Data Clustering using Graph based Algorithm Approach

Ranjit Kumar¹, Devendra Kumar Mishra² & Rachna Dhaka³

¹²CSE, MDU, Haryana, India
³IT, MDU, Haryana, India

Abstract: In this paper, we discuss a graph based unsupervised clustering process which can cluster data object points in a data set using an initial input value $t$ and considering angles formed by certain data points to form clusters. The value of $t$ is calculated prior clustering from the data set itself which is then used as an input for clustering process. We have also taken some example data sets to arrive at clusters and pointed out the advantages and disadvantages that the algorithm has w.r.t. different kind of data set.

Keywords: Clustering, Dendrograms

1. INTRODUCTION

The major classes of data clustering technologies are hierarchical and partitioned. These classes are further divided into more specialized types and corresponding algorithms. Each class of clustering techniques has some merits and demerits of their own. Hierarchical clustering methods form dendrograms in which different clustering formations at different stages of partition are easy to depict. Also it is mostly an unsupervised technique. But hierarchical clustering also has a number of demerits. They form crisp clusters which is not always true as a given data object can be closely related to a number of clusters but hierarchical clustering will assign a data object to the cluster with which it has closest relation. Secondly for very big data set the dendrogram may become too complex in itself. Partitional clustering methods are further classified into squared error methods like k-means or graph theoretical methods, mixture resolving and mode seeking. But partitional clustering methods too have its own merits and demerits. The most important demerit of the most commonly used partitional based clustering k-means is that it needs a initial number of clusters among which it distributes the object points depending on the similarity values. Graph theoretical methods are also very popular partitional based clustering technology. Most of the graph theoretical algorithms form MST (minimum spanning tree) and then successively remove the longest edge to form clusters in hierarchical way.

2. CLUSTERING METHOD

Determination of value of $t$:

1. Pick any random point and find the distance of all other points from this point.
2. Find $A_i = D/(n-1)$. Where $D$ is sum of all distances & $n$ is total no. of points and $i$ ∈ {1, 2, 3}.
3. Go to the point that is farthest from this point and repeat step 1 and 2.
4. Go to the midpoint of the above two farthest point and find a data point (if any) within a distance of $min(A_1, A_2)$. Repeat steps 1 through 2 for this point. Call it $A_3$
5. Assign $t = sqrt{min(A_1, A_2, A_3)}$.

Algorithm:

1. Pick a random point $p_i$ from the unconnected points.
2. $C_m = graph(p_i, t)$ /* graph $(p_i, t)$ forms the cluster and return $m$ to name the cluster as $C_m */$
   join($p_i, t$) {
   Join all unconnected points(if any) within threshold $t$ of $p_i$ to $p_i$;
   sortasc($p_i$);
   /* sortasc() creates an array $S[j]$ that contains the sorted list of points joined to $p_i$ according to their distance from it in ascending order */
   for( all $k$ ∈ $S[j]$ ) {
   join($k$, $t$);
   }
   return $m++;$ //m is a global variable & $m$∈{natural numbers}
}
3. Repeat step 1 through 2 till no point which is closer than distance \( t \) from any cluster or any other point is joined.

4. for( all \( C_m \) ) {
   angle(\( C_m \)); /* creates a list \( R_{m,g} \) of points in cluster \( C_m \) which is connected to only two other points and make an angle >90° */
   split( \( R_{m,g} \) ); /* removes the edges for each point in \( R_{m,g} \) which is longer and split the cluster to rename the network which does not contains the point to a new cluster \( C_m \) */
}

5. final() {
   delete remaining single point clusters(if any) to leave them as outliers;
}

2.2. Drawbacks
The algorithm seems to be working well for randomly scattered data points but could not properly derive regular geometrical patterns and clusters in such case are not regular structures.

Advantages over the k-means algorithm:
1. It does not need any initial number of clusters to be provided as in case of k-means.
2. This algorithm does not force all the points to join a cluster like k-means.
3. The clusters are found from the data set without any prior information about the data objects.

But it has certain disadvantages too namely:
1. It is a distance based algorithm so cannot be used for pattern recognition in case of image processing.
2. It forms crisp clusters and so miss out on some benefits of fuzzy clustering method.
3. As the algorithm is weak in recognizing the shapes, it is not good in cases where the clusters to be found are regular shapes like line, square, hexagon etc.

3. CONCLUSION AND FUTURE WORK
In this paper we have applied the basic concepts of graph theoretic clustering to form an algorithm. It has advantages over partitional k-means algorithm. We observed that we can arrive at clusters with just one run. Although the algorithm has some shortcomings with regular geometrical patterns, it arrives at more suitable clusters faster than hierarchical techniques. In future we would like to extend this algorithm to work well with regular geometrical patterns and to derive a better way to find the optimal value for \( t \).
REFERENCES

[1] Hierarchical Clustering of Gene Expression Data Feng Luo, Kun Tang and Latifur Khan Department of Computer Science University of Texas at Dallas Richardson, TX 75083-0688.
