Iris Classification using Agglomerative Hierarchical Clustering Approach

Hnin Hnin Swe
University of Computer, Monywa, Myanmar
hninhninswe84@gmail.com

Abstract

One of the usages of cluster analysis is to understand unknown data, given the value of several selected features. There are a lot of methods have been proposed data understanding. This paper discusses one technique of data clustering which is called the agglomerative hierarchical clustering algorithm. More specifically, this paper applies that algorithm to understand Iris plant data given the size of the sepal and petal. The system results the Iris flowers groups according to the given threshold value. Moreover, the system produces the analysis of the clusters in order to identify the optimal clusters. The proposed system is implemented as the generalized version of the hierarchical agglomerative clustering algorithm, applied on the Iris data set. But this system can be executed on any data set which would like to cluster the objects by means of that algorithm.

Keywords: Data Mining, Clustering, Hierarchical Partitioning, Agglomerative Method, Iris Data Set

1. Introduction

Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Representing data by fewer clusters necessarily loses certain fine details (akin to loss data compression), but achieves simplification. It represents many data objects by few clusters, and hence, it models data by its clusters. Clustering is a useful exploratory tool for multi-dimensional data analysis. When the underlying structure of the data is not readily apparent (i.e., the class structure or the number of groups in the data are unknown), cluster analysis may be applied to uncover this knowledge. Clustering schemes require data objects to be defined in terms of a predefined set of features. Features represent properties of the object that are relevant to the problem solving task. Clustering methods break the observation into distinct non overlapping groups. There are many different clustering methods. This unit will illustrate one of them, hierarchical clustering which contain agglomerative and divisive algorithm. This paper is used agglomerative algorithm.

Iris plant is widely distributed throughout the North Temperate Zone. Their habitats are considerably varied, ranging from cold regions into the grassy slopes, meadowlands, stream banks and deserts of Europe, the Middle East and northern Africa, Asia and across North America. This plant has a unique shape that can be distinguished with their three colored sepals, and then three, sometimes reduced, petals stand upright, partly behind the sepal bases. The sepals and the petals differ from each other. They are united at their base into a floral tube, that lies above the ovary. There are six subgenus of Iris, i.e. Iris, Limniris, Xiphium, Nepalensis, and Scorpiris. All species of Iris discussed in this paper are belonged to subgenus Limniris.

1.1. Iris setosa

This species grown in Alaska, the Youkon, Japan, China and Northern Asia. The bloom period is from June to August. This flower has various colors of petals, and can be distinguished to others by its small-sized petal, which is normally less than 2 cm. Another interesting characteristic is the leaf, which is 1.5 to 2.5 cm width and 30 to 60 cm length.

1.2. Iris virginica

Iris virginica is common along the coastal plain from Florida to Georgia. The flower consists of 3 horizontal sepals and 3 erect petals, which is relatively large (4 to 6 cm in diameter). Each plant has 2 to 6 flowers that bloom from April to May upon a single, erect, 30 to 90 cm tall stalk. Leaves are 1 to 3 cm wide and are sometimes longer than the flower stalk.

1.3. Iris versicolor

The native distribution of Iris versicolor spans from Newfoundland to Manitoba, south to Florida and Arkansas. They bloom from May to July. The three petaled flowers are often finely variegated with yellow, green, and white. Iris versicolor's sword-like leaves emerge from thick horizontal root stock which are covered with fibrous roots. This flower will grow to heights of 120 cm in spreading clumps. The individual leaves are somewhat shorter than the entire plant.

There are several features that discriminate Iris. The simplest and most commonly used is the flower, which is the size of sepal and petal. Other possible features are the leaf, the stalk, and the size of whole plant. Since the available data are only the flower's size, we summarize the sample data set as follows.

Table 1. Size information of Iris flowers

Sepal (length)	Sepal (width)	Petal (length)	Petal (width)	Class label
5.1	3.5	1.4	0.2	Iris-Setosa
7.1	3.0	5.9	2.1	Iris-Virginica
5.4	3.4	1.5	0.4	Iris-Setosa
6.4	3.2	4.5	1.5	Iris-Versuicoloor
6.3	3.3	4.7	1.6	Iris-Versuicoloor
7.3	2.9	6.3	1.8	Iris-Virginica
4.4	2.9	1.4	0.2	Iris-Setosa
4.99	3.1	1.5	0.1	Iris-Setosa
5.8	2.8	5.1	2.4	Iris-Virginica
5.6	2.9	3.6	1.3	Iris-Versuicoloor
5.9	3.2	5.7	2.3	Iris-Virginica
5.0	3.4	4.5	1.6	Iris-Versuicoloor
7.2	3.0	5.8	1.6	Iris-Virginica
4.8	3.4	1.9	0.2	Iris-Setosa

2. Related work

A.K.Poernomo applied C4.5 algorithm for classifying Iris data. In his experiment, the accuracy was quite promising which is ranged between 94% and 96%. However, even the simplest technique, i.e. OneR, could have better accuracy. The main weakness is that his algorithm can only divide the space in orthogonal space. Therefore, even if the objects are linearly separable, this algorithm might not be able to separate them. That algorithm also tried to do binary split as much as possible, which is not always to be the best. In this data, it is seen that even OneR can classify more accurately. He also tried to overcome this weakness by combining this technique with PCA. However, the result turns out not to be promising. Other possible methods to enhance this technique is to kernelize the data, or to use other techniques which makes the data to be linearly separable in orthogonal space. Another possible way is to allow non-orthogonal split [12].

3. Agglomerative clustering

Cluster analysis groups data objects based on information found in the data that describes the objects and their relationships. The goal is that the objects within a groups be similar (or related) to one another 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. Cluster analysis is related to other techniques that are used to divide data objects into groups. For instance, clustering can be regarded as a form of classification in that it creates a labeling of objects with class (cluster) labels. However, it derives these labels only from the data. In contrast, classification in the sense is supervised classification; i.e., new, unlabeled objects are assigned a class label using a model developed from objects with known class labels [5].

Hierarchical clustering based on linkage metrics results in clusters of proper (convex) shapes. Hierarchical clustering frequently deals with the matrix of distances (dissimilarities) or similarities between training points. It is sometimes called connectivity matrix. Linkage metrics are constructed (see below) from elements of this matrix.

In hierarchical clustering the goal is to produce a hierarchical series of nested clusters, ranging from clusters of individual points at the bottom to an all-inclusive cluster at the top. Hierarchical clustering techniques proceed by either a series of successive mergers or a series of successive divisions. Hierachical methods result in a nested sequence of clusters which can be graphically represented with a tree, called a dendrogram. We can distinguish between two main types of hierarchical clustering algorithms: agglomerative algorithms and divisive algorithms.

A diagram is called a dendogram graphically represents this hierarchy and is an inverted tree that describes the order in which points are merged (bottom-up view) or clusters are split (top-down view). One of the attractions of hierarchical techniques is that they correspond to taxonomies that are very common in the biological sciences, e.g., kingdom, phylum, genus, species (Some cluster analysis work occurs under the name of “mathematical taxonomy.”). Another attractive feature is that hierarchical techniques do not assume any particular number of clusters. Instead any desired number of clusters can be obtained by “cutting” the dendogram at the proper level. Finally, hierarchical techniques are thought to produce better quality clusters [6].

Hierarchical clustering is an iterative procedure in which n data points are partitioned into groups which may vary from a single cluster containing all n points, to n clusters each containing a single point. Hierarchical clustering techniques can be divided into "agglomerative" and "divisive" methods. In the former, clusters initially containing one element each are successively fused to generate larger clusters. At each step, the clusters to be fused are those that are, according to some predefined metric, most similar ("closest") to each other. In the latter, a large cluster is divided into successively smaller clusters. Both fusions and divisions are irreversible.

Agglomerative (bottom-up) methods start with the individual objects. This means that they initially consider each object as belonging to a different cluster. The most similar objects are first grouped, and these initial groups are then merged according to their similarities. Eventually, all sub groups are fused into a single cluster. In order to merge clusters we need to define the similarity between our merged cluster and all other clusters. There are several methods to measure the similarity between clusters. These methods are sometimes referred to as linkage methods. There does not seem to be consensus as to which is the best method. Below we define four widely used rules for joining clusters. In the lecture we will explain each method at the hand of a simple example.

3.1 Algorithm of the agglomerative clustering

1. Assign each object to a separate cluster.
2. Repeat
3. (Re) evaluate all pair-wise distances between clusters (e.g. Euclidean distance measures).
4. Construct a distance matrix using the distance values.
5. Look for the pair of clusters with the shortest distance.
6. Remove the pair from the matrix and merge t hem .
7. Until the distance matrix is reduced to a single Element or terminate condition.

The distance between two points defined as the square root of the sum of the squares of the differences between the corresponding coordinates of the points; for example, in two-dimensional Euclidean geometry, the Euclidean distance between two points **a** = (**a**$_x$, **a**$_y$) and **b** = (**b**$_x$, **b**$_y$) is defined as:

$$d(\mathbf{a},\mathbf{b}) = \sqrt{(\mathbf{a}_x - \mathbf{b}_x)^2 + (\mathbf{a}_y - \mathbf{b}_y)^2}$$

Hierarchical clustering initializes a cluster system as a set of singleton clusters (agglomerative case) or a single cluster of all points (divisive case) and proceeds iteratively with merging or splitting of the most appropriate cluster(s) until the stopping criterion is achieved. The appropriateness of a cluster(s) for merging/splitting depends on the (dis) similarity of cluster(s) elements. This reflects a general presumption that clusters consist of similar points. An important example of dissimilarity between two points is the distance between them. Other proximity measures are discussed in the section General Algorithm Issues.

To merge or split subsets of points rather than individual points, the distance between individual points has to be generalized to the distance between subsets. Such derived proximity measure is called a linkage metric. The type of the linkage metric used significantly affects hierarchical algorithms, since it reflects the particular concept of closeness and connectivity. Major inter-cluster linkage metrics [7, 8] include single link, average link, and complete link. The underlying dissimilarity measure (usually, distance) is computed for every pair of points with one point in the first set and another point in the second set. A specific operation such as minimum (single link), average (average link), or maximum (complete link) is applied to pair-wise dissimilarity measures.

There are three components that primarily govern the clustering process:

- distance or similarity metric,
- the control algorithm, and
- the criterion function.

The distance metric typically uses an objective measure, e.g., the Euclidean distance or the Mahalanobis distance, to define the proximity between pairs of objects in the data set. The control algorithm can be agglomerative, where the partition structure is constructed bottom-up through successive merging of atomic clusters into larger groups, or divisive, where clusters are formed top-down from one large cluster that includes all of the data objects, and then is successively subdivided into smaller groups. The criterion function is used for evaluating the goodness of a partition structure once it is formed; typically the mean square error is used in numeric partitioning schemes. Regardless of the distance metric, control algorithm or criterion function, the most defining characteristic of a clustering scheme is the type of data the scheme is intended to cluster. For static data clustering, e.g., each data is represented as a vector of feature values with one value per feature, clustering schemes have been developed for numeric, nominal and mixed data types

4. Proposed system design

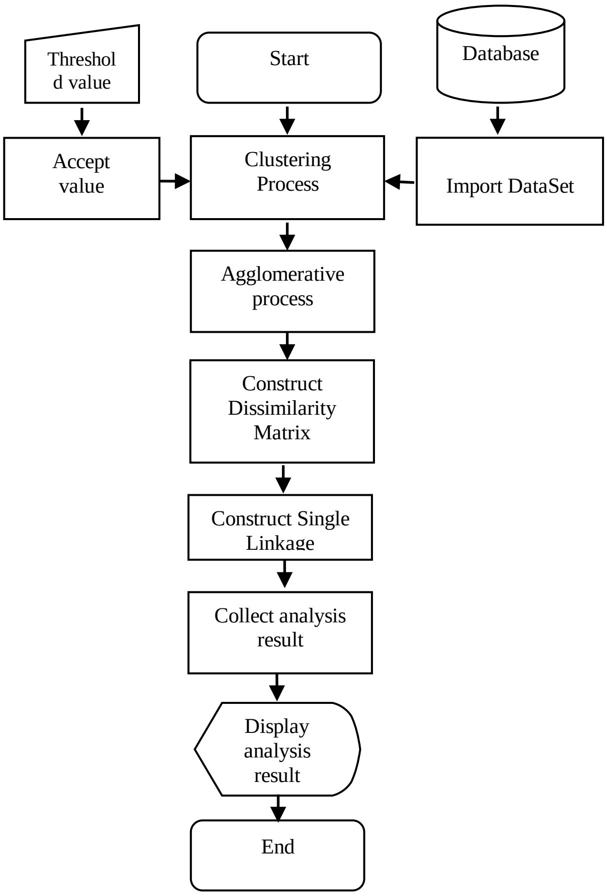


Figure 1. System flow diagram

5. Implementation

Here are six different tabs in the main menu. These are File, Edit, Analyze, Calculate, Algorithm and Help.

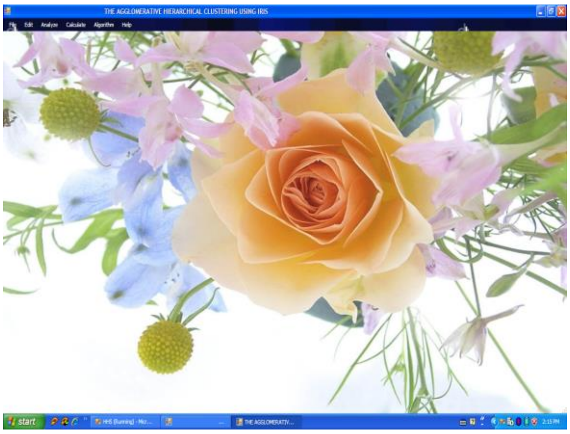
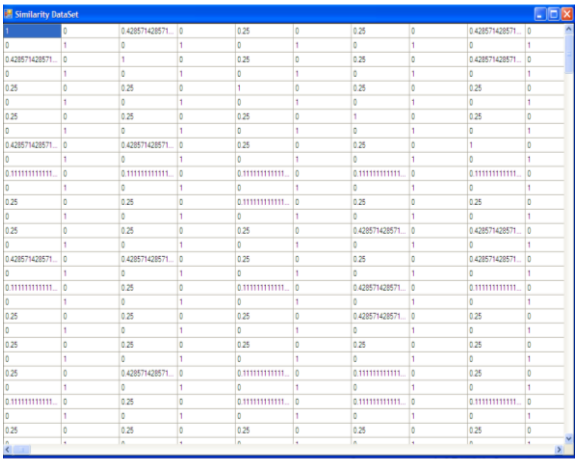


Figure 2. Main menu

5.1. The similarity of data set

The following frame implements the calculation of the similarity of objects from the data set based on the Euclidian distance. According to the agglomerative hierarchical clustering algorithm, the matrix is the NxN objects from Iris data set. Thus, this view shows the calculated similarity values of each objects form the data set.

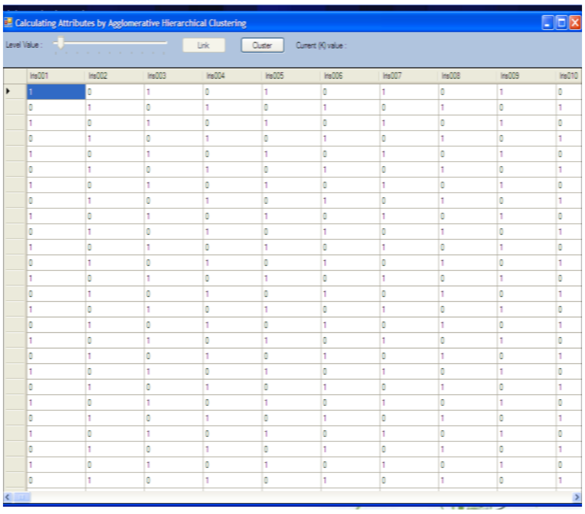

Similarity DataSet

1	0	0.428571428571...	0	0.25	0	0.25	0	0.428571428571...	0
0	1	0	1	0	1	0	1	0	1
0.428571428571...	0	1	0	0.25	0	0.25	0	0.428571428571...	0
0	1	0	1	0	1	0	1	0	1
0.25	0	0.25	0	1	0	0.25	0	0.25	0
0	1	0	1	0	1	0	1	0	1
0.25	0	0.25	0	0.25	0	1	0	0.25	0
0	1	0	1	0	1	0	1	0	1
0.428571428571...	0	0.428571428571...	0	0.25	0	0.25	0	1	0
0	1	0	1	0	1	0	1	0	1
0.111111111111...	0	0.111111111111...	0	0.111111111111...	0	0.111111111111...	0	0.111111111111...	0
0	1	0	1	0	1	0	1	0	1
0.25	0	0.25	0	0.111111111111...	0	0.25	0	0.25	0
0	1	0	1	0	1	0	1	0	1
0.25	0	0.25	0	0.25	0	0.428571428571...	0	0.428571428571...	0
0	1	0	1	0	1	0	1	0	1
0.428571428571...	0	0.428571428571...	0	0.25	0	0.25	0	0.428571428571...	0
0	1	0	1	0	1	0	1	0	1
0.111111111111...	0	0.25	0	0.111111111111...	0	0.428571428571...	0	0.111111111111...	0
0	1	0	1	0	1	0	1	0	1
0.25	0	0.25	0	0.25	0	0.428571428571...	0	0.25	0
0	1	0	1	0	1	0	1	0	1
0.25	0	0.25	0	0.25	0	0.25	0	0.25	0
0	1	0	1	0	1	0	1	0	1
0.25	0	0.428571428571...	0	0.111111111111...	0	0.111111111111...	0	0.25	0
0	1	0	1	0	1	0	1	0	1
0.111111111111...	0	0.25	0	0.111111111111...	0	0.111111111111...	0	0.111111111111...	0
0	1	0	1	0	1	0	1	0	1
0.25	0	0.25	0	0.25	0	0.25	0	0.25	0

Figure 3. Similarity dataset form

5.2. The agglomerative hierarchical clustering algorithm

This frame implements the agglomerative hierarchical clustering algorithm using Iris data set. The input parameter of the application is the level value of the similarity (or) the threshold value of the similarity which is defined by the user. Then, link this value to the data set in order to cluster the data object in this analysis.



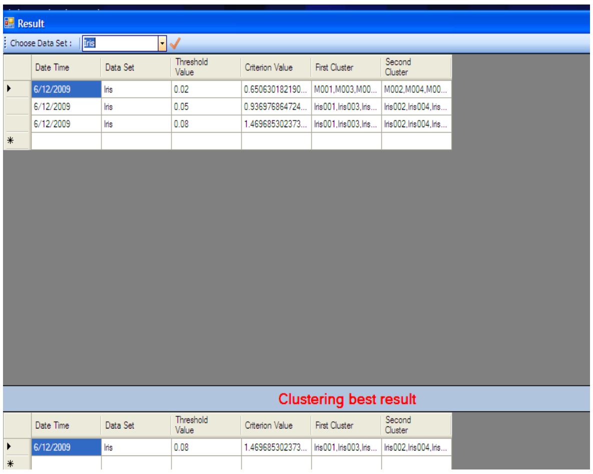
Calculating Attributes by Agglomerative Hierarchical Clustering
Level Value :
Link
Cluster
Current (K) value :

	Iris001	Iris002	Iris003	Iris004	Iris005	Iris006	Iris007	Iris008	Iris009	Iris010
▶	1	0	1	0	1	0	1	0	1	0
	0	1	0	1	0	1	0	1	0	1
	1	0	1	0	1	0	1	0	1	0
	0	1	0	1	0	1	0	1	0	1
	1	0	1	0	1	0	1	0	1	0
	0	1	0	1	0	1	0	1	0	1
	1	0	1	0	1	0	1	0	1	0
	0	1	0	1	0	1	0	1	0	1
	1	0	1	0	1	0	1	0	1	0
	0	1	0	1	0	1	0	1	0	1
	1	0	1	0	1	0	1	0	1	0
	0	1	0	1	0	1	0	1	0	1
	1	0	1	0	1	0	1	0	1	0
	0	1	0	1	0	1	0	1	0	1
	1	0	1	0	1	0	1	0	1	0
	0	1	0	1	0	1	0	1	0	1
	1	0	1	0	1	0	1	0	1	0
	0	1	0	1	0	1	0	1	0	1
	1	0	1	0	1	0	1	0	1	0
	0	1	0	1	0	1	0	1	0	1
	1	0	1	0	1	0	1	0	1	0
	0	1	0	1	0	1	0	1	0	1
	1	0	1	0	1	0	1	0	1	0
	0	1	0	1	0	1	0	1	0	1
	1	0	1	0	1	0	1	0	1	0
	0	1	0	1	0	1	0	1	0	1

Figure 4. Calculating attributes by agglomerative hierarchical clustering form

5.3. The result form of the analysis

The analysis results are shown in the following figure. There are two separate sections in this figure. The first section is the results of the agglomerative hierarchical clustering process. The records will be added at each time of the process execution. The second section is the best clustering results of the analysis among the executions.



Choose Data Set : Iris

	Date Time	Data Set	Threshold Value	Criterion Value	First Cluster	Second Cluster
▶	6/12/2009	Iris	0.02	0.650630182190...	M001,M003,M00...	M002,M004,M00...
	6/12/2009	Iris	0.05	0.936976864724...	Iris001,Iris003,Iris...	Iris002,Iris004,Iris...
	6/12/2009	Iris	0.08	1.469685302373...	Iris001,Iris003,Iris...	Iris002,Iris004,Iris...
*						

Clustering best result

	Date Time	Data Set	Threshold Value	Criterion Value	First Cluster	Second Cluster
▶	6/12/2009	Iris	0.08	1.469685302373...	Iris001,Iris003,Iris...	Iris002,Iris004,Iris...
*						

Figure 5. Result Form

6. Conclusion

This paper implements a simple and efficient clustering algorithm based on the agglomerative (bottom-up) method. This algorithm is easy to implement, applying on an Iris data set. Moreover, the software is also analyzes the other data set such as Myanmar's Native Orchid, Car data set and Industry data set to understand the nature of the cluster analysis. This system also allows the users to view the process knowledge of agglomerative hierarchical clustering algorithm.

7. References

[1] J.Han and M.Kamber,"Data Mining Concept and Techniques", Morgan Kaufmann, 2000.

[2] J.Han, A.Tung and K.Kamber, "Spatial Clustering Geographic Data Mining and Knowledge Discovery", Taylor and Francis, 21, 2001.

[3] E.Anderson, "The Irises of the Gasp´e peninsula", Bulletin of the American Iris Society, 1935.

[4] R.C.Dubes and A.K.Jain, "Algorithm for Clustering Data", Prentice Hall Upper Saddle River, N.J., 1988.

[5] R.A.Fisher,"The use of multiple measurements in taxonomic problems", Annals of Eugenics, 1936.

[6] S. Guha, R. Rastogi, and K. Shim, "CURE: An efficient clustering algorithm for large databases", In Proceeding of the 1998 ACM SIGMOD International Conference on Management of Data Engineering, 1998.

[7] F.Murtach, "Multidimensional Clustering Algorithms", Physica-Verlag, Vienna, 1985.

[8] C. Olson, "Parallel algorithms for hierarchical clustering", Parall Computing, 1995.

[9] T.Zhang, R.Ramakrishnan and M.Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases", Proceeding of the ACM SIGMOD Record, 1996.