Karlsruher Institut für Technologie

Institut für Programmstrukturen und Datenorganisation (IPD)
Lehrstuhl für Systeme der Informationsverwaltung, Prof. Böhm

Masterarbeit

Studying the Placement of Outliers Hidden in Subspaces to Prevent Catastrophic Failures or Sabotage

Masterarbeit

In general, outliers are data objects that deviate from other (usual) data objects. This deviation typically is not consistent across attribute subspaces. Consider for example data on the temperature of an engine and the speed the engine is running at. These two measures should clearly be highly correlated. I. e., a combination of a high temperature and a low speed is an outlier. However, when we examine both one-dimensional subspaces on their own, the values might be in usual ranges of temperature and speed. Depending on the subspaces searched for outliers, this is what we refer to as hidden outlier: Its outlierness in the two-dimensional full space is hidden by the one-dimensional subspaces. If outliers are hidden, they may be missed when detecting them, which might have drastic consequences. For example, the situation above might lead to an engine failure that could have been avoided.

A pure analytical analysis of hidden outliers is impractical, leastwise due to the variety of outlier detection methods. Hence, we recently developed an approach to place hidden outliers in datasets. Although the algorithm can place hidden outliers, the systematical development and evaluation of new design alternatives could lead to significant improvements. We are already aware of some issues of the developed placement that should be improved. These are

- The placement relies on a sampling procedure that is related to uniform sampling across the whole data space. This may produce unrealistic objects, e. g., objects that are clearly not generated from the process the data originated from, e. g., when the data is a time series. The objects might also be unrealistic when consulting domain knowledge, e. g. if fuel suction is zero ⇒ engine cannot be running.
- Another problem is the dependency of the placing procedure on the outlier detection and subspaces used in the placing process. I. e., how hidden do the placed objects remain when subspaces or detection methods used change?

Based on these issues, the central question behind this assignment is: "**What is an effective and efficient way of placing realistic hidden outliers, and how can this be evaluated?**". The work towards answering this question can be divided into smaller task as follows

- Identify a technique to compare placed hidden objects with usual data objects
- Analyze the validity of points placed with the existing approach
    - using different real-world and artificial datasets
    - considering changes in the parameters, e. g. the detection method
- Using the results from the previous analysis, determine the space of constraints that improve the validity of placed points. E. g. simple constraints on single attributes or complex ones including some dependencies (temperature and engine speed)
- Develop a method to integrate any such constraint into the placement. This method should incorporate that the constraints might be provided by domain experts or other users

With this work you will develop your skills in Big Data Analytics as well as regarding the design and evaluation of algorithms.

Ansprechpartner

Georg Steinbuß, M. Sc. georg.steinbuss@kit.edu +49 721 608-43911 Raum: 363

Am Fasanengarten 5 76131 Karlsruhe Gebäude: 50.34