Isolation-based Anomaly Detection

Anomalies are data points that are few and different. As a result of these properties, we show that, anomalies are susceptible to a mechanism called isolation. This paper proposes a method called Isolation Forest (iForest) which detects anomalies purely based on the concept of isolation without employing any distance or density measure - fundamentally different from all existing methods.

As a result, iForest is able to exploit subsampling (i) to achieve a low linear time-complexity and a small memory-requirement, and (ii) to deal with the effects of swamping and masking effectively. Our empirical evaluation shows that iForest outperforms ORCA, one-class SVM, LOF and Random Forests in terms of AUC, processing time, and it is robust against masking and swamping effects. iForest also works well in high dimensional problems containing a large number of irrelevant attributes, and when anomalies are not available in training sample.

1. INTRODUCTION

Anomalies are data patterns that have different data characteristics from normal instances. The ability to detect anomalies has significant relevance, and anomalies often provides critical and actionable information in various application domains. For example, anomalies in credit card transactions could signify fraudulent use of credit cards. An anomalous sport in an astronomy image could indicate the discovery of a new star. An unusual computer network traffic pattern could stand for an unauthorised access. These applications demand anomaly detection algorithms with high detection accuracy and fast execution.

Most existing anomaly detection approaches, including classification-based methods, Replicator Neural Network (RNN), one-class SVM and clustering-based methods, construct a profile of normal instances, then identify anomalies as those that do not conform to the normal profile. Their anomaly detection abilities are usually a 'side-effect' or by-product of an algorithm originally designed for a purpose other than anomaly detection (such as classification or clustering). This leads to two major drawbacks: (i) these approaches are not optimized to detect anomalies - as a consequence, these approaches often under-perform resulting in too many false alarms (having normal

instances identified as anomalies) or too few anomalies being detected; (ii) many existing methods are constrained to low dimensional data and small data size because of the legacy of their original algorithm.

This paper proposes a different approach that detects anomalies by isolating instances, without relying on any distance or density measure. To achieve this, our proposed method takes advantage of two quantitative properties of anomalies: i) they are the minority consisting of **few** instances, and ii) they have attribute-values that are very **different** from those of normal instances. In other words, anomalies are 'few and different', which make them more susceptible to a mechanism we called Isolation. Isolation can be implemented by any means that separates instances. We opt to use a binary tree structure called isolation tree (iTree), which can be constructed effectively to isolate instances. Because of the susceptibility to isolation, anomalies are more likely to be isolated closer to the root of an iTree; whereas normal points are more likely to be isolated at the deeper end of an iTree. This forms the basis of our method to detect anomalies. Although, this is a very simple mechanism, we show in this paper that it is both effective an efficient in detecting anomalies.

The proposed method, called Isolation Forest (iForest), builds an ensemble of iTrees for a given data set; anomalies are those instances which have short average path lengths on the iTrees. There are two training parameters and one evaluation parameter in this method: the training parameters are the number of trees to build and subsampling size; the evaluation parameter is the tree height limit during evaluation. We show that iForest's detection accuracy converges quickly with a very small number of trees; it only requires a small subsampling size to achieve high detection accuracy with high efficiency; and the different height limits are used to cater for anomaly clusters of different density.

2. ISOLATION AND ISOLATION TREES

In this paper, the term *isolation* means 'separating an instance from the rest of the instances'. In general, an isolation-based method measures individual instances' susceptibility to be isolated; and anomalies are those that have the highest susceptibility. To realize the ideal of isolation, we turn to a data structure that naturally isolates data. In randomly generated binary trees where instances are recursively partitioned, these trees

produce noticeable shorter paths for anomalies since (a) in the regions occupied by anomalies, less anomalies result in a smaller number of partitions - shorter paths in a tree structure, and (b) instances with distinguishable attribute - values are more likely to be separated early in the partitioning process. Hence, when a forest of random trees collectively produce shorter path lengths for some particular points, they are highly likely to be anomalies.

Definition: Isolation Tree. Let T be a node of an isolation tree. T Is either an external-node with no child, or an internal-node with one test and exactly two daughter nodes (T_l, T_r). A test at node T consists of an attribute q and a split value p such that the test $q < p$ determines the traversal of a data point to either T_l or T_r.

Let $X = \{x_1, ..., x_n\}$ be the given data set of a d-variate distribution. A sample of ψ instances $X' \subset X$ is used to build an isolation tree (iTree). We recursively divide X' by randomly selecting an attribute q and a split value p, until either: (i) the node has only one instance or (ii) all data at the node have the same values. An iTree is a *proper binary tree*, where each node in the tree has exactly zero or two daughter nodes. Assuming all instances are distinct, each instance is isolated to an external node when an iTree is fully grown, in which case the number of external nodes is ψ and the number of internal nodes is $\psi - 1$; the total number of nodes of an iTrees is $2\psi - 1$; and thus the memory requirement is bounded and only grows linearly with ψ.

The task of anomaly detection is to provide a ranking that reflects the degree of anomaly. Using iTrees, the way to detect anomalies is to sort data points according to their average path lengths; and anomalies are points that are ranked at the top of the list. We define path length as follow:

Definition: Path Length $h(x)$ of a point x is measured by the number of edges x traverses an iTree from the root node until the traversal is terminated at an external node.

We employ path length as a measure of the degree of susceptibility to isolation:

- short path length means high susceptibility to isolation,
- long path length means low susceptibility to isolation.

3. ISOLATION, DENSITY AND DISTANCE MEASURES

In this paper, we assert that path-length-based isolation is more appropriate for the task of anomaly detection than the basic density and distance measures.

Using basic density measures, the assumption is that '*Normal points occur in dense regions, while anomalies occur in sparse regions*'. Using basic distance measures, the basic assumption is that '*Normal point is close to its neighbours and anomaly is far from its neighbours*'.

There are violations to these assumptions, e.g., high density and short distance do not always imply normal instances; likewise low density and long distance do not always imply anomalies. When density or distance is measured in a local context, which is often the case, points with high density or short distance could be anomalies in the global context of the entire data set. However, there is no ambiguity in path-length-based isolation and we demonstrate that in the following three paragraphs.

In density based anomaly detection, anomalies are defined to be data points in regions of low density. Density is commonly measured as (a) the reciprocal of the average distance to the k-nearest neighbours (the inverse distance) and (b) the count of points within a given fixed radius.

In distance based anomaly detection, anomalies are defined to be data points which are distant from all other points. Two common ways to define distance-based anomaly score are (i) the distance to k^{th} nearest neighbour and (ii) the average distance to k-nearest neighbours. One of the weaknesses in these density and distance measures is their inability to handle data sets with regions of different densities. Also, for these methods to detect dense anomaly clusters, k has to be larger than the size of the largest anomaly cluster. This creates a search problem: finding an appropriate k to use. Note that a large k increases the computation substantially.

On the surface, the function of an isolation measure is similar to a density measure or a distance measure, i.e., isolation ranks scattered outlying points higher than normal points. However, we find that path length based isolation behaves differently form a density or distance measure, under data with different distributions. Path length, however is able to address this situation by giving the isolated dense points shorter path lengths. The main reason for this is that path length is grown in adaptive context, in

which the context of each partitioning is different, from the first partition (the root node) in the context of the entire data set, to the last partition (the leaf node) in the context of local data-points. However, density ($k-nn$) and $k^{th}nn$ distance only concern with k neighbours (local context) and fail to take the context of the entire data set into consideration.

In summary, we have compared three fundamental approaches to detect anomalies; they are isolation, density and distance. We find that the isolation measure (path length) is able to detect both clustered and scattered anomalies; whereas both distance and density measures can only detect scattered anomalies. While there are many ways to enhance the basic distance and density measures, the isolation measure is better because no further 'adjustment' to the basic measure is required to detect both clustered and scattered anomalies.