On the Two-level Hybrid Clustering Algorithm

Eng Yeow Cheu, Chee Keong Kwoh, Zonglin Zhou
Bioinformatics Research Centre,
School of Computer Engineering,
Nanyang Technological University, Singapore 639798
ezlzhou@ntu.edu.sg

ABSTRACT

In this paper, we design the hybrid clustering algorithms, which involve two level clustering. At each of the levels, users can select the k-means, hierarchical or SOM clustering techniques. Unlike the existing cluster analysis techniques, the hybrid clustering approach developed here represents the original data set using a smaller set of prototype vectors (cluster means), which allows efficient use of a clustering algorithm to divide the prototype into groups at the first level. Since the clustering at the first level provides data abstraction first, it reduces the number of samples for the second level clustering. The reduction of the number of samples, hence, the reduction of computational cost is especially important when hierarchical clustering is used in the second stage. The prototypes clustered at the first level are local averages of the data and therefore less sensitive to random variations than the original data. The empirical evaluation of the two-level hybrid clustering algorithms is made at four data sets

1. INTRODUCTION

Over the years, extensive research has been carried out in determining the optimal cluster analysis. Techniques for clustering have been developed very rapidly, spurred mostly by the availability of computers to carry out awesome calculations involved. These research efforts have resulted in a number of well-known algorithms, and variants are continuously being developed, each addressing specific shortcomings of their ancestors.

In this paper, three general methods are selected, namely (1) k-means, an iterative partitioning method, (2) agglomerative hierarchical clustering, a method that builds a hierarchical clustering tree from bottom-up, (3) Self-Organizing Map (SOM), a prominent unsupervised neural network model mapping high-dimensional data onto a two-dimensional plane. Our hybrid clustering techniques are designed based on them. Analysis of differences in performance of the three general methods and our hybrid clustering algorithms is also given.

2. CLUSTERING ALGORITHMS

There are many different algorithms that are available today, and the two of the algorithms that we investigate, fall into two general categories: hierarchical and nonhierarchical. The third is an unsupervised clustering method - SOM, used to find clusters in the input data, and identify an unknown data vector with one of the clusters [1].

2.1. HIERARCHICAL CLUSTERING PROCEDURE

There are basically two types of hierarchical clustering procedures – agglomerative and divisive. In agglomerative hierarchical methods, each observation starts out as its own cluster. In subsequent steps, the two closest clusters are combined into a new aggregate cluster, thus reducing the number of clusters by one in each step. Two groups of individuals formed at an earlier stage may join together in a new cluster. Eventually, all individuals are fused into one large cluster.

In divisive methods, an initial single group of objects is divided into two subgroups such that the objects in one subgroup are “far from” the objects in the other. These subgroups are then further divided into dissimilar subgroups; the process continues until there are as many subgroups as objects (each object forms a cluster).

In both hierarchical methods, a hierarchy of a tree-like structure is constructed and usually represented as a dendrogram or tree graph. The dendrogram illustrates the mergers or divisions that have been made at successive levels.

In particular, Wishart [6] contends that the “top down” decision tree approach has inherently greater risk of mis-classification by inefficiently splitting on a single variable than the “bottom up” approach. Each classification generated in a decision tree is univariate by definition, and this limits the range of possible segments available for consideration. By comparison, the agglomerative approach is multivariate and exploratory, and allows for more feasible segments to be investigated in terms of the actual distribution of the scatter. Hence, this project concentrates on agglomerative hierarchical algorithms mainly (divisive methods act almost as agglomerative methods in reverse).

The following are the steps in the agglomerative hierarchical clustering algorithm for grouping N objects:

1. Start with N clusters, each containing a single entity and an $N \text{ x } N$ symmetric matrix of distances (or similarities) $D = \{d_{jk}\}$.

2. Search the distance matrix for the nearest (most similar) pair of clusters. Let the distance between "most similar" clusters U and V be D_{uv}.

1. Merge clusters U and V. Label the newly formed cluster (UV). Update the entries in the distance matrix by
   a. deleting the rows and columns corresponding to clusters U and V and
   b. adding a row and column giving the distances between cluster (UV) and the remaining clusters.

Repeat Steps 2 and 3 a total of $N-1$ times. (All objects will be in a single cluster after the algorithm terminates.) Record the identity of clusters that are merged and the levels (distances or similarities) at which the mergers take place.

2.2. Variations of Hierarchical Algorithm

This section describes the various variants of agglomerative hierarchical clustering algorithms - single linkage, complete linkage, average linkage and Ward's method (ESS).

2.2.1. Linkage Methods

The inputs to a linkage algorithm can be distances or similarities between pairs of objects. Single linkage, complete linkage and average linkage are the three linkage-based hierarchical clustering algorithms implemented.

Table 1: Between-clusters distances

Between-clusters distance	$d(Q_k, Q_l)$
Single linkage	$d_s = \min_{i,j}\{\lVert x_i - x_j \rVert\}$
Complete linkage	$d_c = \max_{i,j}\{\lVert x_i - x_j \rVert\}$
Average linkage	$d_a = \dfrac{\Sigma_{i,j}\lVert x_i - x_j \rVert}{N_k N_l}$

Between-clusters distance $d(Q_k, Q_l)$; $x_i \in Q_k, x_j \in Q_l, k \neq l$. N_k is the number of samples in cluster Q_k.

Table 1 shows the between-clusters distance definition for each of the linkage methods. In this case, dissimilarity coefficient is employed. The selection of the distance criterion or similarity coefficient depends on application.

Single Linkage: Groups are formed from the individual entities by merging nearest neighbours, where the term nearest neighbour connotes the smallest distance or largest similarity.

Complete Linkage: The distance (similarity) between clusters is determined by the distance (similarity) between the two elements, one from each cluster, which are most distant (or least similar).

Average Linkage: Average linkage treats the distance between two clusters as the average distance between all pairs of items where one member of a pair belongs to each cluster.

2.2.2 Ward's Method (Euclidean Sum of Squares)

In Ward's method, the distance between two clusters is the sum of squares between the two clusters summed over all variables. At each stage in the clustering procedure, the within-cluster sum of squares is minimized over all partitions obtainable by combining two clusters from the previous stage.

The Euclidean Sum of Squares (ESS), E_p, for a cluster p is given by:

$$E_p = \frac{\sum_{i \in p} c_i \sum_j w_j (x_{ij} - \mu_{pj})^2}{\sum_j w_j}$$

where x_{ij} is the value of variable j in case i within cluster p, c_i is an optional differential weight for case i, w_j is an optional differential weight for variable j, and μ_{pj} is the mean of variable j for cluster p.

The total ESS for all clusters p is $E = \sum_p E_p$ and the increase in the Euclidean Sum of Squares $I_{p \cup q}$ at the union of two clusters p and q is:

$$I_{p \cup q} = E_{p \cup q} - E_p - E_q$$

Ward considers hierarchical clustering procedures based on minimizing the 'loss of information' from joining two groups. This method is usually implemented with loss of information taken to be an increase in an error sum of squares criterion. At each step, union of every possible pair of clusters is considered, and the two clusters whose combination results in the smallest increase in ESS are joined.

2.2. NONHIERARCHICAL CLUSTERING PROCEDURE

Nonhierarchical procedures do not involve the tree-like construction process. Instead, these methods assign objects into clusters once the number of clusters to be formed is specified. The number of clusters may be either be specified in advance or determined as part of the clustering procedure. Nonhierarchical methods start from either from (1) an initial partition of items into groups or (2) an initial set of seed points, which will form the nuclei of clusters.

Nonhierarchical clustering procedures are frequently referred to as K-means clustering. MacQueen [5] suggests the term *K-means* for describing an algorithm of his that assigns each item to the cluster having the nearest centroid (mean). In its simplest form, the process is composed of three steps:

1. Partition the items into k initial clusters. (or specify k initial centroids (seed points))

2. Proceed through the list of items, assigning an item to the cluster who centroid (mean) is nearest. (Distance is usually computed using Euclidean distance with either standardized or unstandardized observations.)
Recalculate the centroid for the cluster receiving the new item and for the cluster losing the item.

3. Repeat Step 2 until no more reassignments take place.

Because a matrix of distances (similarities) does not have to be determined, and the basic data do not have to be stored during the computer run, nonhierarchical methods can be applied to larger data sets than can hierarchical techniques.

2.3. SELF-ORGANIZING MAP (SOM)

The Self-Organizing Map (SOM) is an unsupervised neural network mapping high dimensional input data onto a usually two-dimensional output space while reserving relations between the data items. The cluster structure within the data as well as the inter-cluster similarity is visible from the resulting topology preserving mapping [3, 4].

The SOM consists of units (neurons), which are arranged as a two-dimensional rectangular or hexagonal grid. During the training process vectors from the data set are presented to the map in random order. The unit most similar to a chosen vector is selected as the winner and adopted to match the vector even better. Then units in the neighborhood of the winner are slightly adopted as well. The trained SOM provides a mapping of the data space onto a two-dimensional plain in such a way that similar data points are located close to each other.

3. THE EMPIRICAL STUDY

In the empirical section, the software for all the clustering algorithms evaluated in this paper is available at [2].

Data set 1 is artificially generated to see how the algorithms perform when there are two well-separated but non-homogeneous clusters. The hybrid approach on data set 1 is performed using Ward's hierarchical clustering and single linkage hierarchical clustering. During the first stage of the hybrid approach, Ward's method is used to find ten smaller clusters on the standardized data set 1. As can be seen from Figure 1, ten small clusters are found. No smaller cluster is formed with elements in both elongated clusters of data set 1.

During the 2nd stage single linkage hierarchical clustering, cluster analysis is performed on the ten cluster means. The cluster means are treated as new input vectors to the 2nd stage. This hybrid approach utilizes the property of Ward's method and single linkage hierarchical clustering. Ward's method tends to find relatively equal sizes and hyper-spherical clusters whereas single linkage clustering tends to form long elongated cluster. In this test by combining the features of both clustering methods, the two elongated clusters of data set 1 are found in Figure 2.

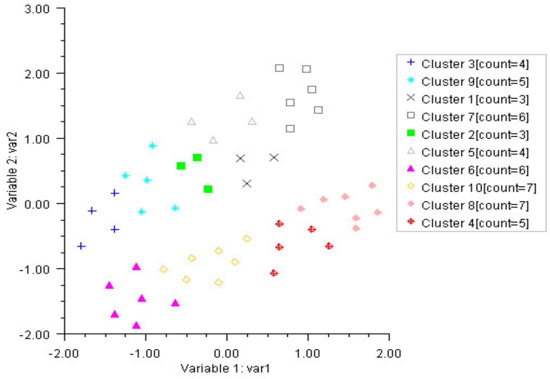

Figure1: Result after 1st stage Ward's hierarchical clustering on data set 1

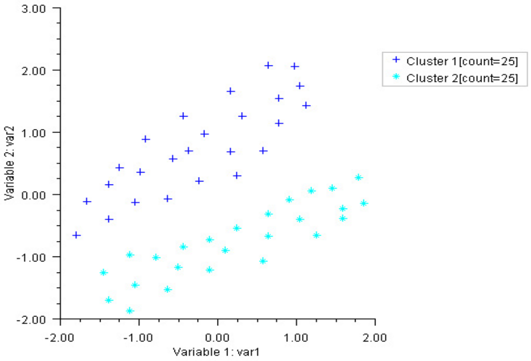

Figure2: Result after 2nd stage single linkage hierarchical clustering on data set 1

Data set 2 contains three classes of 50 instances each, where each class refers to a type of iris plant. Each instance has four continuous attributes. One class is linearly separable from the other two; the latter are not linearly separable from each other. Table 2 summarizes the results achieved by each of the clustering techniques carried out in this experimental setup, including two two-level hybrid clustering algorithms.

Table 2: Results of the clustering techniques on raw data set 2

Clustering Method	Percentage of samples correctly clustered
K-means	89.3%
Single linkage	68%
Complete linkage	96%
Average linkage	74%
Ward's method	89.3%
Hybrid 1st stage - SOM 2nd stage – K-means	92.6%
Hybrid 1st stage – K-means 2nd stage – Complete linkage	82%

Data set 3 contains two classes of 690 samples. In this dataset, there is a good mix of attributes: continuous, nominal with small numbers of values, and nominal with larger numbers of values.

In Table 3, the results achieved by direct clustering on this data set using complete linkage, and average linkage hierarchical clustering technique are not as good as the result achieved using the hybrid approach. In this experiment setup, hybrid approach clustering utilising SOM and complete linkage hierarchical clustering achieves a better result than complete linkage clustering on data set 3. A better result is also achieved using hybrid approach clustering utilising SOM and average linkage hierarchical clustering than direct average linkage hierarchical clustering on data set 3.

Table 3: Results of the clustering techniques on data set 3

Clustering Method	Percentage of samples correctly clustered
K-means	84%
Single linkage	55%
Complete linkage	55%
Average linkage	55%
Ward's method	79%
Hybrid 1st stage – K-means 2nd stage – Single linkage	55%
Hybrid 1st stage – K-means 2nd stage – Complete linkage	55%
Hybrid 1st stage - SOM 2nd stage – Complete linkage	80%
Hybrid 1st stage - SOM 2nd stage – Average linkage	76%
Hybrid 1st stage - SOM 2nd stage – K-means	84%

Data set 4 contains two classes of samples where one class is the group of patients diagnosed positively for diabetes. Each sample has eight continuous attributes.

In this experiment setup, the results in Table 4 achieved by all the clustering techniques are about the same. There is a slight improvement using hybrid approach utilizing K-means clustering and complete linkage hierarchical clustering when it is compared to the result achieved using complete linkage hierarchical clustering on data set 4.

Table 4: Results of each of the clustering techniques on data set 4

Clustering Method	Percentage of samples correctly clustered
K-means	70%
Single linkage	65%
Complete linkage	67%
Average linkage	65%
Ward's method	66%
Hybrid 1st stage – K-means 2nd stage – Single linkage	65%
Hybrid 1st stage – K-means 2nd stage – Complete linkage	70%
Hybrid 1st stage - SOM 2nd stage – Complete linkage	63%
Hybrid 1st stage - SOM 2nd stage – Average linkage	65%
Hybrid 1st stage - SOM 2nd stage – K-means	65%

4. CONCLUSIONS

We compared on the four data sets the performance of the two-level hybrid clustering algorithms against the other clustering algorithms: k-mean, SOM, single linkage, complete linkage, average linkage, and Ward's hierarchical clustering. The two-level hybrid clustering algorithms hit the highest percentage of samples correctly clustered on all the data sets as compared to each of the other clustering algorithms alone. In particular, in data set 1, the hybrid approach using ward's method in the first stage and single linkage hierarchical clustering in the second stage is able to find the two well-separated non-homogeneous clusters of the data set, whereas other clustering methods, other than single linkage clustering, are not able to find the clusters for this type of data set.

REFERENCES

[1] M. S. Aldenderfer and R. K. Blashfield, *Cluster analysis.* Beverly Hills: Sage Publications, 1984.

[2] Clustan Clustering Software, Available: www.clustan.com, 2003.

[3] T. Kohonen, "Self-organizing maps: Optimization approaches", *In proceedings of the international conference on artificial neural networks,* Finland, pp. 981-990, 1991.

[4] T. Kohonen, J. Hynninen, J. Kangas, and J. Laaksonen, *The Self-Organizing Map Program Package*, Laboratory of Computer and Information Science, Helsinki University of Technology, 1995

[5] Macqueen, "Some methods for classification and analysis of multivariate observation", *Proc. 5th Berkeley Symp.,* I, pp. 281-297, 1967.

[6] D. Wishart, "Efficient hierarchical cluster analysis for data mining and knowledge discovery", Presented at *the Interface 1998.* Minneapolis, USA, 1998.