\left\|x - z_i\right\|^2 \qquad (4)$$

Where N is the number of data items in the dataset, K is the number of clusters, and Z_i is the cluster centre of cluster C_i.

Inter-cluster distance: This is the minimum of the pair wise distance between any two cluster centers given by

$$Inter = \min\left(\left\|z_i - z_j\right\|^2\right),$$
$$.i = 1,2,....,K-1 \qquad (5)$$
$$j = i+1,...,K$$

In the evaluation of the clustering algorithms, the validity index used is that proposed by Ray and Turi [5] as follows

$$validity = \frac{Intra}{Inter} \qquad (6)$$

These equations are in conformance of the definitions of Intra and Inter cluster distances that the separation increases in its value and compactness decreases as with the increase in the number of clusters. This is useful in estimating the optimal number of clusters.

4. EXPERIMENTAL RESULTS

Please use In this section a series of experiments are conducted on both synthetic and real datasets to demonstrate the validity. The reason for choosing the synthetic datasets is that they are easy to control and can be designed to contain a certain number of clusters. Three synthetic datasets were generated. The MST algorithm is tested on 2D datasets and compared with Hierarchical agglomerative and k-means algorithm. The 2D datasets used are Dataset1, Dataset2, and Dataset3 as show in the Fig. 2. The details of these datasets are presented in table 2.

Dataset1 (Fig.2 (a)) and Dataset2 (Fig. 2 (b)) have 3 and 2 well separated clusters respectively. Dataset3 has three clusters with some outliers (Fig. 2 (c). Moreover the sizes of the clusters are different. Experiments were also conducted on real world data sets: Iris, Soybean and Breast tissue data sets. The description of all these data sets is given in UCI machine learning repository [33]. The details are presented in the table II. The raw data sets are used in the experiments, for getting the accurate results while solving clustering problems. If the data is normalized, although it is usual to get the better clustering results, the clustering results not only depend on clustering methods, but also depend on normalization methods. Therefore, we decided not to normalize the data in order to ensure that the clustering results absolutely depend on the accuracy of clustering methods.

The results of the above datasets are depicted in table 3. The validity ratio based on MST clustering for Dataset3 is 0.1209, where as for K-Means this ratio is 0.2163. In presence of outliers, MST based clustering algorithm shows better results than K-Means algorithm. The validity ratio for Dataset1 and Dataset2 is same irrespective of the algorithm used as the clusters are very well defined.

Table 2. Datasets

Name	**No. of attributes**	**Data Size**	**No. of Clusters**
DataSet1	2	60	3
DataSet2	2	40	2
DataSet3	2	163	3
Iris	4	150	3
Soybean	35	47	4
Breast Tissue	9	106	6

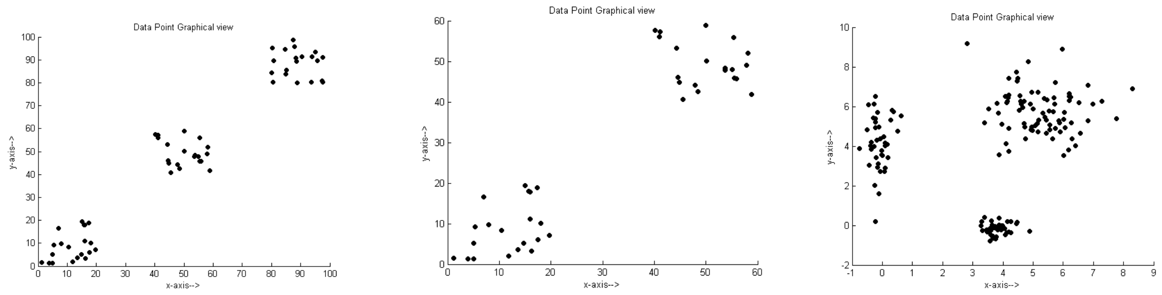
Three real datasets from UCI Machine Learning Repository are employed to test the validity of algorithms. The datasets are Iris, soybean and Breast Tissue. The performance of different clustering algorithms on these datasets is shown in the table III. For the Iris dataset, of the three hierarchical clustering procedures the single-link and average link procedures showed the minimum validity ratio of 0.1488. The MST clustering algorithm given a value of 0.2373. The MST based partitional algorithm is better than standard k-Means algorithm for Iris dataset. The clustering performance of MST based clustering is better than the other algorithms under consideration for Soybean dataset. The k-means performed poorly for Breast Tissue dataset. The above results show that the performance of MST based clustering algorithm is better than the standard K-means algorithm for all the datasets used in this paper.

5. CONCLUSIONS

The clustering aims at recognizing and digs out significant groups in underlying data. Thus based on a clustering criterion the data are grouped so that data points in a cluster are more similar to each other than points in different clusters. Since clustering is applied in many fields, a number of clustering techniques and algorithms have been proposed. In this paper we discussed and compared different categories in which algorithms can be classified (i.e., partitional, hierarchical and grid-based clustering) and we presented representative algorithms of each category. We concluded the discussion on clustering algorithms by a comparative presentation using the validity ratio that MST clustering algorithm performs better than the classical K-means partitional algorithm. The experimental results also showed that the hierarchical trees produced by partitional algorithms are better than those produced by agglomerative hierarchical clustering algorithms. This paper demonstrated that MST based clustering algorithm outperforms other clustering algorithms, including Hirarchial and k-means on most of the benchmark data sets from the UCI repository. We intend to further look at the potentials of MST based clustering algorithm in various data mining domains where cluster boundaries are inherently irregular. We will continue to study the rich properties of the MST clustering techniques and identify new challenges of applying those techniques in practice.

6. REFERENCES

[1] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, David Botstein, “Cluster analysis and display of genome-wide expression patterns”, Proc. Natl. Acad. Sci. USA 95, 14863-14867.



(a) (b) (c)

Fig 2. Synthetic Data sets (a) Dataset1, (b) Dataset2, (c) Dataset3

Table 3. Validity Ratios

Name	Single-Link	Complete-Link	Average-Link	MST	K-Means
DataSet1	0.0224	0.0224	0.0224	0.0223	0.0224
DataSet2	0.02289	0.02289	0.02289	0.02289	0.02289
DataSet3	0.1461	0.0600	0.1461	0.1209	0.2163
Iris	0.1488	0.2013	0.1488	0.2373	0.4332
Soyabin	0.1903	0.1903	0.1903	0.1859	0.1903
Breast Tissue	0.0307	0.0307	0.0307	0.0307	0.7061

[2] Jain, A. K., Murty, M. N., and Flynn, P. J., 1999, "Data Clustering: A Review", ACM Computing Surveys, vol. 31, no. 3, pp. 264-323.

[3] Jain, A.K. and Dubes, R.C., Algorithms for Clustering Data, Prentice-Hall advanced reference series, Prentice-Hall. Englewood Cliffs, NJ, USA, 1988

[4] C. T. Zahn. “Graph-theoretic methods for detecting and describing gestalt clusters”. IEEE Trans. Comput., 20:68–86, 1971.

[5] S. Ray, R.H. Turi. “ Determination of Number of Clusters in K–Means Clustering and application in color Image Segmentation”, Proc. 4th Intl. Conf. ICAPRDT ‘99, Calcutta India, pp. 137-143 (1999).

[6] Hichem Frigui and Raghu Krishnapuram, “Clustering by Competitive Agglomeration”, Journal: Pattern Recognition-PR, Vol. 30, No. 7, pp. 1109-1119, 1997.

[7] Osama Abu Abbas, “Comparisons Between Data Clustering Algorithms,” The International Arab Journal of Information Technology, vol. 5, no. 3, pp. 320-325, 2008.

[8] Sung Young Jung, and Taek-Soo Kim, “An Agglomerative Hierarchical Clustering Using Partial Maximum Array and Incremental Similarity Computation Method”, Proceedings of the 2001 IEEE International Conference on Data Mining, p.265-272, November 29-December 02, 2001.

[9] “Cluster analysis” in http://en.wikipedia.org/wiki/Cluster_Analysis, last accessed on 29th April, 2011.

[10] “Hierarchical Clustering Algorithms” in http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/hierarchical.html, last accessed on 12th May, 2011

[11] Margaret H.Dunham “Data Mining Introductory and Advance Topics”, Low price Edition – Pearson Education, Delhi, 2003.

[12] S. Guha, R. Rastogi, and K. Shim, “CURE: An efficient clustering algorithm for large databases,” in Proc. ACM SIGMOD Int. Conf. Management of Data, 1998, pp. 73–84.

[13] Guha, S., Rastogi, R., and Shim, K. “ROCK: A robust clustering algorithm for categorical attributes,” Inf. Syst., vol. 25, no. 5, pp. 345–366, 2000.

[14] G. Karypis, E. Han, and V. Kumar, “Chameleon: Hierarchical clustering using dynamic modeling,” IEEE Computer, vol. 32, no. 8, pp. 68–75, Aug. 1999.

[15] T. Zhang, R. Ramakrishnan, and M. Livny, “BIRCH: An efficient data clustering method for very large databases,” in Proc. ACM SIGMOD Conf. Management of Data, 1996, pp. 103–114.

[16] Ying Xu, Victor Olman, Dong Xu, “Clustering gene expression data using a graph theoretic approach: an application of minimum spanning trees”, Bioinformatics, Vol 18, No 4, Pages 536-545, 2002.

[17] R. Prim. “Shortest connection networks and some generalizations”. Bell System Technical Journal, 36:1389–1401, 1957.

[18] J.B. Kruskal, “On the shortest spanning subtree of a graph and the traveling salesman problem”. American Mathematical Society, 7(48–50), 1956.

[19] B.H.V.S.Ramakrishnam Raju, V. Valli Kumari, “Parameter Free Minimum Spanning Tree (PFMST) based clustering algorithm”, PDCTA 2011, CCIS 203, pp. 552-560, Springer, 2011. (yet to be published)

[20] J.C. Gower and G.J.S. Ross, "Minimal spanning trees and single linkage cluster analysis". Applied Statistics, 18:54–64, 1969.

[21] C.C. Gotlieb and S. Kumar, "Semantic clustering of index terms". Journal of the ACM, 15(4):493–513, 1968.

[22] F.B. Backer and L.J. Hubert. "A graph-theoretic approach to goodness-of-fit in complete-link hierarchical clustering", Journal of the American Statistical Association, 71:870–878, 1976.

[23] J.G. Augustson and J. Minker, "An analysis of some graph theoretical clustering techniques", Journal of ACM, 17(4):571–588, 1970.

[24] V.V. Raghavan and C.T. Yu. "A comparison of the stability characteristics of some graph theoretic clustering methods", IEEE Transactions on Pattern Analysis and Machine Intelligence, 3:393–402, 1980.

[25] Kazumasa Ozawa, "A stratificational overlapping cluster scheme". Pattern Recognition, vol. 18, no. 3-4, pp. 279-286, 1985.

[26] G.T. Toussaint, "The relative neighborhood graph of a finite planar set", Pattern Recognition, 12:261–268, 1980.

[27] R. Yager, "Intelligent control of the hierarchical agglomerative clustering process," IEEE Trans. Syst., Man, Cybern., vol. 30, no. 6, pp.835–845, 2000.

[28] E. Hartuv and R. Shamir, "A clustering algorithm based on graph connectivity", Inf. Process. Lett., vol. 76, pp. 175–181, 2000.

[29] G. Nagy, "State of the art in pattern recognition". In Proceedings of the IEEE, volume 56, pages 836–862, 1968.

[30] W.H.E Day, Complexity theory: An introduction for practitioners of classification. In Clustering and Classification, P.Arabie and L. Hubert, Eds. World Scientific Publishing Co., Inc., River Edge, NJ. 1992

[31] K. Krishna and M. Murty, "Genetic K-means algorithm," IEEE Trans.Syst., Man, Cybern. B, Cybern., vol. 29, no. 3, pp. 433–439, Jun. 1999.

[32] A. K. Jain, Algorithms for Clustering Data, Prentice Hall, Englewood Cliffs, New Jersey, 1988.

[33] UCI Machine Learning Repository: Data Sets, http://archive.ics.uci.edu/ml/datasets.html, last accessed on 10th May, 2011.