

EINLADUNG

Zeit: Mittwoch, 09. Juni 2010, 16:15 Uhr

Ort: Seminarraum 5056, Ahornstr. 55

Referent: Dipl.-Inform. Emmanuel Müller

Titel: Efficient Knowledge Discovery
in Subspaces of High Dimensional Databases

Abstract:

In many recent applications such as sensor network analysis, customer segmentation or gene expression analysis, tremendous amount of data is gathered. As collecting and storing of data is cheap, users tend to record as much information as possible. However, the valuable knowledge to be learned out of this information is often hidden in subsets of the measured attributes. The discovery of knowledge in such subspaces poses major challenges for today's data mining methods.

Commonly used techniques such as clustering are unable to detect groups of similar objects hidden in subspaces. Subspace clustering aims at detecting clusters in any subspace projection. However, considering all possible subspaces expands the search space significantly. In general, this yields several open challenges for subspace cluster models, their algorithmic computation, and evaluation of results.

In this talk, we discuss our novel methods tackling these challenges. Our subspace clustering models adapt to the intrinsic properties of subspace projections to achieve high quality clusters. We propose novel non-redundant subspace clustering models where each cluster contributes significantly to the extracted knowledge. Our efficient processing schemes avoid the exhaustive search of all subspaces and reduce costly database accesses. We prune large parts of the search space and process only the most promising subspace regions. Overall, our techniques are scalable to large and high dimensional databases providing few and high quality subspace clusters.

As a general contribution to the community, we provide a systematic evaluation study on a broad set of approaches. We show both efficiency and quality characteristics of major paradigms. Furthermore, our evaluation framework is available as open source project and provides a basis for future enhancements in this emerging research area. Thus, this work proposes not only novel methods for efficient cluster and outlier detection in subspace projections, but it is a fundamental basis for repeatable comparison of recent data mining approaches.

Es laden ein: Die Dozenten der Informatik