A Renovated Idea on Minimum Spanning Tree-Based Clustering Algorithm

A. R.Sivakumaran[1], M.Vanitha[2], Dr.P.Manimegalai[3]
[1,2]*Computer Science and Engineering,*
Karpagam College of Engineering, Coimbatore.
[3]*Electronics and Communication Engineering,*
Karpagam University, Coimbatore.

***Abstract*- Minimum Spanning Tree (MST) based clustering algorithms are capable of detecting clusters with irregular boundaries. This paper aims in analyzing and comparing the MST-inspired clustering algorithm with our Renovated MST based clustering algorithm. The MST-inspired clustering algorithm follows Reverse Delete algorithm for generating MST and uses DHCA (Divisive Hierarchical Clustering Algorithm) for formation of clusters. In our Renovated MST based clustering algorithm, we propose a new cycle testing algorithm for testing cycles, in generation of MST and we use Minimum Cost Subgraph property of MST along with cut & cycle property. The timing and efficiency of identifying the clusters are compared for both the algorithms.**

***Index Terms*- Clustering, Minimum Spanning Tree, Divisive Hierarchical Clustering Algorithm, New cycle testing algorithm**

I. INTRODUCTION

Clustering is the process of partitioning the data set into subsets, called clusters, so that the data in each subset share some properties in common. A *cluster* is therefore a collection of objects, which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. Different techniques, such as hierarchical, partitional, and density- and model-based approaches, have been developed for clustering. The MST method is a graphical analysis of an arbitrary set of data points. In such a graph, two points or vertices can be connected either by a direct edge, or by a sequence of edges called a path. The length of a path is the number of edges on it. The degree of link of a vertex is the number of edges that link to this vertex. A loop in a graph is a closed path. A connected graph has one or more paths between every pair of points. A tree is a connected graph with no closed loops.

1.1 SPANNING TREE

Given a connected, undirected graph $G = (V, E)$, where V is the set of nodes, E is the set of edges between pairs of nodes, and a weight $w (u, v)$ specifying weight of the edge (u, v) for each edge $(u, v) \in E$. A spanning tree is an acyclic subgraph of a graph G, which contains all vertices from G. The Minimum Spanning Tree (**MST**) of a weighted graph is minimum weight spanning tree of that graph. The weight for each edge can be the distance between its two end points. The distance measure may be Euclidean distance, Correlational distance, Mahalanobis distance[5] etc. A graph may have many spanning trees; for example, the complete graph on four vertices has sixteen spanning trees .

1.2 PROPERTIES OF MINIMIUM SPANNING TREE

Possible multiplicity: There may be several minimum spanning trees of the same weight having a minimum number of edges. If all the edge weights of a given graph are the same, then every spanning tree of that graph is minimum. *Uniqueness*: If each edge has a distinct weight, then there will be only one unique minimum spanning tree. *Minimum-cost subgraph*: If the weights are non-negative, then a minimum spanning tree is in fact the minimum-cost subgraph connecting all vertices, since subgraphs containing cycles necessarily have more total weight. *Cycle property*: For any cycle C in the graph [2], if the weight of an edge e of C is larger than the weights of other edges of C, then this edge cannot belong to an MST. *Cut property*: For any cut C in the graph, if the weight of an edge e of C is smaller than the weights of other edges of C, then this edge belongs to all MSTs of the graph. *Minimum-cost edge*: If the edge of a

graph with the minimum cost e is unique, then this edge is included in any MST.

1.3 MST ALGORITHMS

Kruskal's algorithm, Prim's algorithm and Reverse Delete algorithm are some of the MST algorithms. In the Kruskal's algorithm, all the edges are sorted into a nondecreasing order by their weights, and the construction of an MST starts with n trees, i.e., every vertex being its own tree. In the Prim's algorithm, the construction of an MST starts with some root node t and the tree T greedily grows from t outward. At each step, among all the edges between the nodes in the tree T and those not in the tree yet, the node and the edge associated with the smallest weight to the tree T are added. "Reverse Delete" algorithm starts with the full graph and deletes edges in order of nonincreasing weights based on the cycle property as long as doing so does not disconnect the graph.
MST-based clustering algorithms consist of three steps: First, A minimum spanning tree is constructed using either the Prim's algorithm or the Kruskal's algorithm. Second, The inconsistent edges are removed to get a set of connected components or clusters.In final Step 2 is repeated until some terminating condition is satisfied.

II. RELATED WORK

Hybrid Minimal Spanning Tree and Mixture of Gaussians based Clustering Algorithm[4] aims to increase the robustness and consistency of the clustering results and to decrease the influence of the user on the clustering results. Conditions for cutting MST: (i) Based on Distance between the vertices (ii) Based on distance of the separated sub trees (iii) Based on the separation which is specified with the goodness of the obtained partitions. The HEMST (Hierarchical Euclidean) based Clustering Algorithm[5] produces a K-Partition of a set of points for any given K value and MSDR (Maximun Standard deviation Reduction) clustering Algorithm partitions a point set into a group of clusters by maximizing the overall standard deviation reduction without the given K value. These algorithms can be applied to image color clustering. Color Clustering (with or without the given K value).

In Zahn's original work[3], the inconsistent edges are defined to be those whose weights are significantly larger than the average weight of the nearby edges in the tree. The performance of this clustering algorithm is affected by the size of the nearby neighborhood. Image Segmentation using MST[6] has the difficulty of partition an image into connected regions of similar texture or similar colors/gray-levels. Two classes of algorithms used here are (i) Boundary Detection-Based Approaches (ii) Region growing and Clustering-based approaches. Using (i) Hierarchical Clustering (ii) K-means clustering and iii) Clustering through self-organizing maps, Clustering of Gene Expression data[7] is performed. They are implemented as EXCAVATOR (EXpression data Clustering Analysis and VisualizATiOn Resource).

III. MST-INSPIRED CLUSTERING ALGORITHM

It uses an efficient implementation of the cut and the cycle property of the minimum spanning trees[2] and has much better performance than $O(N^2)$ time. The following two observations motivate design. (i) MST-Based Clustering Algorithms can be more efficient if longest edges can be identified quickly (ii) Prim's algorithm can be efficiently applied to each individual size-reduced cluster. Sequentially stored data items in the data structure consist of two arrays (i) Distance array (ii) Index array. The distance array is used to record the distance of each data point to some other data point in the sequentially stored data set. The index array records the index of the data item at the other end of the distance in the distance array. It is proposed to follow the sequential initialization (SI) by multiple runs of a recursive procedure known as DHCA (Divisive Hierarchical Clustering Algorithm)

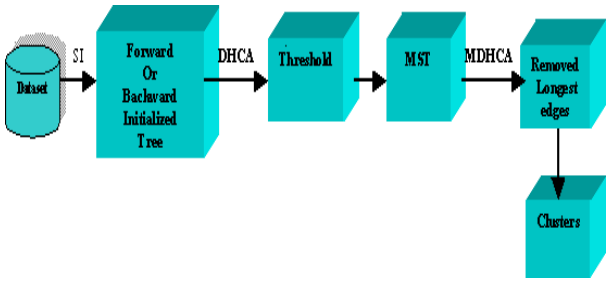
3.1 DIVISIVE HIERARCHICAL CLUSTERING ALGORITHM

A divisive clustering starts with one cluster of all data points and recursively splits the most appropriate cluster[1]. The process continues until a stopping criterion (frequently, the requested number k of clusters) is achieved. Initially all the data objects are considered in one cluster. Then for each object the degree of irrelevance is measured and the most irrelevant data object is split from the main cluster and a new cluster is formed with only that data object in it. The highest degree of irrelevance of an object corresponds to the one that is most distant from all other objects in that cluster. The most irrelevant object splits off and forms a new cluster. This is equivalent to splitting the cluster with the largest diameter. The process continues until it satisfies certain termination condition, such as a desired number of clusters are formed or the distance between two closest clusters is above a certain threshold distance.

3.2 MDHCA ALGORITHM

Divisive hierarchical methods face the difficulty of making a right decision of splitting at a high level. Such possibilities can be greatly reduced by multiple runs of DHCA. After the DHCA updates, to check whether the longest edge in the current spanning tree is associated with the true shortest edge that connects the two partitions a flag array is used. It marks all the points on one side of the longest edge to be 1 and all the points on the other side to be 0. Then the DHCA can be applied multiple times with the partition centers being chosen only from the data points marked either 1 or 0, but not both. This procedure is called as marked DHCA (MDHCA). MDHCA is an efficient method to implement the cycle property

MST Inspired Clustering Algorithm



MST-Inspired clustering algorithm can be summarized as following [1]:

Step1:Start with a spanning tree built by the SI Step2: Calculate the mean and the standard deviation of the edge weights in the current distance array and use their sum as the threshold. Partially refine the spanning tree by running our DHCA multiple times until the percentage threshold difference between two consecutively updated distance arrays is below 10^{-6}. Step3:Identify and verify the longest edge candidates by running MDHCA until two consecutive longest edge distances converge to the same value at the same places. Step4: Remove this longest edge found in step 3. Step 5: If the number of clusters in the data set is preset or if the difference between two consecutively removed longest edges has a percentage decrement larger than 50 percent of the previous one, we stop. Otherwise go to Step 3.

IV. PROPOSED ALGORITHM

A Renovated MST Clustering Algorithm

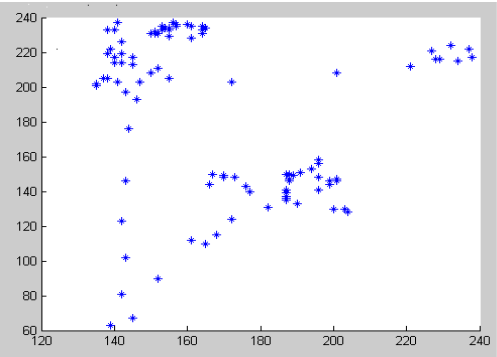
The Minimum Spanning Tree algorithm of a weighted graph with renovated idea is proposed. Here, we use a new cycle testing algorithm [9] for testing cycles, in generation of Minimum Spanning Tree. The reason behind this is to optimize the execution time for cycle testing. Also, we describe Minimum Spanning Tree algorithms for weighted graph with minimum cost sub graph property in addition to cycle and cut property. We apply here new concept for explanation of minimum Spanning tree with better time complexity.

A new cycle testing algorithm for testing cycles in generation of Minimum Spanning Tree

Step 1: From the combination of edges and incidence matrix, obtain degree of each node contributed by the edges under consideration.
Step 2: Test whether at least two nodes of degree one. If not, go to step 6. Otherwise continue.
Step 3: Test whether at least three nodes of degree more than one? If not go to step 5
Step 4: Delete pendant edges, if exists of n-1 edges and modify the degree of the nodes accordingly and go to step 2. Otherwise go to step 6.
Step 5: Edge combinations are tree
Step 6: TerminateV. ALGORITHM PERFORMANCE

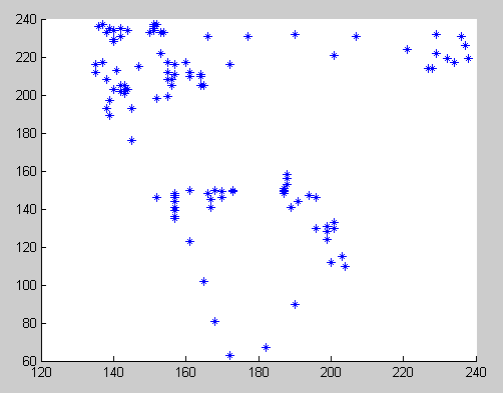
The result of Divisive Hierarchical Clustering using Kruskal Algorithm is compared with DHCA using Reverse Delete Algorithm for image data. The following graphs show the performance of above algorithms.

Fig: Formation of clusters for DHCA with Kruskal



The result shows that performance of DHCA using Reverse delete is better than performance of DHCA using Kruskal. The resulting clusters are clearly formed incase of Reverse delete algorithm

Fig: Formation of clusters for DHCA with Reverse delete



In the proposed system, the minimum cost subgraph property of MST is implemented along with cycle testing algorithm. If the weights are non-negative, then a minimum spanning tree is in fact the minimum-cost subgraph connecting all vertices, since subgraphs containing cycles necessarily have more total weight

VI. CONCLUSION

The existing system uses graph property of MST and renovated algorithm in construction of MST to improve the performance than DHCA and MDHCA. It also reduces the time complexity. In the future other rich properties may be applied to be useful in very large datasets and comparison can be done with the existing system

REFERENCES

[1] A Divide-and-Conquer Approach for Minimum Spanning Tree-Based Clustering Xiaochun Wang, Xiali Wang, and D. Mitchell Wilkes. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 7, JULY 2009 945

[2] Katriel, P. Sanders, and J.L. Traff, "A Practical Minimum Spanning Tree Algorithm Using the Cycle Property," Proc. 11th European Symp. Algorithms (ESA '03), vol. 2832, pp. 679-690, 2003.

[3] C.T. Zahn, "Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters," IEEE Trans. Computers, vol. 20, no. 1, pp. 68-86, Jan. 1971.

[4] A. Vathy-Fogarassy, A. Kiss, and J. Abonyi, "Hybrid Minimal Spanning Tree and Mixture of Gaussians Based Clustering Algorithm," Foundations of Information and Knowledge Systems, pp. 313-330, Springer, 2006

[5] O. Grygorash, Y. Zhou, and Z. Jorgensen, "Minimum Spanning Tree-Based Clustering Algorithms," Proc. IEEE Int'l Conf. Tools with Artificial Intelligence, pp. 73-81, 2006.

[6] Y. Xu and E.C. Uberbacher, "2D Image Segmentation Using Minimum Spanning Trees," Image and Vision Computing, vol. 15, pp. 47-57, 1997.

[7] Y. Xu, V. Olman, and D. Xu, "Clustering Gene Expression Data Using a Graph-Theoretic Approach: An Application of Minimum Spanning Trees," Bioinformatics, vol. 18, no. 4, pp. 536-545, 2002.

[8]. Renovation of Minimum Spanning Tree Algorithms of Weighted Graph *ACM Ubiquity, Volume 9, Issue 7 February 19 – February 25, 2008*