Clustering Analysis of Data with High Dimensionality

Athman Bouguettaya
CSIRO ICT Center, Australia

Qi Yu
Virginia Tech, USA

INTRODUCTION

Clustering analysis has been widely applied in diverse fields such as data mining, access structures, knowledge discovery, software engineering, organization of information systems, and machine learning. The main objective of cluster analysis is to create groups of objects based on the degree of their association (Kaufman & Rousseeuw, 1990; Romesburg, 1990).

There are two major categories of clustering algorithms with respect to the output structure: partitional and hierarchical (Romesburg, 1990). K-means is a representative of the partitional algorithms. The output of this algorithm is a flat structure of clusters. The K-means is a very attractive algorithm because of its simplicity and efficiency, which make it one of the favorite choices to handle large datasets. On the flip side, it has a dependency on the initial choice of number of clusters. This choice may not be optimal, as it should be made in the very beginning, when there may not exist an informal expectation of what the number of natural clusters would be. Hierarchical clustering algorithms produce a hierarchical structure often presented graphically as a dendrogram. There are two main types of hierarchical algorithms: agglomerative and divisive. The agglomerative method uses a bottom-up approach, i.e., starts with the individual objects, each considered to be in its own cluster, and then merges the clusters until the desired number of clusters is achieved. The divisive method uses the opposite approach, i.e., starts with all objects in one cluster and divides them into separate clusters. The clusters form a tree with each higher level showing higher degree of dissimilarity. The height of the merging point in the tree represents the similarity distance at which the objects merge in one cluster. The agglomerative algorithms are usually able to generate high-quality clusters but suffer a high computational complexity compared with divisive algorithms.

In this paper, we focus on investigating the behavior of agglomerative hierarchical algorithms. We further divide these algorithms into two major categories: group based and single-object based clustering methods. Typical examples for the former category include Unweighted Pair-Group using Arithmetic averages (UPGMA), Centroid Linkage, and WARDS, etc. Single LINKage (SLINK) clustering and Complete LINKage clustering (CLINK) fall into the second category. We choose UPGMA and SLINK as the representatives of each category and the comparison of these two representative techniques could also reflect some similarity and difference between these two sets of clustering methods. The study examines three key issues for clustering analysis: (1) the computation of the degree of association between different objects; (2) the designation of an acceptable criterion to evaluate how good and/or successful a clustering method is; and (3) the adaptability of the clustering method used under different statistical distributions of data including random, skewed, concentrated around certain regions, etc. Two different statistical distributions are used to express how data objects are drawn from a 50-dimensional space. This also differentiates our work from some previous ones, where a limited number of dimensions for data features (typically up to three) are considered (Bouguettaya, 1996; Bouguettaya & LeViet, 1998). In addition, three types of distances are used to compare the resultant clustering trees: Euclidean, Canberra Metric, and Bray-Curtis distances. The results of an exhaustive set of experiments that involve data derived from 50-dimensional space are presented. These experiments indicate a surprisingly high level of similarity between the two clustering techniques under most combinations of parameter settings.

The remainder of this paper is organized as follows. Section 2 discusses the clustering techniques used in our evaluation and describes the various distributions

used to derive our experimental data. Section 3 outlines the experimental methodology and Section 4 presents a summary of our results. Finally, concluding remarks are drawn in Section 5.

BACKGROUND

In this section, we outline a set of key elements for conducting clustering analysis. These include *distances of similarity, coefficients of correlation, clustering methods, and statistical distributions of data objects*. In what follows, we will give a detailed discussion of each of these elements. Finally, we present a general algorithm, which outlines the procedure of constructing clustering in our study.

Distances of Similarity

To cluster data objects in a database system or in any other environment, some means of quantifying the degree of associations between items is needed. This can be a measure of distances or similarities. There are a number of similarity measures available and the choice may have an effect on the results obtained. Multi-dimensional objects may use relative or normalized weight to convert their distance to an arbitrary scale so they can be compared. Once the objects are defined in the same measurement space as the points, it is then possible to compute the degree of similarity. In this respect, the smaller the distance the more similar two objects are. The most popular choice in computing distance is the Euclidean distance with:

$$d(i,j) = \sqrt{(x_{i_1} - x_{j_1})^2 + (x_{i_2} - x_{j_2})^2 + ... + (x_{i_n} - x_{j_n})^2} \quad (1)$$

Euclidean distance belongs to the family of Minkowski's distances, which is defined as

$$d(i,j) = (|x_{i_1} - x_{j_1}|^m + |x_{i_2} - x_{j_2}|^m + ... + |x_{i_n} - x_{j_n}|^m)^{\frac{1}{m}} \quad (2)$$

When m = 2, Minkowski's distance becomes Euclidean distance. Another widely used distance, called Manhattan distance, is also a special case of Minkowski's distance (when m is set to 1).

In addition to Euclidean distance, we also use another two types of distances to investigate how this element could affect clustering analysis: *Canberra Metric* and *Bray-Curtis distances*. Canberra Metric distance, a(i, j), has a range between 0.0 and 1.0. The data objects i and j are identical when a(i, j) takes value 0.0. Specifically, a(i, j) is defined as:

$$a(i,j) = \frac{1}{n}\left(\frac{|x_{i_1} - x_{j_1}|}{(x_{i_1} + x_{j_1})} + \frac{|x_{i_2} - x_{j_2}|}{(x_{i_2} + x_{j_2})} + ... + \frac{|x_{i_n} - x_{j_n}|}{(x_{i_n} + x_{j_n})}\right) \quad (3)$$

Similarly, Bray-Curtis distance, b(i, j), also has values ranged from 0.0 to 1.0. The value 0.0 indicates the maximum similarity between two data objects. b(i, j) is defined as:

$$b(i,j) = \frac{|x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + ... + |x_{i_n} - x_{j_n}|}{(x_{i_1} + x_{j_1}) + (x_{i_2} + x_{j_2}) + ... + (x_{i_n} + x_{j_n})} \quad (4)$$

Coefficients of Correlation

Coefficients of correlation are the measurements that describe the strength of the relationship between two variables X and Y. It essentially answers the question *"how similar are X and Y?"*. In our study, coefficients of correlation will be used to compare outcomes (i.e., hierarchical trees) of different clustering techniques. The values of the coefficients of correlation range from 0 to 1 where the value 0 points to *no similarity* and the value 1 points *high similarity*. The coefficient of correlation is used to find the similarity among (clustering) objects. The correlation r of two random variables X and Y where: X = (x_1, x_2, x_3, . . . , x_n) and Y = (y_1, y_2, y_3, . . . , y_n) is given by the formula:

$$r = \frac{|E(X,Y) - E(X) \times E(Y)|}{\sqrt{(E(X^2) - E^2(X)}\sqrt{(E(Y^2) - E^2(Y)}} \quad (5)$$

where

$E(X) = (\sum_{i=1}^{n} x_i)/n$,

$E(Y) = (\sum_{i=1}^{n} y_i)/n$, and