Progress on Trajectory Anomaly Detection in Mining Massive Trajectory Data

A.K.L.Sriniwasa

Abstract

This report replicates the results of two papers on advanced anomaly detection models: Optimal Isolation Forest (OPTiF) and Deep Isolation Forests (DIF). OPTiF enhances traditional isolation methods by incorporating clustering-based learning to optimize isolation quality, mathematically and experimentally proving that the optimal branching factor for an isolation tree is Euler's number, e. DIF leverages neural networks for non-linear isolation, addressing the limitations of traditional and extended isolation forests. Both models are benchmarked against standard Isolation Forests using various datasets, demonstrating improvements in detection performance, and highlighting their respective advantages and disadvantages.

Paper 1: OPTiForest

Introduction

The paper is an isolation-based method called *Optimal Isolation Forest* (*OPTiF*)[1] which incorporates clustering based learning to hash which enables more information to be learned from data for better isolation quality. The paper also seeks to establish a theory on isolation efficiency to determine the optimal branching factor for an isolation tree.

Motivation

Despite the increasing importance of Anomaly Detection in various fields, and despite there being several anomaly detection methods, no theoretical work has answered the fundamental question of the optimal tree structure with respect to branching factor. This paper proves both mathematically and experimentally that the optimal branching factor has a value equal to the Euler's number e(which roughly equals 2.718). It also designed a practical optimal isolation forest OptIForest incorporating clustering based learning to hash which enables more information to be learned from data for better isolation quality.

Methodology

A practical isolation tree consists of two parts with the upper layers resulting from clustering and the lower layers from LSHiForest[2]. As our goal is to arrange the clusters into a natural hierarchy to form an isolation tree, agglomerative hierarchical clustering is adopted in our approach. The quality of clustering at upper levels is more sensitive to anomaly detection than that at lower levels. For the lower levels, the basic idea of our approach is to use an isolation tree efficiently produced in LSHiForest to initialise the clusters so that the clustering process begins with bigger initial clusters rather than the ones with a single data instance.

Experiments

First, datasets described in the next section were collected, and it was proven that the optimal branching factor was e. Then, the datasets were benchmarked against Isolation Forests[3] and the AUC ROC and AUC PR scores were compared. We also compare the runtime of the code.

Datasets

5 datasets were collected, and have been linked below. Of these 5 datasets, 4 of them namely ad, shuttle, Cardio, and Ionosphere were used in the original paper. The last dataset Pageblocks_16 was used in place of the backdoor dataset from the original paper used in the original paper.

Results

Of the 5 datasets used, ad, Pageblocks_16 and Ionosphere showed a significant increase of about 7-10% in their performance with the OPTiF model. Shuttle and Cardio however had significantly worse performance with OPTiF. However, for all 5 datasets, the time taken for OPTiF to run was significantly longer, with Ionosphere and Cardio taking almost 10,000 times longer to run compared to Isolation Forests, and the other 3 taking between 1001000 times as long to run.

Advantages and Disadvantages of this model

Advantages: The model in general produces better scores for most types of datasets compared to Isolation Forests.

Disadvantages: The time taken to run the model is very large, and lower thresholds in general take longer to run.

Paper 2: Deep Isolation Forests

Introduction

Deep Isolation Forests[4] (DIF) represent a promising advancement in anomaly detection, leveraging neural networks to achieve non-linear isolation. The DIF model offers strong representation power and has demonstrated promising results in various applications. The paper introduces the theory behind the DIF model, benchmarking it against Isolation Forests and Extended Isolation Forests[5] using tabular, graphical and time series datasets, in addition to experimentally proving its linear complexity both spatially and temporally. We have however only considered the tabular datasets part for replication.

Motivation

The Isolation Forest (IF) model operates by constructing isolation trees (iTrees) that isolate anomalies based on random partitioning. Despite its efficiency and simplicity, the IF model struggles with certain types of anomalies, especially those requiring complex, non-linear separations. Additionally, IF can create ghost regions, areas mistakenly marked as anomalies due to artifacts introduced during the isolation process. These limitations necessitate the development of more advanced models. Extended Isolation Forests (EIF) attempt to address some of these issues by employing hyperplane-based isolation, allowing for the consideration of multiple dimensions simultaneously. While this enhances the model's ability to handle more complex anomalies, EIFs still suffer from scalability issues and high memory usage, particularly with high-dimensional data. The need for a model that combines the strengths of these approaches while mitigating their weaknesses led to the development of the DIF model. Deep Isolation Forests leverage the power of deep learning to overcome the limitations of traditional isolation forests. By utilizing neural networks, DIF can achieve non-linear isolation, enhancing its ability to detect anomalies in complex datasets. The DIF model constructs a random representation ensemble with multiple isolation trees, enabling it to capture intricate patterns in the data that traditional models might miss.

Methodology

The Deep Isolation Forests (DIF) model enhances anomaly detection by combining isolation forest techniques with deep learning. It constructs a random representation ensemble using multiple isolation trees (iTrees), leveraging deep learning to capture complex, non-linear patterns in data. The process begins with the Computation Efficiency Representation Ensemble (CERE) method, which optimizes the creation of these trees, ensuring efficiency and scalability.

The DIF model evaluates the isolation difficulty of data points using the Deviation Enhanced Anomaly Scoring (DEAS) function. This function improves accuracy by providing a nuanced assessment of anomalies based on how challenging it is to isolate each data point within the iTrees. By integrating deep learning for complex pattern recognition and maintaining nearly linear time complexity, the DIF model offers a robust, scalable, and precise approach to anomaly detection.

Experiments

Three of the datasets mentioned in the original paper have been used in the first series of experiments. These datasets were benchmarked against the Isolation Forest and Extended Isolation Forest models. The metrics used for the benchmarking were the AUC-ROC score and the AUC-PR scores. After this, several synthetic datasets were generated, varying dimensions and sizes. These datasets were trained on all three algorithms, and their time complexity was compared.

Datasets

Three of the datasets mentioned in the original paper have been used, namely Fraud, Pageblocks and Shuttle. Fraud is for fraudulent credit card transaction detection. Pageblocks and Shuttle are from an anomaly benchmark study. For the synthetic datasets, 18 were generated. In the first 9, the number of data points was fixed at 5000 and the number of dimensions was varied from 16 to 4096 in multiples of 2. For the other 9 datasets, the number of features was fixed at 32, and the number of datapoints was varied from 1000 to 256,000 in multiples of 2. All synthetic datasets had 5% anomalies.

Results

DIF applied on Shuttle showed a 10% increase in both the metrics when compared against EIF, and an increase of nearly 8% when compared to the IF model on both metrics. DIF model applied on both Fraud and Pageblocks had no significant difference in the AUC-ROC score(less than 0.5%), but a significant increase was noticed in the AUC-PR scores. More

specifically for the AUC-PR scores, Fraud showed a 1% improvement compared to EIF and a 20% improvement from IF, and Pageblocks showed a 1% increase compared to EIF and a 7.1% increase when compared to the IF models.

For the time complexity, the log(Dimensions) vs log(T) and log(size) vs log(T) plots were all linear, thus proving that all three models are linear. In running time, IF was the fastest model, followed by DIF, with EIF being the slowest of all three models.

Advantages and Disadvantages of the model

Advantages: DIF is a linear time complexity deep learning model that fixes several of the issues present in the IF and EIF models.

Disadvantages: Being a deep learning model, the computational resources required are much more significant compared to the other models.

Conclusion

In conclusion, the replication of results from the OPTiF and DIF models demonstrates significant advancements in anomaly detection. OPTiF, with its optimal branching factor and clustering-based approach, shows enhanced isolation quality but suffers from longer computation times. DIF, leveraging deep learning for non-linear isolation, provides robust and scalable anomaly detection, though it requires substantial computational resources. Both models outperform traditional Isolation Forests in specific scenarios, confirming their potential for complex anomaly detection tasks. Future work should focus on optimizing these models for better efficiency and broader applicability in diverse real-world datasets.

Citations and References

[1] https://arxiv.org/abs/2306.12703

[2] X. Zhang et al., "LSHiForest: A Generic Framework for Fast Tree Isolation Based Ensemble Anomaly Analysis," 2017 IEEE 33rd International Conference on Data Engineering (ICDE), San Diego, CA, USA, 2017, pp. 983-994, doi: 10.1109/ICDE.2017.145. keywords: {Vegetation;Current measurement;Big Data;Algorithm design and analysis;Data mining;Benchmark testing;Feature extraction},

[3] F. T. Liu, K. M. Ting and Z. -H. Zhou, "Isolation Forest," 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 2008, pp. 413-422, doi: 10.1109/ICDM.2008.17. keywords: {Application software;Credit cards;Detectors;Constraint optimization;Data mining;Information technology;Laboratories;Isolation technology;Performance evaluation;Astronomy;anomaly detection;outlier detection;novelty detection;isolation forest;binary trees;model based},

[4] https://arxiv.org/pdf/2206.06602

[5] S. Hariri, M. C. Kind and R. J. Brunner, "Extended Isolation Forest," in IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 4, pp. 1479-1489, 1 April 2021, doi: 10.1109/TKDE.2019.2947676. keywords: {Forestry;Vegetation;Distributed databases;Anomaly detection;Standards;Clustering algorithms;Heating systems;Anomaly detection;isolation forest},