Robust Random Cut Forest Based Anomaly Detection On Streams

Sudipto Guha SUDIPTO@CIS.UPENN.EDU
University of Pennsylvania, Philadelphia, PA 19104.

Nina Mishra NMISHRA@AMAZON.COM
Amazon, Palo Alto, CA 94303.

Gourav Roy GOURAVR@AMAZON.COM
Amazon, Bangalore, India 560055.

Okke Schrijvers OKKES@CS.STANFORD.EDU
Stanford University, Palo Alto, CA 94305.

Abstract

In this paper we focus on the anomaly detection problem for dynamic data streams through the lens of random cut forests. We investigate a robust random cut data structure that can be used as a sketch or synopsis of the input stream. We provide a plausible definition of non-parametric anomalies based on the influence of an unseen point on the remainder of the data, i.e., the externality imposed by that point. We show how the sketch can be efficiently updated in a dynamic data stream. We demonstrate the viability of the algorithm on publicly available real data.

1. Introduction

Anomaly detection is one of the cornerstone problems in data mining. Even though the problem has been well studied over the last few decades, the emerging explosion of data from the internet of things and sensors leads us to reconsider the problem. In most of these contexts the data is streaming and well-understood prior models do not exist. Furthermore the input streams need not be append only, there may be corrections, updates and a variety of other dynamic changes. Two central questions in this regard are (1) how do we define anomalies? and (2) what data structure do we use to efficiently detect anomalies over dynamic data streams? In this paper we initiate the formal study of both of these questions. For (1), we view the problem from the perspective of model complexity and say that a point is an anomaly if the complexity of the model increases substantially with the inclusion of the point. The labeling of a point is data dependent and corresponds to the externality imposed by the point in explaining the remainder of the data. We extend this notion of externality to handle "outlier masking" that often arises from duplicates and near duplicate records. Note that the notion of model complexity has to be amenable to efficient computation in dynamic data streams. This relates question (1) to question (2) which we discuss in greater detail next. However it is worth noting that anomaly detection is not well understood even in the simpler context of static batch processing and (2) remains relevant in the batch setting as well.

For question (2), we explore a randomized approach, akin to (Liu et al., 2012), due in part to the practical success reported in (Emmott et al., 2013). Randomization is a powerful tool and known to be valuable in supervised learning (Breiman, 2001). But its technical exploration in the context of anomaly detection is not well-understood and the same comment applies to the algorithm put forth in (Liu et al., 2012). Moreover that algorithm has several limitations as described in Section 4.1. In particular, we show that in the presence of irrelevant dimensions, crucial anomalies are missed. In addition, it is unclear how to extend this work to a stream. Prior work attempted solutions (Tan et al., 2011) that extend to streaming, however those were not found to be effective (Emmott et al., 2013). To address these limitations, we put forward a sketch or synopsis termed *robust random cut forest* (RRCF) formally defined as follows.

Definition 1 *A* **robust random cut tree** *(RRCT) on point set S is generated as follows:*

1. *Choose a random dimension proportional to* $\frac{\ell_i}{\sum_j \ell_j}$, *where* $\ell_i = \max_{x\in S} x_i - min_{x\in S} x_i$.
2. *Choose* $X_i \sim$ *Uniform*$[\min_{x\in S} x_i,\, \max_{x\in S} x_i]$
3. *Let* $S_1 = \{x | x \in S, x_i \le X_i\}$ *and* $S_2 = S \setminus S_1$ *and recurse on* S_1 *and* S_2.

Proceedings of the 33^{rd} *International Conference on Machine Learning*, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).