Clustering Techniques

Sheng Ma
IBM T.J. Watson Research Center, USA

Tao Li
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 (Hartigan, 1975; Jain & Dubes, 1988), 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 an 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 $\boldsymbol{W}$ be a set of n entities, finding a partition of $\boldsymbol{W}$ into groups, such that the entities within each group are similar to each other, while entities belonging to different groups are dissimilar. The entities usually are 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 are 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 is the (physical) representation of the given dataset? What kind of attributes (e.g., numerical, categorical or ordinal) are there?
- **Data Generation:** The formal model for describing the generation of the dataset. 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? A 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 used widely to find local minima of optimization.
- **Cluster Validation and Interpretation:** Cluster validation evaluates the clustering results and judges the cluster structures. Interpretation often is necessary for applications. Since there is no label information, clusters are sometimes justified by ad hoc methods (such as exploratory analysis), based on specific application areas.

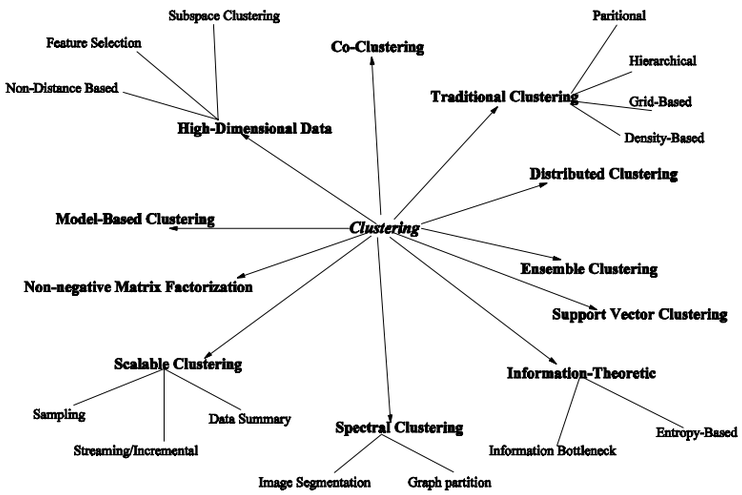
For a given clustering problem, the five components are tightly coupled. The formal model is induced from the physical representation of the data; the formal model, along with the objective function, determines the clustering capability, and the optimization procedure decides how efficiently and how 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 of evaluating the optimization procedure.

MAIN THRUST

We review some of the current clustering techniques in this section. Figure 1 gives a 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 as partitional, hierarchical, density-based, and grid-based (Han & Kamber, 2000). Partitional clustering attempts to directly decom-

Figure 1. Summary of clustering techniques



pose the dataset 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 grouping the neighboring points of a dataset 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 partition-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 (Ng, Jordan & Weiss, 2001; Weiss, 1999), closely related to the latent semantics index (LSI), uses selected eigenvectors of the data affinity matrix to obtain a data representation that easily can be 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, and Bialek (1999) and Slonim and Tishby (2000) developed 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 so forth.

Another interesting development is co-clustering, which conducts simultaneous, iterative clustering of both data points and their attributes (features) through utilizing the canonical duality contained in the point-by-attribute data representation. The idea of co-clustering of data points and attributes dates back to Anderberg (1973) and Nishisato (1980). Govaert (1985) researches simultaneous block clustering of the rows and columns of the contingency table. The idea of co-clustering also has been applied to cluster gene expression and experiments (Cheng & Church, 2000). Dhillon (2001) presents a co-clustering algorithm for documents and words using bipartite graph formulation and a spectral heuristic. Recently, Dhillon, et al. (2003) proposed an information-theoretic co-clustering method for a two-dimensional contingency table. By viewing the non-negative contingency table as a joint probability distribution between two discrete random variables, the optimal co-clustering then maximizes the mutual information between the clustered random variables. Li and Ma (2004) recently developed Iterative Feature and Data (IFD) clustering by rep-