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 $W$ 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 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-