Outlier Detection and Description in Complex Databases

Wednesday, September 3, 2014 9:00 AM (40 minutes)

Outlier analysis is an important data mining task that aims to detect unexpected, rare, and suspicious objects in large and complex databases. Consistency checks in sensor networks, fraud detection in financial transactions, and emergency detection in health surveillance are only some of today's application domains for outlier analysis. As measuring and storing of data has become cheap, in all of these applications, objects are described by a large variety of different measures and relationships between objects. However, out of these complex databases, for each object only a small subset of relevant measures and relationships provides the meaningful information for outlier detection. The residual information is irrelevant for this object, and with the growing amount of irrelevant information traditional outlier mining approaches fail to detect outliers.

To address this problem, recent subspace search techniques focus on a selection of subspace projections. The objective is to find multiple subsets (i.e. subspaces) of the given attributes, which show a significant deviation between an outlier and regular objects. Thus, subspace search allows: (1) A clear distinction between clustered objects and outliers. (2) A description of outlier reasons by the selected subspaces. However, it lacks flexibility in handling different outlier characteristics that have been invented for different application domains and proposed as formal outlier models in the literature.

This talk will cover a flexible subspace selection scheme allowing instantiations with different outlier models. We utilize the differences of outlier scores in random subspaces to perform a combinatorial refinement of relevant subspaces. Our refinement allows an individual selection of subspaces for each outlier, which is tailored to the underlying outlier model. This flexibility ensures that the approach directly benefits from any research progress in future outlier models. It allows search for relevant subspaces individually for each outlier, and hence, enables to describe each outlier by its specific outlier properties.

Presenter: Dr MÜLLER (KIT), Emmanuel

Session Classification: Plenary talks