Clustering Algorithm for Arbitrary Data Sets

Yu-Chen Song
Inner Mongolia University of Science and Technology, China

Hai-Dong Meng
Inner Mongolia University of Science and Technology, China

INTRODUCTION

Clustering analysis is an intrinsic component of numerous applications, including pattern recognition, life sciences, image processing, web data analysis, earth sciences, and climate research. As an example, consider the biology domain. In any living cell that undergoes a biological process, different subsets of its genes are expressed in different stages of the process. To facilitate a deeper understanding of these processes, a clustering algorithm was developed (Ben-Dor, Shamir, & Yakhini, 1999) that enabled detailed analysis of gene expression data. Recent advances in proteomics technologies, such as two-hybrid, phage display and mass spectrometry, have enabled the creation of detailed maps of biomolecular interaction networks. To further understanding in this area, a clustering mechanism that detects densely connected regions in large protein-protein interaction networks that may represent molecular complexes was constructed (Bader & Hogue, 2003). In the interpretation of remote sensing images, clustering algorithms (Sander, Ester, Kriegel, & Xu, 1998) have been employed to recognize and understand the content of such images. In the management of web directories, document annotation is an important task. Given a predefined taxonomy, the objective is to identify a category related to the content of an unclassified document. Self-Organizing Maps have been harnessed to influence the learning process with knowledge encoded within a taxonomy (Adami, Avesani, & Sona, 2005). Earth scientists are interested in discovering areas of the ocean that have a demonstrable effect on climatic events on land, and the SNN clustering technique (Ertöz, Steinbach, & Kumar, 2002) is one example of a technique that has been adopted in this domain. Also, scientists have developed climate indices, which are time series that summarize the behavior of selected regions of the Earth's oceans and atmosphere. Clustering techniques have proved crucial in the production of climate indices (Steinbach, Tan, Kumar, Klooster, & Potter, 2003).

In many application domains, clusters of data are of arbitrary shape, size and density, and the number of clusters is unknown. In such scenarios, traditional clustering algorithms, including partitioning methods, hierarchical methods, density-based methods and grid-based methods, cannot identify clusters efficiently or accurately. Obviously, this is a critical limitation. In the following sections, a number of clustering methods are presented and discussed, after which the design of an algorithm based on Density and Density-reachable (CADD) is presented. CADD seeks to remedy some of the deficiencies of classical clustering approaches by robustly clustering data that is of arbitrary shape, size, and density in an effective and efficient manner.

BACKGROUND

Clustering aims to identify groups of objects (clusters) that satisfy some specific criteria, or share some common attribute. Clustering is a rich and diverse domain, and many concepts have been developed as the understanding of clustering develops and matures (Tan, Steinbach, & Kumar, 2006). As an example, consider spatial distribution. A typology of clusters based on this includes: Well-separated clusters, Center-based clusters, Contiguity-based clusters, and Density-based clusters. Given the diversity of domains in which clustering can be applied, and the diverse characteristic and requirements of each, it is not surprising that numerous clustering algorithms have been developed. The interested reader is referred to the academic literature (Qiu, Zhang, & Shen, 2005), (Ertöz, Steinbach, & Kumar, 2003), (Zhao, Song, Xie, & Song, 2003), (Ayad & Kamel, 2003), (Karypis, Han, & Kumar, 1999) for further information.

Though the range of clustering algorithms that have been developed is broad, it is posisble to classify them according the broad approach or method adopted by each:

- A partitioning method creates an initial set of k partitions, where the parameter k is the number of partitions to be constructed. Then it uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. Typical partitioning methods include K-means , K-medoids, CLARANS, and their derivatives.
- A hierarchical method creates a hierarchical decomposition of the given set of data objects. The method can be classified as being either agglomerative (bottom-up) or divisive (top-down), based on how the hierarchical decomposition is formed. To compensate for the rigidity of the merge or split, the quality of hierarchical agglomeration can be improved by analyzing object linkages at each hierarchical partition - an approach adopted by the CURE and Chameleon algorithms, or by integrating other clustering techniques such as iterative relocation – an approach adopted by BIRCH.
- A density-based method clusters objects based on the concept of density. It either grows the cluster according to the density of the neighborhood objects (an approach adopted by DBSCAN), or according to some density function (such that used by DENCLUE).
- A grid-based method first quantizes the object space into a finite number of cells thus forming a grid structure, and then performs clustering on the grid structure. STING is a typical example of a grid-based method based on statistical information stored in grid cells. CLIQUE and Wave Cluster are examples of two clustering algorithms that are both grid-based and density-based.
- A model-based method hypothesizes a model for each of the clusters and finds the best fit of the data to that model. Typical model-based methods involve statistical approaches (such as COBWER, CLASSIT, and AutoClass).

In essence, practically all clustering algorithms attempt to cluster data by trying to optimize some objective function.

DEVELOPMENT OF A CLUSTERING AGORITHM

Before the development of a clustering algorithm can be considered, it is necessary to consider some problems intrinsic to clustering. In data mining, efforts have focused on finding methods for efficient and effective cluster analysis in large databases. Active themes of research focus on the effectiveness of methods for clustering complex shapes and types of data, high-dimensional clustering techniques, scalability of clustering methods, and methods for clustering mixed numerical and categorical data in large databases. When clustering algorithms are analyzed, it is obvious that there are some intrinsic weaknesses that affect their applicability:

- Reliability to parameter selection - for partitioning methods and hierarchical methods, it is necessary to input parameters both for the number of clusters and the initial centroids of the clusters. This is difficult for unsupervised data mining when there is lack of relevant domain knowledge (Song & Meng, 2005 July), (Song & Meng, 2005 June). At the same time, different random initializations for number of clusters and centroids of clusters produce diverse clustering results, indicating a lack of stability on the part of the clustering method.
- Sensitivity to noise and outliers - noise and outliers can unduly influence the clusters derived by partitioning methods, hierarchical methods, grid-based methods, and model-based methods; however partitioning methods and hierarchical methods are particularly susceptible.
- Selectivity to cluster shapes - partitioning methods, hierarchical methods, and grid-based methods are not suitable for all types of data distribution, and cannot handle non-globular clusters of different shapes, sizes and densities.
- Ability to detect outliers - density-based methods are relatively resistant to noise and can handle clusters of arbitrary shapes and sizes, but can not detect outliers effectively.