HD-Eye: Visual Clustering of High Dimensional Data

Alexander Hinneburg
University of Halle, Germany
hinneburg@informatik.uni-halle.de

Daniel A. Keim
AT&T Research Labs, USA
and University of Constance
keim@research.att.com

Markus Wawryniuk
University of Konstanz, Germany
wawyniu@fmi.uni-konstanz.de

ABSTRACT

Clustering of large data bases is an important research area with a large variety of applications in the data base context. Missing in most of the research efforts are means for guiding the clustering process and understanding the results, which is especially important for high dimensional data. Visualization technology may help to solve this problem since it provides effective support of different clustering paradigms and allows a visual inspection of the results. The *HD-Eye* (high-dim. eye) system shows that a tight integration of advanced clustering algorithms and state-of-the-art visualization techniques is powerful for a better understanding and effective guidance of the clustering process, and therefore can help to significantly improve the clustering results.

Clustering High Dimensional Data

Clustering in large databases of high-dimensional data is an interesting and important, but difficult problem. The clustering problem may be defined as the problem of partitioning the set of data vectors into a number of clusters and noise, such that the data vectors within the clusters are *similar* to each other and the data items which are in different clusters or in the noise partition are *not similar*. The similarity between two data items, however, is difficult to determine and depends on the task and the application.

A large number of specific clustering algorithms have been proposed in the literature of statistics, machine learning, knowledge discovery and databases. A problem of all clustering algorithms is that one clustering paradigm and one parameter setting is usually not sufficient for the whole data set but may only work effectively for a subset of the data, i.e. a subset of the data points or a subset of the dimensions.

The HD-Eye System

The basic idea of the HD-Eye system is to improve the clustering process by allowing the user to directly interact in the crucial steps of the clustering process. The crucial steps of the clustering are **(1)** selection of dimensions to be considered (projections), **(2)** selection of the clustering paradigm and **(3)** partitioning of the data set.

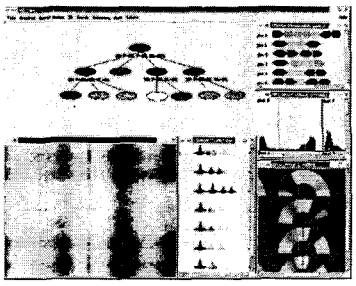

Figure 1: *HD-Eye* screen-shot showing different visualizations of projections and the separator tree.

Since user interaction is crucial for an effective clustering, the HD-Eye system supports a tight integration of advanced clustering algorithms and state-of-the-art visualization techniques, providing a new clustering framework which considers clustering as an iterative process of applying clustering primitives such as density estimators or cluster separators to the data. The strategy is to decompose the data step by step into partitions by combining appropriate density estimation methods with clustering paradigms and applying these combinations to the partitions of the data. The framework concept is similar to decision trees which decompose the data in a divisive, hierarchical way, but in contrast to a purity index (like the gini index) we use statistical information about the clusters and their separation to guide the decomposition.

In the *HD-Eye* system we use novel visualization techniques to help the user identify the most interesting projections and subsets as well as the best separators for partitioning the data. The visualization techniques available in the HD-Eye system include (see figure 1): **(I)** Abstract Iconic Displays (two variants for 1D and multi-dim. projections), **(II)** Color-Based Point Density Displays (1D and 2D) and **(III)** Curve-based Point Density Displays (1D). The *HD-Eye* system allows a visual data explortion by focusing on interesting projections and guiding the important steps of kernel-density-estimation based clustering and The combination of multiple clustering paradigms leads to clustering models which fit well to the intended tasks and the users' interests as well as to a better understanding of the clustering results. Applications include clustering of large image and molecular biology databases.

More information on the *HD-Eye* project can be found via www.inf.uni-konstanz.de/db_vis/

ACM SIGMOD '2002 June 4-6, Madison, Wisconsin, USA