1. Name: Yong Shi
2. Email: yongshi@cse.buffalo.edu
3. Dissertation Title: Dynamic Data Mining on Multi-Dimensional Data
4. Committee Chair: Dr. Aidong Zhang
5. Committee Members: Dr. Xin He, Dr. Jinhui Xu, Dr. Ling Bian

Abstract

The generation of multi-dimensional data has proceeded at an explosive rate in many disciplines with the advance of modern technology, which greatly increases the challenges of comprehending and interpreting the resulting mass of data. Existing data analysis techniques have difficulty in handling multi-dimensional data. Multi-dimensional data has been a challenge for data analysis because of the inherent sparsity of the points.

A first step toward addressing this challenge is the use of clustering techniques, which is essential in the data mining process to reveal natural structures and identify interesting patterns in the underlying data. Cluster analysis is used to identify homogeneous and well-separated groups of objects in databases. There are also a lot of approaches designed for outlier detection. In many situations, clusters and outliers are concepts whose meanings are inseparable to each other, especially for those data sets with noise. It is well acknowledged that in the real world a large proportion of data has irrelevant features which may cause a reduction in the accuracy of some algorithms. One of the well known techniques for improving the data analysis performance is the method of dimension reduction which is often used in clustering, classification, and many other machine learning and data mining applications. There are also many approaches proposed to index high-dimensional data sets for efficient querying. Although most of them can efficiently support nearest neighbor search for low dimensional data sets, they degrade rapidly when dimensionality goes higher. Also the dynamic insertion of new data can cause original structures no longer handle the data sets efficiently since it may greatly increase the amount of data accessed for a query.

In this dissertation, we study the problems mentioned above. We proposed a novel data preprocessing technique called shrinking which optimizes the inner structure of data. We then proposed a shrinking-based clustering algorithm for multi-dimensional data and extended the algorithm to the dimension reduction field, resulting in a shrinking-based dimension reduction algorithm. We also proposed a cluster-outlier iterative detection algorithm to detect the clusters and outliers in another perspective for noisy data sets. Apart from the shrinking-based data analysis and cluster-outlier interactive relationship exploration research, we designed a new indexing structure, ClusterTree+, for time-related high-dimensional data, which eliminates obsolete data dynamically and keeps the data in the most updated status so as to further promote the efficiency and effectiveness of data insertion, query and update.