On Clustering Techniques

Sheng Ma
Machine Learning for Systems IBM T. J. Watson Research Center, USA

Tao Li
School of Computer Science, Florida International University, USA

INTRODUCTION

Clustering data into sensible groupings, as a fundamental and effective tool for efficient data organization, summarization, understanding and learning, has been the subject of active research in several fields such as statistics (Jain & Dubes, 1988; Hartigan, 1975), machine learning (Dempster, Laird & Rubin, 1977), Information theory (Linde, Buzo & Gray, 1980), databases (Guha, Rastogi & Shim, 1998; Zhang, Ramakrishnan & Livny, 1996) and Bioinformatics (Cheng & Church, 2000) from various perspectives and with various approaches and focuses. From application perspective, clustering techniques have been employed in a wide variety of applications such as customer segregation, hierarchal document organization, image segmentation, microarray data analysis and psychology experiments.

Intuitively, the clustering problem can be described as follows: Let W be a set of n entities, finding a partition of W into groups such that the entities within each group are **similar** to each other while entities belonging to different groups are **dissimilar**. The entities are usually described by a set of measurements (attributes). Clustering does not use category information that labels the objects with prior identifiers. The absence of label information distinguishes cluster analysis from classification and indicates that the goals of clustering is just finding a hidden structure or compact representation of data instead of discriminating future data into categories.

BACKGROUND

Generally clustering problems are determined by five basic components:

- **Data representation:** What's the (physical) representation of the given data set? What kind of attributes (e.g., numerical, categorical or ordinal)?
- **Data generation:** The formal model for describing the generation of the data set. For example, Gaussian mixture model is a model for data generation.
- **Criterion/objective function:** What are the objective functions or criteria that the clustering solutions should aim to optimize? Typical examples include entropy, maximum likelihood and within-class or between-class distance (Li, Ma & Ogihara, 2004a).
- **Optimization procedure:** What is the optimization procedure for finding the solutions? Clustering problem is known to be NP-complete (Brucker, 1977) and many approximation procedures have been developed. For instance, Expectation-Maximization (EM) type algorithms have been widely used to find local minima of optimization.
- **Cluster validation and interpretation**: Cluster validation evaluates the clustering results and judges the cluster structures. Interpretation is often necessary for applications. Since there is no label information, clusters are sometimes justifies by ad hoc methods (such as exploratory analysis) based on specific application areas.

For a given clustering problem, the five components are tightly coupled. The formal model is induced from the physical representation of the data, the formal model along with the objective function determines the clustering capability and the optimization procedure decides how efficiently and effectively the clustering results can be obtained. The choice of the optimization procedure depends on the first three components. Validation of cluster structures is a way of verifying assumptions on data generation and evaluating the optimization procedure.

MAIN THRUST

We review some of current clustering techniques in the section. Figure 1 gives the summary of clustering techniques. The following further discusses traditional clustering techniques, spectral based analysis, model-based clustering and co-clustering.

Traditional clustering techniques focus on one-sided clustering and they can be classified into partitional, hierarchical, density-based, and grid-based (Han & Kamber, 2000). Partitional clustering attempts to directly decompose the data set into disjoint classes such that the data points in a class are nearer to one another than the data points in other classes. Hierarchical clustering proceeds successively by building a tree of clusters. Density-based clustering is to group the neighboring points of a data set into classes based on density conditions. Grid-based clustering quantizes the data space into a finite number of cells that form a grid-structure and then performs clustering on the grid structure. Most of these algorithms use distance functions as objective criteria and are not effective in high dimensional spaces.

As an example, we take a closer look at K-means algorithms. The typical K-means type algorithm is a widely-used partitional-based clustering approach. Basically, it first chooses a set of K data points as initial cluster representatives (e.g., centers), and then performs an iterative process that alternates between assigning the data points to clusters based on their distances to the cluster representatives and updating the cluster representatives based on new cluster assignments. The iterative optimization procedure of K-means algorithm is a special form of EM-type procedure. The K-means type algorithm treats each attribute equally and computes the distances between data points and cluster representatives to determine cluster memberships.

A lot of algorithms have been developed recently to address the efficiency and performance issues presented in traditional clustering algorithms. Spectral analysis has been shown to tightly relate to clustering task. Spectral clustering (Weiss, 1999; Ng, Jordan & Weiss, 2001), closely related to the latent semantics index (LSI), uses selected eigenvectors of the data affinity matrix to obtain a data representation that can be easily clustered or embedded in a low-dimensional space. Model-based clustering attempts to learn generative models, by which the cluster structure is determined, from the data. (Tishby, Pereira & Bialek, 1999; Slonim & Tishby, 2000) develop information bottleneck formulation, in which given the empirical joint distribution of two variables, one variable is compressed so that the mutual information about the other is preserved as much as possible. Other recent developments of clustering techniques include ensemble clustering, support vector clustering, matrix factorization, high-dimensional data clustering, distributed clustering and etc.

Figure 1. Summary of clustering techniques