Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Hybrid Clustering Algorithm

B. Chandra
Indian Institute of Technology Delhi,
Delhi, India
bchandra104@yahoo.co.in

***Abstract*—The paper presents a new graph based clustering algorithm. Traditional clustering algorithms have the drawback that it takes large number of iterations in order to come up with the desired number of clusters. The advantage of this approach is that the size of the dataset is reduced using graph based clustering approach and the required number of clusters is generated using K means algorithm. The proposed algorithm consists of two phases, the first phase being constructing the graph and de-associating the graphs into connected sub graphs which denote the number of sub groups within the data. In the second phase in order to group the sub graphs that are close to each other K means algorithm is employed.**

***Keywords*—Partition Clustering, k-Means, Density Based Clustering.**

I. INTRODUCTION

Clustering is one of the most techniques in data mining. Clustering helps in discovering groups and identifying interesting distributions and patterns in the underlying data. Clustering aims to organize a collection of data items into clusters, such that items within a cluster are more "similar" to each other than they are to items in the other clusters. Clustering algorithms take as input some parameters (e.g. number of clusters, density of clusters) and attempt to define the best partitioning of a data set for the given parameters. According to the method adopted to define clusters, the algorithms can be broadly classified into the following types [5]: Partition clustering, Hierarchical clustering, Density-based clustering, Grid-based clustering and clustering using graph theoretic approach. For Partition clustering the number of clusters needs to be predefined. The number of clusters formed in case of Hierarchical clustering depends on the threshold. The threshold is a heuristic parameter that depends on the domain knowledge of the person employing clustering. In Density based clustering algorithm, the number of clusters generated depends on the minimum density fixed apriori. Grid-based clustering is used for clustering spatial data, where spatial data is quantized and a hierarchical structure is formed. In this paper we propose a hybrid clustering algorithm by combining Graph theory based Clustering and Partition Clustering. The proposed approach has two phases. In the first phase the connected directed graph is generated and later cycles are disconnected from the graph to form smaller sub graphs. In the second phase medians of points in each of the sub graphs is computed and K means clustering is applied to obtain the required number of clusters. The performance of the proposed algorithm has been illustrated on several popularly known datasets form the UCI Machine Learning repository and it has been compared with performance of K means algorithm.

II. OVERVIEW OF EXISTING ALGORITHMS

K-Means is a commonly used Partitioning based clustering algorithm [6]. K means is an iterative algorithm which starts with a fixed number of cluster centers. Each point in the data set is assigned to the respective clusters based on which cluster center is the nearest. Once all the points are assigned the cluster centers are recomputed and the same process is repeated again. The algorithm stops when the cluster centers do not change. Another algorithm of this category is *PAM* (*Partitioning Around Medoids*). In PAM clustering algorithm a representative point called the mediod is assigned for each cluster. The most centrally located point within the clusters is chosen as the mediod for each of the c clusters. Then, each of the non-selected points is grouped with the medoid based on similarity measure. Medoids are then swapped with other non-selected points until all points qualify as medoid. PAM is an expensive algorithm as regards finding the medoids, as it compares an object with entire dataset [8]. CLARA (Clustering Large Applications) is an implementation of PAM for various subsets of the dataset and then outputs the best clustering out of these samples [8]. CLARANS (Clustering Large Applications based on Randomized Search), combines the sampling techniques with PAM. Hierarchical clustering algorithms can be subdivided into Agglomerative algorithms and divisive algorithms. In Agglomerative algorithms clusters are formed by merging the two closest clusters into one while Divisive algorithms split a large cluster into two clusters. The number of clusters depended on the values of the threshold used for the similarity measure used. BIRCH [14] uses a hierarchical data structure called CF-tree for partitioning the incoming data points. CF-tree is a height balanced tree, which stores the clustering features. It is based on two parameters, branching factor B and threshold T, which referred to the diameter of a cluster. BIRCH can typically find a good clustering with a single scan of the data and improve the quality further with a few additional scans. Each node in CF-tree can hold a limited number of entries due to its size and hence it does not always correspond to a natural cluster. Different clusters are generated for different orders of the same input data's it is order sensitive. In CURE [2] a combination of random sampling and partition clustering is used to handle large databases. Each cluster is represented by a certain number of points which are generated by selecting well-scattered points. These points are then shrunk towards the cluster centroid by a specified fraction. *ROCK* [3] is a robust clustering algorithm for Boolean and

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

categorical data that uses the concept of a point's neighbours and links for generating clusters. Density based algorithms typically regard clusters as dense regions of objects in the data space that are separated by regions of low density. *DBSCAN* [1] is a density based algorithm. The key idea in *DBSCAN* is that for each point in a cluster, the neighbourhood of a given radius has to contain at least a minimum number of points. In [4] another density-based clustering algorithm, *DENCLUE*, is proposed. This algorithm introduces a new approach to cluster large multimedia databases. DENCLUE clustering is based on two parameters, *N* which determines the influence of a data point in its neighbourhood and '< ' describes whether a density-attractor is significant, allowing a reduction of the number of density-attractors and helping to improve the performance. Spatial data are clustered using Grid-based algorithms which quantise the space into a finite number of cells and then do all operations on the quantised space. *STING* (Statistical Information Grid-based method) is a grid based clustering algorithm. It divides the spatial area into rectangular cells using a hierarchical structure. *STING* [11] computes various statistical parameters for the dataset and uses it to generate a hierarchical structure of the grid cells so as to represent the clustering information at different levels. WaveCluster [9] is the latest grid-based algorithm that is based on signal processing techniques (wavelet transformation) to convert the spatial data into frequency domain. Clusters are identified by finding the dense regions in the transformed domain. A-priori knowledge about the exact number of clusters is not required in WaveCluster. Some clustering algorithms that use graph theoretic approach have also been explored in the recent past. Graph-theoretic clustering algorithms basically consist of searching for certain combinatorial structures in the similarity graph, such as a minimum spanning tree [12, 13] or a minimum cut [10, 15] and concepts of generalizing a maximal complete sub graph to edge-weighted graphs [7] based on notion of a maximal clique. Graph based clustering algorithm have been primarily applied to cluster pixels within images at achieve efficient segmentation. Zhenyu and Richard demonstrated its application to the image segmentation problem [15]. The data to be clustered are represented by a weight undirected adjacency graph G. The weights reflect the similarity between the linked vertices. Clustering is achieved by removing edges of the graph to form mutually exclusive sub graphs such that the largest inter-subgraph maximum flow is minimized. EXCAVATOR (EXpression data Clustering Analysis and VisualizATiOn Resource) is Minimum Spanning Tree (MST) based approach suggested for clustering gene expression data [12]. The MST tree is partitioned by cutting particular set of edges ('long' edges from either local or global point of view). Shi et al proposed the normalized cut criterion [10], for generating sub graph of the image for segmentation. The normalized cut criterion measures both the total dissimilarity between the different groups as well as the total similarity within the groups. Rather than focusing on local features and their consistencies in the image data, the approach is aimed at extracting the global impression of an image. The proposed algorithm uses a graph theoretic approach to identify sub clusters within the data and then use K means clustering algorithm to cluster the representatives of these sub clusters.

III. PROPOSED ALGORITHM

The proposed algorithm is aimed at finding initial subgroups of within the data using graphs and later use K means algorithm to find the required number of clusters. The correlation between the patterns is used as the distance measure. The first point is chosen as the starting point. A directed edge is drawn to the point which has maximum correlation value with reference to the chosen point. This process is repeated recursively from the current point till a cycle is obtained. After this the next point form the remaining dataset is chosen and the same procedure is repeated till all the points in the dataset are exhausted. Once the entire graph is constructed, the graph is traversed to find cycles. The cycles are disconnected for the graph to form sub graphs. These sub graphs partition the data into small subgroups. Medians of each of these subgroups are taken as representatives for the subgroups and K means algorithm is employed to find the required number of clusters. The algorithm for constructing the graph and then further sub graphs is given as follows:

Algorithm
Input: $L_n \leftarrow$ *set of points*
Output: *Clusters*

Begin
Calculate distance matrix Dnxn of the normalized input data;
$S_n \leftarrow L_n$;
Curr $\leftarrow$ *first point in* L_n;
Prev $\leftarrow$ *Curr;*
$L_n = L_n \backslash Curr$;
Repeat
// For all points 'x' except Curr && Prev consider d(Curr,x) from Dnn;
$y \leftarrow max\{d(Curr,x), x \in S_n \,\&\, x \neq Curr,\ x \neq prev\}$;
// max function used above is for correlation distance measure;
// use min function for Euclidean distance measure;
$L_n = L_n - \{y\}$;
Prev $\leftarrow$ *Curr;*
Curr $\leftarrow y$;
Draw. Directed edge from Prev to Curr;
If (Curr is in Graph G)
Curr $\leftarrow$ *a point 'x' from* L_n
Prev $\leftarrow$*Curr;*
$L_n = L_n - \{x\}$;
End;
Until L_n *is not empty;*
// grouping into clusters
$i \leftarrow 1$;
Repeat
Traverse graph G from a vertex and until a cycle is identified;
$C_i \leftarrow$ *{set of all vertices of the Cycle};*
Remove all adjacent edges incident with points of C_i *& that vertex from G*
$i \leftarrow i+1$;

Until a cycle present in G;
For each connected component K remaining in G
$C_i \leftarrow$ *{set of all vertices of K};*
$i \leftarrow i+1$;
End;
End;

Illustration : Let the dataset be D= { G1=(5.1,3.5,1.4); G2=(4.9,3,1.4); G3=(4.7,3.2,1.3); G4=(4.6,3.1,1.5); G5=(5,3.6,1.4); G6=(5.4,3.9,1.7); G7=(4.6,3.4,1.4); G8=(5,3.4,1.5); G9=(4.4,2.9,1.4); and G10=(4.9,3.1,1.5)}

The distance matrix using correlation distance measure is shown in Table 1.

The graph constructed for the above distance matrix is given in Figure 1. G1 is chosen as the first starting point. Observing the distance matrix, G3 has maximum correlation with G1 and hence we draw a directed edge from G1 to G3. Now we look for the next point in the dataset (except G1 and G3) which has maximum correlation with G3. G8 has maximum correlation with G3 and hence we draw a directed edge from G3 to G8. The same process is carried out with G8 to find a point in the dataset other than G3 and G8. G1 is the point which satisfies the requirement and hence we draw a directed edge from G8 to G1. This forms a cycle and hence we pick another point in the dataset which has not been used for graph generated earlier. G2 is the point which satisfies this criterion. We repeat the above procedure until all the points are used for graph generation.

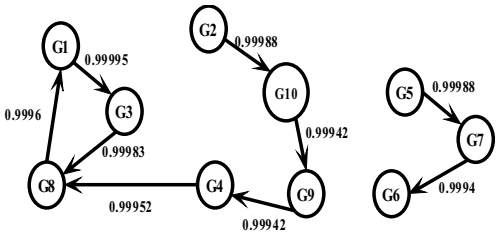

Figure 1: Graph for Dataset D

Once the graphs are constructed, the graph is traversed again to find out cycles within the graph. Each cycle is separated for the initial graph by cutting of the edges which connects the cycle to the initial graph. Here the edge between G4 and G8 would be cut, hence giving 3 different sub graphs (clusters) (see Figure 2)

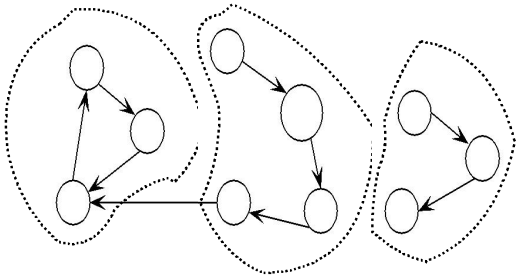

Figure 2: Sub Graphs after disassociating cycles

The proposed algorithm thus provides 3 clusters for dataset D. For a large dataset it may be possible that a large number of smaller clusters are formed. The purity of the clusters is decided during the development of sub graphs by disconnecting cycles form the connected graph. The median of each of the points in each of the sub clusters are computed and these sets of points are presented to the K means algorithm. The K means algorithm starts by partitioning the inputs points into k initial sets by assigning each point to the nearest cluster centers. The centroid of the points in each of the k sets is calculated and the algorithm is repeated until convergence which is attained when the points no longer switch clusters (i.e. the cluster centers do not change).

In the above example, three sub graphs consisting of points: {G1, G3 and G8}, {G2, G4, G9, G10} and {G5, G6, G7} are formed. The median of each of the points in the sub graph are taken as representatives for further clustering using K means. Let D_G be the dataset composed of the medians of the sub graphs of Dataset D. D_G = {(5, 3.4, 1.4), (4.75, 3.05, 1.45), (5, 3.6, 1.4)}. On applying K means algorithm on D_G, patterns 1 and 3 combine to form one cluster and pattern 2 forms the second cluster.

K means algorithm is used to merge each of these sub clusters to form the required number of clusters. The advantage of this approach is that K means algorithm is applied to the median of the sub clusters rather than the entire dataset. K means algorithm iterates till the cluster centers do not change. In each iteration a point shifts from one cluster to another, thus requiring more time to stabilize. In the proposed approach, smaller clusters formed after the graph is partitioned represent subgroups within the data points, which almost belong to the same class. By taking the median of the points belonging to each subgroup, the original size of the dataset is drastically reduced. K means algorithm when applied to the reduced sized dataset results into faster convergence.

IV. RESULTS

The performance of the proposed algorithm was evaluated using various datasets available in the UCI Machine Learning Repository. The results are compared with that of K means algorithm. (See Table 2.)

It could be observed that K means algorithm takes more number of iterations compared to that of the proposed algorithm. The proposed algorithm groups similar patterns in

the first phase to reduce the size of the dataset significantly. The proposed algorithm gives better classification accuracy for majority of datasets and it is observed that there is drastic reduction in the number of iterations taken for convergence using the proposed algorithm.

V. CONCLUSIONS

The proposed algorithm shows considerable reduction in the number of iteration required for convergence. Phase one of the proposed algorithm is an effective means of reducing the size of the dataset without compromising the overall final classification accuracy after clustering.

REFERENCES

[1] M. Ester, H.P. Kriegel., J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In Proceeding of 2nd Int. Conf. On Knowledge Discovery and Data Mining, Portland,1996, pp. 226–23.

[2] Guha, S., Rastogi, R., and Shim K., "CURE: An Efficient Clustering Algorithm for Large Databases". In Proceedings of the *ACM SIGMOD* Conference, 1998.

[3] S. Guha, R. Rastogi, and K. Shim, "ROCK: A Robust Clustering Algorithm for Categorical Attributes". In Proceedings of the *IEEE Conference on Data Engineering*, 1999.

[4] A. Hinneburg, and D. Keim, "An Efficient Approach to Clustering in Large Multimedia Databases with Noise". In Proceedings of *KDD Conference*, 1998.

[5] A.K. Jain., M.N. Murty, and P.J.Flyn, "Data Clustering: A Review". *ACM Computing Surveys*,1999, 31(3), pp 264–323.

[6] J.B. MacQueen,, "Some Methods for Classification and Analysis of Multivariate Observations". In Proceedings of 5th Berkley *Symposium on Mathematical Statistics and Probability*, 1967, vol.1: Statistics, pp. 281–297.

[7] P. Massimiliano and M. Pelillo, "A New Graph-Theoretic Approach to Clustering and Segmentation", Proceedings of the 2003 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR'03), 2003, 1063-6919/03.

[8] R. Ng, and J.Han, "Efficient and Effective Clustering Methods for Spatial Data Mining". In Proceeding's of the 20th *VLDB Conferenc*e, Santiago, Chile, 1994.

[9] C. Sheikholeslami, S. Chatterjee, and A. Zhang, "WaveCluster: A-MultiResolution Clustering Approach for Very Large Spatial Database". In Proceedings of 24th *VLDB Conference*, New York, USA,1998.

[10] J.Shi and J.Malik, "Normalized cuts and image segmentation". *IEEE Trans. Pattern Anal. Machine Intell.*, 2000, 22(8): pp.888–905.

[11] W. Wang, J.Yang, and R.Muntz, "STING: A Statistical Information Grid Approach to Spatial Data" Mining. In Proceedings of 23rd *VLDB Conference*,1997.

[12] X.Ying, V. Olman and D. Xu, "Clustering gene expression data using a graph theoretic approach: an application of minimum spanning trees", *Bioinformatics*, 2002, vol.18, No. 4, pp 536-545.

[13] C.T. Zahn., "Graph-theoretic methods for detecting and describing gestalt clusters". *IEEE Transactions on Computing*, 1971, vol.20, pp.68–86.

[14] T.Zhang, R.Ramakrishnman and M.Linvy, "BIRCH: An Efficient Method for Very Large Databases". *ACM SIGMOD*, Montreal, Canada, 1996.

[15] W. Zhenyu and L.Richard, "An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation". *IEEE Trans. Pattern Anal. Machine Intell.,* 1993, 15(11):1101–1113.