A Data Distribution View of Clustering Algorithms

Junjie Wu
Tsinghua University, China

Jian Chen
Tsinghua University, China

Hui Xiong
Rutgers University, USA

INTRODUCTION

Cluster analysis (Jain & Dubes, 1988) provides insight into the data by dividing the objects into groups (clusters), such that objects in a cluster are more similar to each other than objects in other clusters. Cluster analysis has long played an important role in a wide variety of fields, such as psychology, bioinformatics, pattern recognition, information retrieval, machine learning, and data mining. Many clustering algorithms, such as K-means and Unweighted Pair Group Method with Arithmetic Mean (UPGMA), have been well-established.

A recent research focus on clustering analysis is to understand the strength and weakness of various clustering algorithms with respect to data factors. Indeed, people have identified some data characteristics that may strongly affect clustering analysis including high dimensionality and sparseness, the large size, noise, types of attributes and data sets, and scales of attributes (Tan, Steinbach, & Kumar, 2005). However, further investigation is expected to reveal whether and how the data distributions can have the impact on the performance of clustering algorithms. Along this line, we study clustering algorithms by answering three questions:

1. What are the systematic differences between the distributions of the resultant clusters by different clustering algorithms?
2. How can the distribution of the "true" cluster sizes make impact on the performances of clustering algorithms?
3. How to choose an appropriate clustering algorithm in practice?

The answers to these questions can guide us for the better understanding and the use of clustering methods. This is noteworthy, since 1) in theory, people seldom realized that there are strong relationships between the clustering algorithms and the cluster size distributions, and 2) in practice, how to choose an appropriate clustering algorithm is still a challenging task, especially after an algorithm boom in data mining area. This chapter thus tries to fill this void initially. To this end, we carefully select two widely used categories of clustering algorithms, i.e., K-means and Agglomerative Hierarchical Clustering (AHC), as the representative algorithms for illustration. In the chapter, we first show that K-means tends to generate the clusters with a relatively uniform distribution on the cluster sizes. Then we demonstrate that UPGMA, one of the robust AHC methods, acts in an opposite way to K-means; that is, UPGMA tends to generate the clusters with high variation on the cluster sizes. Indeed, the experimental results indicate that the variations of the resultant cluster sizes by K-means and UPGMA, measured by the Coefficient of Variation (CV), are in the specific intervals, say [0.3, 1.0] and [1.0, 2.5] respectively. Finally, we put together K-means and UPGMA for a further comparison, and propose some rules for the better choice of the clustering schemes from the data distribution point of view.

BACKGROUND

People have investigated clustering algorithms from various perspectives. Many data factors, which may strongly affect the performances of clustering schemes, have been identified and addressed. Among them the

high dimensionality, the large size, and the existence of noise and outliers are typically the major concerns.

First, it has been well recognized that high dimensionality can make negative impact on various clustering algorithms which use Euclidean distance (Tan, Steinbach, & Kumar, 2005). To meet this challenge, one research direction is to make use of dimensionality reduction techniques, such as Multidimensional Scaling, Principal Component Analysis, and Singular Value Decomposition (Kent, Bibby, & Mardia, 2006). A detailed discussion on various dimensionality reduction techniques for document data sets has been provided by Tang et al. (2005). Another direction is to redefine the notions of proximity, e.g., by the Shared Nearest Neighbors similarity (Jarvis & Patrick, 1973). Some similarity measures, e.g., cosine, have also shown appealing effects on clustering document data sets (Zhao & Karypis, 2004).

Second, many clustering algorithms that work well for small or medium-size data sets are unable to handle large data sets. For instance, AHC is very expensive in terms of its computational and storage requirements. Along this line, a discussion of scaling K-means to large data sets was provided by Bradley et al. (1998). Also, Ghosh (2003) discussed the scalability of clustering methods in depth. A more broad discussion of specific clustering techniques can be found in the paper by Murtagh (2000). The representative techniques include CURE (Guha, Rastogi, & Shim, 1998), BIRCH (Zhang, Ramakrishnan, & Livny, 1996), CLARANS (Ng & Han, 2002), etc.

Third, outliers and noise in the data can also degrade the performance of clustering algorithms, such as K-means and AHC. To deal with this problem, one research direction is to incorporate some outlier removal techniques before conducting clustering. The representative techniques include LOF (Breunig, Kriegel, Ng, & Sander, 2000), HCleaner (Xiong, Pandey, Steinbach, & Kumar, 2006), etc. Another research direction is to handle outliers during the clustering process. There have been several techniques designed for such purpose, e.g., DBSCAN (Ester, Kriegel, Sander, & Xu, 1996), Chameleon (Karypis, Han, & Kumar, 1999), SNN density-based clustering (Ertoz, Steinbach, & Kumar, 2001), and CURE (Guha, Rastogi, & Shim, 1998).

In this chapter, however, we focus on understanding the impact of the data distribution, i.e., the distribution of the "true" cluster sizes, on the performances of K-means and AHC, which is a natural extension of our previous work (Xiong, Wu, & Chen, 2006; Wu, Xiong, Wu, & Chen, 2007). Also, we propose some useful rules for the better choice of clustering algorithms in practice.

MAIN FOCUS

Here, we explore the relationship between the data distribution and the clustering algorithms. Specifically, we first introduce the statistic, i.e., Coefficient of Variation (CV), to characterize the distribution of the cluster sizes. Then, we illustrate the effects of K-means clustering and AHC on the distribution of the cluster sizes, respectively. Finally, we compare the two effects and point out how to properly utilize the clustering algorithms on data sets with different "true" cluster distributions. Due to the complexity of this problem, we also conduct extensive experiments on data sets from different application domains. The results further verify our points.

A Measure of Data Dispersion Degree

Here we introduce the Coefficient of Variation (CV) (DeGroot & Schervish, 2001), which measures the dispersion degree of a data set. The CV is defined as the ratio of the standard deviation to the mean. Given a set of data objects $X = \{x_1, x_2, ..., x_n\}$, we have $CV = s/\bar{x}$ where

$$\bar{x} = \sum_{i=1}^{n} x_i / n \text{ and}$$

$$s = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 / (n-1)}.$$

Note that there are some other statistics, such as standard deviation and skewness (DeGroot & Schervish, 2001), which can also be used to characterize the dispersion of a data distribution. However, the standard deviation has no scalability; that is, the dispersion of the original data and stratified sample data is not equal if the standard deviation is used. Indeed, this does not agree with our intuition. Meanwhile, skewness cannot catch the dispersion in the situation that the data is symmetric but has high variance. In contrast, the CV is a dimensionless number that allows comparison of