Abstract

Rapid advances in data capture, transmission and storage technologies have enabled modern business and science to collect increasingly large volumes of data. Data mining is the technique of analyzing such large datasets in order to reveal embedded patterns (regularities and relationships) that are nontrivial. Clustering is one of the primary data analysis tasks in data mining. Clustering techniques partition a set of objects so that objects with similar characteristics are grouped together and different groups contain objects with dissimilar characteristics. It is either used as a stand alone tool to get insight into the data distribution pattern of a dataset or as a preprocessing step for other data mining algorithms operating on the detected clusters.

The attributes used to describe data objects can be quantitative, qualitative or mixture of both. The types of attributes determine the clustering techniques to be used to analyze the data.

Data mining applications place special requirements on clustering algorithms including : scalability, ability to find clusters embedded in subspaces of high dimensional data, ability to find clusters with widely varying sizes, shapes and densities, ability to deal with mixture of attribute types, and insensitivity to the order of input records.

We have developed separate algorithms for clustering numeric, categorical and mixed-type data satisfying these requirements. Two application specific techniques are also developed. The following are the different algorithms included in the thesis.

1. *An Improved Sampling-Based DBSCAN for Large Spatial Databases.*

2. *A Parallelization of Density-Based Clustering Technique on Distributed Memory Multicomputers.*

3. *DDSC: A Density Differentiated Spatial Clustering Technique* to detect clusters with widely differing densities.

4. *CatSub: Clustering Categorical Data Based on Subspace.*

5. *SMIC: A Subspace Preferenced Mixed Type Data Clustering Technique* to find clusters in large high dimensional datasets with mixture of numeric and categorical attributes.

6. *Biclustering Gene Expression Data Using A Node Addition Algorithm.*

7. *A Clustering Based Technique For Network Intrusion Detection.*

Experimental results establish the validity of the algorithms proposed.

Keywords: *Clustering, query sampling, distributed clustering, variable density, categorical data, mixed-type data, incremental clustering, scalability, outliers, gene expression data, biclustering, intrusion detection.*