An overview of modern day Anomaly Detection in Univariate Time-Series

Brandonlee Santos*
(Dated: February 6, 2023)

Abstract

Due to the prevalence of anomalies in data within the fields of fraud detection, cybersecurity, ecosystem disturbances, medical diagnosis and etc, statistical methods have been developed to soundly and systematically decipher anomalies, namely Box-Plot and the Extreme Studentized deviate hypothesis test(cite). Recently, due to the rise of Machine Learning, techniques from density clustering and deep learning have also gained popularity such as DBSCAN, LOF(Local Outlier Factor) AE(Auto Encoders) and VAE(Variational Auto Encoders) (cite). In this paper potential algorithmic schemes are provided for univariate Time-Series and an investigation is done comparing classical statistical and modern day machine learning techniques for labeling anomalies.

Keywords: first keyword, second keyword, third keyword

I. INTRODUCTION AND OUTLINE

It is trivial to visually discover an anomaly on a graph, hence it is considered incumbent for exploratory data analysis. 20th century statisticians John Tukey and Frank E. Grubbs formulated commonly used techniques in statistics; namely Tukey's IQR and Grubbs hypothesis test, despite this, the notion of an anomaly has yet to be mathematically defined(cite tukeys eda and cite og grubb test paper). Although more creative sophisticated benchmarks have been thought of to distinguish between so called anomalies and outliers, it is a computational framework and not inherently supported by a statistical or mathematical theory(cite classification framework). Meanwhile, other papers argue that in fact outliers and anomalies should not be distinguished(cite Charu Aggarwal). Following from this assumption, in this paper I define an anomaly as data point which differs substantially from the distribution, where the anomalies are a notably smaller subspace of the original distribution; this idea was formally defined by Hawkins capturing the idea of 'global' outliers. Originally distribution and clustering methods were at the frontiers of anomaly detection; however, Knorr and Ng proposed the notion of using distance based algorithms, which inspired the advent of LOF also loosely inspired by developments in clustering methods but does not require declaration of clusters. One of the main criticisms echoed in the original LOF paper is that cluster based methods fail to optimize for outlier detection and rather optimizes for clustering. Despite the massive hype behind sophisticated machine learning models for detecting outliers in Time-Series, this paper suggests that the classical statistical Time-Series outlier techniques outperform the state of the art modern Machine Learning models. Regardless of this, the necessity of regression models to obey the Gauss Markov assumptions and the stationarity assumption associated with the Box-Jenkins methodology for SARIMAX models would greatly restrict potential datasets, therefore, in this paper for the classical statistic anomaly techniques I will only consider the Box- Plot and Generalized Extreme Studentized hypothesis test. The fields of Computer Vision and Natural Language Processing have been revolutionized in the 10 years due to the success of deep learning and the development of the attention is all you need mechanism(cite). This paper was heavily inspired from (cite VAE), where AE's and VAE's with different dynamic scoring functions are compared in depth; the conclusion was that the FC AE was the most successful at detecting anomalies, and because the dynamic scoring function has such a large impact on results it seems crucial to focus on the significance of the model by not implementing a scoring function, even though clearly utilizing these scoring functions will typically yield better results(cite VAE). Ensemble methods have also shown to have drastically impacted the trajectory of data science(cite). The original Isolation Forest paper demonstrated that LOF, ORCA and Random Forests had notably worse AUC scores across different datasets for anomaly detection then Isolation Forest (cite iforest). Although one of the main benefits of Isolation Forest is for higher dimensional data and computational speed; however in this experiment it will be judged upon its accuracy on univariate Time-Series. Because of the general research consensus and to the best of my knowledge the ensemble method Isolation Forest, the unsupervised cluster density techniques such as DBSCAN and LOF(Local Outlier Factor), and deep learning techniques FC AE, LSTM VAE have yet to be compared to the classical statistical techniques and each other, thus an investigation of these anomaly techniques is warranted. In this paper a brief introduction to these algorithms is used to provide some intuition, an empirical study is done comparing anomaly

* Correspondence email address: Santos.Brandonlee1@gmail.com

detection effectiveness for these algorithms on different datasets and finally based upon these results a systematic methodology is provided for dealing anomalies.

II. OUTLIER DETECTION IN CLASSICAL STATISTICS

The Box-Plot became popularized with the notorious Tukey's exploratory analysis(cite). Tukey wanted to decompose the summary statistics into a schematic visual plot and carefully chose for anomalies to be points considered $1.5 * IQR$ below the Q_1 and above the Q_3, referred to as the outer fences. There has been many innovations based on Tukey's Box-Plot over the years, such as MEDCOUPLE accounting for more complicated datasets in spite of these Tukey's Box-Plot has remained ' a rule of thumb'.
The Generalized Extreme Studentized Deviate tests the following hypothesis for a dataset X

H_0 : There are no outliers in X

H_1 : There are up to r outliers in X

The test statistic depends on the sample mean and variance $\bar{x}$ and s and is given by

$$R_i = \frac{max_i(x_i - \bar{x})}{s}$$

The procedure is as follows: delete the observation that maximizes the numerator and then redo the computation for n-1 observations and repeat until the upper bound of outliers is detected, which will generate a set of r different test statistics. For each test statistic, r critical values are calculated as such

$$\lambda_i = \frac{(n-i)t_p^2}{\sqrt{n-i+1+t_p^2(n-i+1)}}$$

where $t_{p,v}$ approximately follows student t distribution and p is given by $p = 1 - \frac{\alpha}{2(n-i+1)}$. The number of outliers is found by finding the biggest test statistic which satisfies $R_i > \lambda_i$. In the original paper Rosner demonstrates the sample size should be very accurate for large samples $n > 25$. The ESD test is based on the Grubb's test for detecting single anomalies in Univariate Time-Series but requires a specification of a hyper-parameter for the upper bound of anomalies present(cite rosner). Similarly, the Tietjen Moore test is another generalization of Grubb's test for multiple anomalies in; however the issue with Grubb's and Tietjen Moore test is that specifying the amount of outliers drastically impacts the results as opposed to setting the upper bound. Intuitively the ESD test can effectively be thought of as applying the Grubb's test several times; but, notably, the ESD adjusts the critical values according to the bound provided.

III. ISOLATION FOREST

The main attraction of ensemble methods is that they do not rely on computing a distance metric in higher dimensional space or density estimation. This results in large benefits for anomaly detection accuracy but also computational speed and feasibility; the Isolation Forest is linear in computational time. The original Isolation Forest paper by Liu and Zhou(cite) demonstrates over LOF and ORCA. Other papers have demonstrated its potential usage in the field of finance. The general mechanism of the isolation forest is to use an ensemble of trees to consider the average path length of anomalous data points; this points should be 'isolated' in the tree, hence the name Isolation Forest. The intuition behind this is that since there are fewer occurrences of anomalous points these points should generate smaller paths due to two different reasons; First the bootstrapping aggregation procedure subsamples from the data and therefore a smaller number of partitions will induce a smaller path in a tree and second, the anomalous points are more likely to be filtered out earlier due to there extremity. The algorithm for the iTree is given below in Table 1 Given a sample time-series dataset

Data: X - training data, e- current tree height, l - height threshold
Result: an iTree
if $e \geq l$ *or* $|X| \leq 1$ **then**
  return exNode(Size $\leftarrow |X|$)
end
else
  Let Q be a a list of attributes of X
  randomly select an attribute q from Q
  randomly split a point p between a max and min within the attribute q
  $X_l = filter(X, q < p)$
  $X_r = filter(X, q \geq p)$
  return inNode(Left $\leftarrow$ iTree(X_l,e+1, l),
  Right $\leftarrow$ iTree(X_r,e+1, l),
  SplitAttribute $\leftarrow$ q,
  SplitValue $\leftarrow$ p)
end

Algorithm 1: iTree

X the typically decision tree procedure is performed where an attribute q is randomly selected along with a split value q to divide the data into a left and right tree. In the original paper this idea is emphasized in an experiment where the authors randomly partition a normal and anomalous point to demonstrate the average path length is notably shorter for anomalous data. The anomaly score is derived from the path length of a

binary tree in data structures; given a dataset of n instances the average path length of a binary tree is given by $c(n) = 2H(n-1)-(2(n-1)/n)$ where H(i) is the Harmonic number to compute the anomaly score given an ensemble of isolation trees for a point x the score is given by $s(x,n) = 2^{\frac{E(h(x))}{c(n)}}$ where $E(h(x)) = \frac{1}{L}\Sigma_{i=1}^{L} h_i(x)$ is the average score across an ensemble of isolation trees. The authors note that the isolation forest work well with small sampling of the data unlike many other methods which require large sampling of the data. The issues of swamping and masking in anomaly detection may occur due to oversampling; this has thought to be occurring because of anomalies being able to 'normalize' themselves; in fact masking and swamping is resolved by subsampling as data points are easier isolated if sample size is lower and different anomalies are detected by different iTree's as desired in the ensemble regime. The steps for obtaining the average path length of each tree is listed in the algorithm 2 table.

Data: x - an instance, T - an iTree, e - current path length; initialized to 0
Result: path length of x
if *T is an external node* **then**
  return e+ c(T.size)
end
a ← T.SplitAtt
if $x_a < T.splitValue$ **then**
  return PathLength(x.T.left,e+1)
end
else
  $x_a \geq$ T.splitValue
end
return PathLength(x, T.right, e+1)

Algorithm 2: PathLength

Where the ceiling height hyper-parameter for the max height is adjusted for subsampling size of the data and is given by $l = \log_2(\psi)$. In another paper they consider window sizes of time-series data; exploiting this idea of using smaller subsamples of the data to better detect anomalies. Concept drift is a phenomenon that occurs when the data being is updated live, since the model is trained on older data it may struggle to detect current data being predicted on; because of this the authors suggest that the sliding window hyper-parameter be found through brute force search. To deal with concept drift they retrain the model on a new window and either selectively keep certain models or disregard depending upon performance. They suggest that having a fixed window size for streaming data greatly hinders performance. In Table 1 in the appendix the procedure for synthesizing an iTree is listed. Lastly the steps for combining the iTree's into a forest off is given in algorithm table 3.

The authors suggest adjusting the subsample size ψ by accounting for kurtosis as it is sensitive to the presence of anomalies

Data: X - input data, t - number of trees, ψ - sub-sampling size
Result: a set of t iTrees
Initialize Forest set height $l = ceiling(\log_2 \psi)$ **for** *i=1 to t do* **do**
  $X' \leftarrow sample(X, \psi)$
  $Forest \leftarrow Forest \bigcup iTree(X', 0, l)$
end
return Forest

Algorithm 3: iForest

IV. LOCAL OUTLIER FACTOR

Let p be an element of a dataset X suppose $1 \leq MinPts \leq |D|$ and let $C_1, C_2, ...C_n$ be a partition of $N_{Minpts(p)}$. If defining $\eta_i = \frac{|C_i|}{N_{MinPts(p)}}$ and notions of $direct^i_{min}(p) = min(reach - dist(p,q)$,$direct^i_{max}(p) = max(reach - dist(p,q))$, $indirect^i_{min}(p) = min(reach - dist(p,o))$ and $indirect^i_{max}(p) = max(reach - dist(p,o))$ where q is a point in local neighborhood and q is a point in a nearby neighborhood; then the following bounds are imposed $LOF(p) \geq (\Sigma_{i=1}^{n} direct^i_{min}(p))(\Sigma_{i=1}^{n} \frac{\eta_i}{indirect^i_{max}}(p)$)and $LOF(p) \leq (\Sigma_{i=1}^{n} direct^i_{max}(p))(\Sigma_{i=1}^{n} \frac{\eta_i}{indirect^i_{min}}$; a much more sophisticated proof is provided in (LOF paper Breunig Kiegel) which is similar to proofs in elementary topological and advanced real analysis courses. Conceptually the LOF can be thought of as the reachability distances in the neighborhood of p relative to those in the indirect neighborhood.

ACKNOWLEDGEMENTS

[1] D. J. Griffiths,k *Introduction to Electrodynamics* (Cambridge University Press, Cambridge, 2017).

[2] A. Bobrinha, Revista Brasileira de Lorem Ipsum **23**, 179 (2002).

[3] R. P. Feynman, R. B. Leighton and M. Sands, *Lições de Física de Feynman* (Editora Bookman, Porto Alegre, 2008).

[4] J. D. Jackson, *Classical Electrodynamics* (John Wiley & Sons, Danvers, 1999).

Appendix: Appendix