Flexible Subspace Search for Outlier Detection and Description

Emmanuel Müller

University of Antwerp, Belgium
emmanuel.mueller@ua.ac.be

Karlsruhe Institute of Technology, Germany
emmanuel.mueller@kit.edu

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 attributes. However, for each object only a few relevant attributes provide the meaningful information for outlier detection, the residual attributes are irrelevant for this object. Traditional outlier mining approaches fail to detect outliers in such high dimensional databases as they consider a large number of irrelevant attributes. This effect is subsumed by the well-known *curse of dimensionality* and various sub-effects for outlier mining in high dimensional databases.

To address this problem, subspace search techniques focus on a selection of subspace projections. The objective is to find multiple subsets of the given attributes (i.e. *subspaces*), which show a significant deviation between an outlier and regular objects. For example in health surveillance, for one patient attributes such as "age" and "skin humidity" might be important to detect the abnormal "dehydration" status of this patient. Other attributes such as "heart beat rate" are irrelevant for the detection of this outlier, but are relevant for the detection of other abnormal patients with a heart disease. From our research perspective, selected subspaces should provide a clear contrast between regular objects and outliers and help the user in the manual verification of unexpected measurements. In our example, providing information about the high deviation in "age" and "skin humidity" while showing normal measurements in all other attributes assists health professionals in verifying this automatically detected outlier. Thus, subspace search should allow: (1) A clear distinction between clustered objects and outliers; (2) a description of outlier reasons by the selected subspaces.

However, flexibility in handling different outlier characteristics is an important issue missing in subspace search. Looking at different application domains and formal outlier models proposed in the literature we find various outlier characteristics. For instance, some models are sensitive to distance deviations; others capture deviation in the local density; yet other models prefer angle-based or statistical deviation. Depending on the outlier model used, different objects in different subspaces have the highest deviation. It is an open research issue to make subspace selection flexible w.r.t. the use of different outlier models. In this work we propose such a flexible and adaptive subspace selection scheme. Our generic processing allows 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. In our empirical evaluation we show the flexibility of our subspace search w.r.t. various outlier models such as distance-based, angle-based, and local-density-based outlier detection.

References to papers covered by the proposed presentation:

Keller F., Müller E., Wixler A., Böhm K.: ***Flexible and Adaptive Subspace Search for Outlier Analysis***
*In Proc. ACM International Conference on Information and Knowledge Management (**CIKM 2013**)*

Nguyen H. V., Müller E., Böhm K.: ***4S: Scalable Subspace Search Scheme Overcoming Traditional Apriori Processing***
*In Proc. IEEE International Conference on Big Data (**BigData 2013**)*