Evaluating the Sensitivity of Isolation Forest Parameters in Anomaly Detection

DT Nicolay 26296918
Computer Science Division
Stellenbosch University
Stellenbosch, South Africa
26296918@sun.ac.za

***Abstract*—TODO**
***Index Terms*—TODO**

I. Introduction

II. Background

A. Isolation Forests

The majority of existing model-based approaches to anomaly detection construct a profile of normal instances, then they identify instances that do not conform to this normal profile as anomalies [1]. Liu et al. (2008) proposed a fundamentally different model-based method that explicitly isolates anomalies instead of profiles normal points, Isolation Forests. Here, isolation refers to separating an instance from the rest of the instances. This is ideal for an anomaly detection problem context, since anomalies are by nature sparse and diverse.

Normal profile methods, since not optimised for anomaly detection, often lead to too many false positives or little to no anomalies detected at all. These methods are also constrained to low dimensional data and small data size since they require significant computational power. Isolation Forests on the other hand take advantage of anomaly datasets consisting of fewer observations for the target class, and anomalies having feature values distinct from the rest of the data. Due to the nature of anomaly observations, they are isolated closer to the root of the tree. This forms the foundation of Isolation Trees.

Isolation Forests involve an ensemble of Isolation Trees where the predicted anomalies are the observations with the shortest average paths across the trees.

B. Control Parameters

There are five control parameters to consider namely: the number of estimators, the maximum samples, the contamination, the maximum features, and whether to first bootstrap sample.

The number of trees parameter determines the number of base estimators in the ensemble. The performance of Isolation Forests converges quickly with a very small number of trees. [1].

The maximum samples describes the number of sample observations to draw from the training data for each base estimator. Only a small sampling size is required to achieve high detection performance with high efficiency [1].

Contamination refers to the proportion of anomalies present in the dataset. It is defined as the number of anomalies divided by total number of observations. When set to a specific value between 0 and 0.5, it determines the threshold on the anomaly scores such that approximately that fraction of the training samples are labelled as outliers. In the original paper, the threshold is automatically fixed at an offset of -0.5, following the original Isolation Forest formulation, where inliers typically yield scores near 0 and outliers near -1, allowing the model to separate them without prior knowledge of the true contamination level.

The maximum features describes how many features are selected at random before tree construction to train each base estimator.

The bootstrap parameter controls the manner in which sample observations are drawn for each tree. This determines whether sampling is done with or without replacement. Bootstrap resampling can lead to marginal improvements across classification metrics [2].

III. Implementation

A. test

IV. Empirical Process

V. Results & Discussion

VI. Conclusions

References

[1] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 2008, pp. 413–422.

[2] H. Choi and K. Jung, "Impact of data distribution and bootstrap setting on anomaly detection using isolation forest in process quality control," *Entropy*, vol. 27, no. 7, p. 761, July 2025.