Subscan - a cluster algorithm for identifying statistically dense subspaces

Hans A. Kestler, Johann M. Kraus
Research group Bioinformatics and Systems Biology
Institute of Neural Information Processing, Ulm University

Cluster analysis is an important technique of explorative data mining. It refers to a collection of statistical methods for learning the structure of data by solely exploring pairwise distances or similarities in feature space. Recent approaches in clustering aim at detecting groups of data points that exist in arbitrary, possibly overlapping subspaces. Generally, subspace clusters are neither exclusive nor exhaustive, i.e. subspace clusters can overlap as well as data points are not forced to participate in clusters. In this context subspace clustering supports the search for meaningful clusters by including dimensionality reduction in the clustering process. Subspace clustering can overcome drawbacks from searching groups in high-dimensional data sets, as often observed in clustering biological or medical data. In the context of microarray data this refers to the hypothesis that only a small number of genes is responsible for different tumor subgroups. We generalize the notion of scan statistics to multi-dimensional space and introduce a new formulation of subspace clusters as aggregated structures from dense intervals reported by single axis scans. Our approach objectifies the search for subspace clusters as the reported clusters are of statistical relevance and are not artifacts observed by chance. Like in hierarchical cluster analysis there are two possible strategies to detect relevant subspace clusters. In a top-down approach, the dimension of clusters identified in the full space is reduced until a minimal subspace supporting the cluster assumption is reached. Using a bottom-up strategy allows the agglomeration of clusters from regions of high density across intervals from different dimensions. We present a bottom-up algorithm to grow high-dimensional subspace clusters from one-dimensional statistically dense seed regions. Our experiments demonstrate the applicability of the approach to both low-dimensional as well as high-dimensional data.