# A Survey on
# Outlier Detection in Data Streams[1]

Aditi Singla (2014CS50277)
Ankush Phulia (2014CS50279)
Vaibhav Bhagee (2014CS50297)

---

# 1 Introduction

In data mining, the task of Outlier Detection is to identify those data points in a dataset which differ significantly from the rest of the points in the dataset and occur in rarity. There are various examples of scenarios where detecting outlier data samples is of use for eg. fraud detection in the financial sector, separation of "noisy" samples from the normal ones to refine the dataset for further processing, detection of "anomalous" event samples in system logs, to prevent large scale system failures etc.
Traditional algorithms have looked at outlier detection in scenarios where the dataset is "static", i.e the data does not change over the period of time and the algorithm has the access to the entire dataset, in memory. However, as the size of the datasets grows large, it might not be always easy to get the entire dataset in memory, for processing. Moreover, there are many scenarios like prevention of system failure, where the data samples like logs, are generated temporally, in a continuous fashion. In such cases, the outlier detection algorithm can never have access to the complete dataset and the analysis for outliers needs to be performed over the "seen" data.
As a part of this survey, we present and discuss various algorithms and techniques, which have been proposed in context of detecting outliers in the setting of streaming data. In particular, we discuss the challenges which are posed when detecting outliers in data streams what approaches are followed to overcome these.

# 2 Motivation

Outlier detection on large datasets has been traditionally looked at alongside clustering and density estimation. Many of the earliest outlier detection algorithms have been the ones performing clustering and in process identifying the "noisy" samples.

However, in high dimensional data, the neighbourhood and clustering based outlier detection algorithms, face issues due to sparsity, failing to escape the curse of dimensionality. In order to avoid the distance computation among the points, exact and approximate algorithms have been proposed, based on projection of data onto smaller dimensions. The basic intuition behind this set of algorithms is that the outlier behaviour of a point is more pronounced in the subspaces of a high dimensional vector space. Hence, these algorithms are known to give better results than the former set of algorithms.

A key assumption which these algorithms make is that the entire dataset is available for processing, in memory. For smaller datasets, this assumption is decent enough to make. However, data has been growing at an exponential rate and modern data problems related to knowledge discovery in databases work on very large datasets, which can be of giga scale or tera scale in size. These

datasets cannot be stored in the main memory for processing and therefore need to be chunked and processed. This is an example of a stream which is of finite size.

In addition to the that, consider the problem of failure detection in systems, which has been described above. In such a typical modern day distributed system, there are multiple micro-services which send their logs to a central process responsible for collection and analysis of logs. In such a case, detection and analysis of log lines corresponding to "anomalous" events, only has access to the logs which have been collected by the process up to that instance of time. Hence, the outlier detection algorithm is looking at a stream of data which is potentially infinite in size.

There have been various techniques which have been proposed to deal with the streaming nature of the data. Among the most popular ones are the ensemble based techniques and subspace partitioning based techniques.

# 3 Problem Formulation

Placeholder for problem formulation

# References

[1] Manzoor, E., Lamba, H., and Akoglu, L. xstream: Outlier detection in feature-evolving data streams. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (New York, NY, USA, 2018), KDD '18, ACM, pp. 1963–1972.