Outlier Detection in Data Streams

A Research Survey

Aditi Singla
Indian Institute of Technology
Delhi

Ankush Phulia
Indian Institute of Technology
Delhi

Vaibhav Bhagee
Indian Institute of Technology
Delhi

ABSTRACT

This paper discusses various techniques to detect outliers from streaming data, where the features or the data itself can evolve over time.

CCS CONCEPTS

• **Information systems → Data stream mining**; • **Computing methodologies → Anomaly detection**; *Ensemble methods*;

KEYWORDS

Outlier Detection, Data Streams, Feature evolution

ACM Reference Format:
Aditi Singla, Ankush Phulia, and Vaibhav Bhagee. 2018. Outlier Detection in Data Streams: A Research Survey. In *Proceedings of Data Mining (COL761).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

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.

COL761, July - November 2018, IIT Delhi

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

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[2], 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 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.

Recent work[3] looks at streams which can evolve over time in multiple ways:

- Feature values of existing data points can change over time
- New unseen features can emerge over a period of time

These type of feature-evolving streams are common in scenarios like data center monitoring, where the jobs, which are submitted over time, can have feature values like number of unsuccessful retries, number of threads, system calls etc. and can have features,

which get added over time, like URLs of log files, repositories etc.

The existing algorithms and techniques for data streams fail to work for feature evolving streams as they make an assumption that the data streams cannot change dimensionality and the values of existing samples cannot be modified. The authors propose an efficient algorithm for this case, which generalises for the static and the row streams as well.

3 PROBLEM FORMULATION

The problem we are looking at essentially, is as follows: Given a data stream $D = \{x1, x2, ...\}$, we want to identify the outlier data points.

There have been various techniques[5][4][1] 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.

3.1 Ensemble based techniques

Ensemble based techniques[1] have gained popularity after their performance in supervised setting of classification in machine learning. The idea behind ensembles is to utilise the power of many. Multiple models can be trained over the same dataset or multiple subsets of the dataset and their outputs or scores can be suitably combined at to produce the output or the score for the given data point. The component of an ensemble can be trained in a way such that they could be dependent on each other, like sequential ensembles or they could be independent of each other, like in feature bagging.

3.2 Partitioning based techniques

Streams have outlier samples in rarity and can evolve over time, besides being on potentially infinite length. Partitioning based techniques[4] essentially construct an ensemble of trees which perform semi supervised, one-class detection of the rare outliers. These techniques attempt to detect outliers by capturing the density of points in multiple different partitions of the subspaces. The key idea to be able to incorporate for the evolving nature of the streams, is to maintain contiguous, latest and reference windows, to perform outlier detection on the unseen data while updating the reference model atomically with the latest one, after the latest window ends.

The feature-evolution setting can be described as follows: Given a data stream D containing updates of the form (id, f, δ) where f could be a feature which has been seen or unseen until that instant and id is the unique identifier corresponding to the data point. δ is the update to the feature value of the data point as described above.

Due to the feature evolving nature of the stream, the set of points need to be in memory all the time. Given the fact that new features can be added, the space allocation for the data points can't be bounded before hand and hence is challenging. Also, the feature updates need to be fast and efficient.

The algorithm proposed by the authors is a constant space and constant time per update algorithm which combines ideas from subspace partitioning and ensembling while incorporating the feature evolving nature of the input stream. The details of the algorithm along with the advantages and disadvantages of ensembling and partitioning based techniques, will be discussed in the further sections.

4 RELATED WORK

4.1 Distance based outlier detection

Many of the early approaches towards anomaly detection in data streams, employ distance based methods. These distance based approaches take into account, the mutual distance of points, in the dataset, to identify outliers as a bi-product of clustering.

However, the outlier detection algorithms outlined above face issues due to sparsity and curse of dimensionality. In an attempt to formulate the algorithms to overcome the above mentioned issues, Aggarwal et al utilise the idea of projecting the data onto the smaller dimensions[2] and propose exact and approximate algorithms, based on the same.

A key challenge to address in this approach is to define the semantics associated with a point being an outlier, in a subspace of the original dataset, besides being computationally efficient.

Outliers in low dimensional projections

In order to overcome the shortcomings of a high number of dimensions, they project the data points onto a lower dimensional space. Every attribute in the dataset is divided into several equi-depth ranges, dividing the complete space into cubes of different dimensions.

Now, they select ranges from k different dimensions and consider the resulting k - dimensional cubes formed. The idea is to identify the most sparse of such k - dimensional cubes. In order to calculate how sparse the cubes are, Aggarwal et al define a metric, called the *sparsity coefficient*, to identify the extent of deviation in the actual number of points within a cube, from its probabilistic estimate.

Let us assume that every attribute is divided into r equi-depth ranges. Then, because of this being an equi-depth split, each range gets $f = 1/r$ fraction of the points. Now, is the distribution of points had been *independent* and $uniform$, then, a k - dimensional cube as described by the authors, would have contained a point with a probability f^k which is Bernoulli distributed.

Considering that the dataset consists of N points, the estimated number of points inside a k - dimensional cube becomes a Normal distributed random variable (by the application of the central limit theorem), with a mean $N.f^k$ and variance $N.f^k.(1 - f^k)$. Thus, the *sparsity coefficient* is defined as,

$$s = (n - N.f^k)/(N.f^k.(1 - f^k))^{0.5}$$

Given this, the problem finding outliers becomes that of finding the cubes with elements having the most negative sparsity coefficients. This search can be conducted in a simple brute force fashion or using evolutionary algorithms described by the authors and discussed below, in brief.

Evolutionary algorithms for Outlier Detection

The key idea here is to encode the information into strings and then use suitable crossover techniques to generate fitter and fitter strings, where the $fitness$ is defined using a suitable fitness function.

For every dimension, the grid range is encoded as a number between 1 and r, where r is the number of equi-depth ranges as described above, or a $*$ to denote a don't care. The sparsity coefficient is used as the fitness function, to evaluate the fitness of the cube, which is encoded cube represents. The sampling probability of a particular string, representing a cube, is a function of how negative is the sparsity coefficient for that cube.

The encoding and the fitness function, described above, are used in conjunction with a simple $two\ point\ crossover$ technique where a random crossover point is chosen for 2 strings and their right parts are crossed over. In this way, the search is carried out until a fixed number of iterations or until the set of outliers stops changing for a specified number of pre-determined iterations.

The outlier detection algorithm described in this section essentially uses the idea of projection of points on lower dimensional subspaces while preserving the semantics of being an outlier. The idea of choosing the k - dimensional cube with the most negative sparsity coefficient, is in some way, an attempt to choose a set of attributes out of the large number of dimensions, which best preserve the semantics associated with the outlier points.

The latter idea has been explored further as the *ensemble* based methods gained popularity in the machine learning community. The former idea has also been explored by various authors while proposing various outlier detection techniques based on $subspace\ partitioning$. Both these techniques have been discussed in significant detail, in the subsequent sections.

REFERENCES

[1] Charu C Aggarwal. 2013. Outlier ensembles: position paper. *ACM SIGKDD Explorations Newsletter* 14, 2 (2013), 49–58.

[2] Charu C Aggarwal and Philip S Yu. 2001. Outlier detection for high dimensional data. In *ACM Sigmod Record*, Vol. 30. ACM, 37–46.

[3] Emaad Manzoor, Hemank Lamba, and Leman Akoglu. 2018. xStream: Outlier Detection in Feature-Evolving Data Streams. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18)*. ACM, New York, NY, USA, 1963–1972. https://doi.org/10.1145/3219819.3220107

[4] Swee Chuan Tan, Kai Ming Ting, and Tony Fei Liu. 2011. Fast anomaly detection for streaming data. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, Vol. 22. 1511.

[5] Luan Tran, Liyue Fan, and Cyrus Shahabi. 2016. Distance-based outlier detection in data streams. *Proceedings of the VLDB Endowment* 9, 12 (2016), 1089–1100.