International Journal of Information Sciences and Application.
ISSN 0974-2255 Volume 5, Number 1 (2013), pp. 33-38

http://www.irphouse.com

Cluster Analysis on Statistical Data using Agglomerative Method

Dr. M. Manimekalai[**], Mrs. M. Anusha[*] and Mrs. G. SriNaganya*

[**]*Director, Department of MCA, Shrimati Indira Gandhi College, Tiruchirappalli, Tamil Nadu, India.*
[*]*Programmer, Department of MCA, Shrimati Indira Gandhi College, Tiruchirappalli, Tamil Nadu, India.*

Abstract

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. Data types in Cluster Analysis[1] are 1.Data Matrix (or object-by-variable structure), 2.Interval-Scaled Variables, 3.Binary Variables, 4.A Categorical Variable, 5.A Discrete ordinal Variable, 6.A Ratio-scaled Variable. Methods used in Clustering: 1.Partitioning Method, 2.Hierarchical Method, 3.Data Density based Method, 4.Grid based Method and 5. Model Based Method. There are two types of Hierarchical Methods[2] in clustering namely Agglomerative hierarchical clustering and Divisive hierarchical clustering.

The scope of this paper is to start out with n clusters for n data points, that is, each cluster consisting of a single data point. One possible approach is to combine a partitioning method like K-means[3] with a agglomerative method. Using a measure of distance, at each step of the method, the method merges two nearest clusters, thus reducing the number of clusters and building successively larger clusters. The process continues until the required number of clusters have been obtained or all the data points are in one cluster.

Keywords: Clusters, Partitioning, Hierarchical, Data Density, Grid

1. Introduction

Cluster analysis groups objects (observations, events) based on the information found in the data describing the objects or their relationships. The goal is that the objects in a group will be similar (or related) to one other and different from (or unrelated to) the objects in other groups. The greater the similarity (or homogeneity) within a group,

and the greater the difference between groups, the “better” or more distinct the clustering.

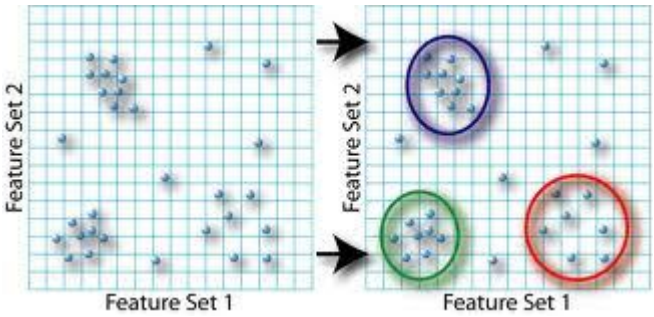

Figure 1: Data Clustering.

To better understand the difficulty of deciding what constitutes a cluster, consider Figures 1. Cluster analysis is a classification of objects from the data, where by classification we mean a labeling of objects with class (group) labels. As such, clustering does not use previously assigned class labels, except perhaps for verification of how well the clustering worked. Thus, cluster analysis is distinct from pattern recognition or the areas of statistics known as discriminate analysis. Decision analysis[4], which seek to find rules for classifying objects for a given set of pre-classified objects. While cluster analysis can be useful in the previously mentioned areas, either directly or as a preliminary means of finding classes, there is much more to these areas than cluster analysis. For example, the decision of what features to use when representing objects is a key activity of fields such as pattern recognition. Cluster Analysis typically takes the features as given and proceeds from there.

2. Agglomerative Method

The agglomerative method is basically a bottom-up approach which involves the flowing steps. An implementation however may include some variation of these steps.

1. Allocate each point to a cluster of its own. Thus we start with n clusters for n objects.
2. Create a distance matrix by computing distances between all pairs of clusters either using, for example, the single-link metric or the complete-link metric. Some other metric may also be used. Sort these distances in ascending order.
3. Find the two clusters that have the smallest distance between them.
4. Remove the pair of objects and merge them.
5. If there is only one cluster left then stop.
6. Compute all distances from the new cluster and update the distance matrix after the merger and go to Step 3.

3. Experimental Results

We now use agglomerative clustering for clustering the data on the following Cluster analysis of training data Example. We will use the centroid method[5] for computing the distances between clusters.

Table 1: Training Data Table

Student	Age	Mark 1	Mark 2	Mark 3
S1	18	73	75	57
S2	18	79	85	75
S3	23	70	70	52
S4	20	55	55	55
S5	22	85	86	87
S6	19	91	90	89
S7	20	70	65	60
S8	21	53	56	59
S9	19	52	82	60
S10	47	75	76	77

4. Step 1 and 2

Allocate each point to a cluster and compute the distance matrix[6] using the centroid method. The distance matrix is symmetric, so we only show half of it in the Table 1. the last column would be for S10 but the diagonal element is zero.

Table 2: Distance matrix for data.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1	0									
S2	34	0								
S3	18	52	0							
S4	42	76	36	0						
S5	57	23	67	95	0					
S6	66	32	82	106	15	0				
S7	18	46	16	30	65	76	0			
S8	44	74	40	8	91	104	28	0		
S9	20	22	36	60	37	46	30	115	0	
S10	52	44	60	90	55	70	60	98	58	0

The above matrix gives the distance of each object with every other object.

5. Step 3 and 4

The smallest distance is 8 between object S4 and object S8. They are combined and removed and we put the combined cluster (C1) where the object S4 was.

The following table is now the new distance matrix. All distances except those with cluster C1 remain unchanged.

Table 3: Distance matrix after merging S4 and S8 into cluster C1.

	S1	S2	S3	C1	S5	S6	S7	S9
S1	0							
S2	34	0						
S3	18	52	0					
C1	41	75	38	0				
S5	57	23	67	93	0			
S6	66	32	82	105	15	0		
S7	18	46	16	29	65	76	0	
S9	20	22	36	59	37	46	30	0
S10	52	44	60	88	55	70	60	58

6. Step 5 and 6

The smallest distance now is 15 between objects S5 and S6. They are combined in a cluster and S5 and S6 are removed.

7. Step 3, 4, 5 and 6

The following table is the updated distance matrix.

Table 4: Distance matrix after merging S5 and S6 into cluster C2.

	S1	S2	S3	C1	C2	S7	S9
S1	0						
S2	34	0					
S3	18	52	0				
C1	41	75	38	0			
C2	61.5	27.5	74.5	97.5	0		
S7	18	46	16	29	69.5	0	
S9	20	22	36	59	41.5	30	0
S10	52	44	60	88	62.5	60	58

We find the shortest distance again.

S3 and S7 are at a distance 16 apart. We merge them and put C3 where S3 was. Although we do not show any further distance matrices, the next step will show that C3 and S1 have the smallest distance. They are then merged and the process continues.

The result of using the agglomerative method could be something like that show in the below figure. The Cluster numbers show an example sequence of cluster mergings. The first one being merging of object S4 and S8 followed by S5 and S6, and so on. As noted earlier, it is possible to use any measure of distance between cluster centroids. Figure 2 shows the overall result of the above method using the hierarchical approach.

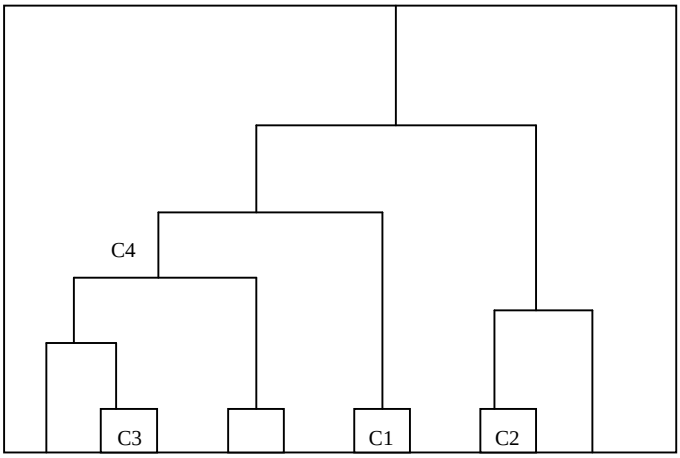

Figure 2: A possible result of using the agglomerative method.

Conclusion

The agglomerative method can be used to better understand the data and helping estimating the number of clusters and the starting seeds. The strength of the cluster analysis is that it will work well with numeric data. The agglomerative method leads to hierarchical clusters in which at each step we build larger and larger clusters that include increasingly dissimilar objects. In Future, a hybrid technology can be adopted by combining agglomerative method with density-based method[7], which leads to better accuracy results.

References

[1] H. Charles Remesburg, “Cluster Analysis For Researchers”, 2004, LULU Press, North Carolina.

[2] Andrew Gelman, Jennifer Hill, “Data Analysis using Regression and Multilevel/ Hierarchical Models”, 2007, Cambridge University Press, NewYork.

[3] Vance Faber, “Clustering and the Continuous k-Means Algorithm”, 1994, Los Alamos Science Number 22.

[4] Andrew Gelman, Matilade Trevisani, Hao Lu and Alexander van Geen, “Direct Data Manipulatin for Local Decision Analysis as Applied to the Problem of Arsenic in Drinking Water from Tube Wells in Bangladesh”, 2004, Risk Analysis, Vol.24, No.6.

[5] D Mariamma, M Gowthami, N Sindhujaa , “New algorithm to get the initial centroids of clusters on multidimensional data”, IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 1, Issue 1, March, 2013

[6] Juha Vesanto and Mila Sulkava, “Distance Matrix Based Clustering of the Self-Organizing Map”, 2002, Neural Networks Research Centre, Helsinki University of Technology, P.O.Box 5400, FIN-02015 HUT, Finland.

[7] Parul Agarwal, M.Afshar Alam, Ranjit Biswas, “Analysing the agglomerative hierarchical Clustering Algorithm for Categorical Attributes”, June 2010, International Journal of Innovation, Management and Technology, Vol. 1, No. 2.