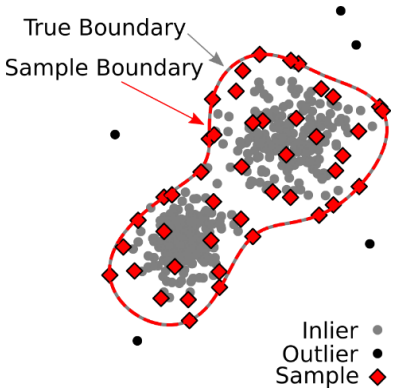
Master's Thesis

Lehrstuhl für Systeme der Informationsverwaltung, Prof. Böhm

Karlsruher Institut für Technologie

**Master's Thesis**

# Scaling Support Vector Data Description sampling to very large data sets




Outlier detection reveals unusual patterns in data. Popular applications for outlier detection are intrusion detection in computer networks or predictive maintenance to identify failures in industrial plants. Nowadays, data collection is cheap and data sets quickly exceed one million observations. Then, manually analyzing the data set is impossible and one must automate the process of finding outliers. One common unsupervised classifier for outlier detection is Support Vector Data Description (SVDD). The basic idea of SVDD is to fit a tight hypersphere around the majority of the observations, the inliers, to distinguish them from outliers. However, a downside of SVDD is that it does not scale well with data set size. Training times quickly become prohibitive with a few thousand observations. To mitigate the issue, literature has proposed "sampling methods" that select a subset of the data, the sample, for training. However, existing sampling methods require data structures that often do not fit into the main memory of modern machines for very large data sets of more than one million observations. Another, related problem is that sampling methods depend on hyperparameters, which must be tuned for each data set individually. This adds additional complexity to selecting a good sample, and currently is in the way of using sampling methods in real world settings.

**The focus of this thesis is to develop a sampling approach to scale SVDD to millions of observations.** In particular, the following research questions are of interest:

- Data sets are of varying size and dimensionality. How can we design an approach that works for arbitrary data sets and automatically adapts without retuning parameters manually?
- Sampling is a tradeoff between the resulting sample size and sample quality. How can we give the user control over this tradeoff?
- Considering limited resources, how close is the accuracy of the classifier trained on a sample compared to a classifier trained on the full data set?

**This results in the following tasks:**

- Literature research on recent advances on scaling kernel methods to large data sets
- Design and implementation of an approach to sample very large data sets for SVDD training. One idea to solve the previously mentioned research questions is to design a sampling method that samples hierarchically. Extending an available implementation at the chair is possible.
- Experimental evaluation of the approach on synthetic and real-world data sets including a comparison to the existing sampling methods

Throughout this work, you will acquire a deep knowledge on state-of-the-art methods to detect unusual observations in large data sets. You train highly demanded skills in development and evaluation of machine-learning algorithms. Knowledge from a lecture such as "Big Data Analytics" is beneficial but not a prerequisite. However, elementary statistical knowledge, programming skills and the ability to accomplish conceptual work are desired.

**Ansprechpartner**

Adrian Englhardt adrian.englhardt@kit.edu +49 721 608-47336 Raum: 340

Am Fasanengarten 5 76131 Karlsruhe Gebäude: 50.34