Final Project Report

Anomaly detection using isolation forest with the help of h20.ai

CSE4003- Cyber security

Submitted by

Name	Reg no
Aditya Firoda	16BCE2184
Garima Sondhi	16BCE2193
Jagruta Advani	16BCE2159
Vartika	16BCI0192

Under the guidance of

Prof. Lavanya K

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

ABSTRACT

Anomaly detection is a method of finding suspicious events or any data items that might create problem for the concerned authorities. Anomalies in a data can indicate problems like security problems, server crashes, etc. Anomaly detection algorithm has a lot of applications in our day to day lives. Some applications where anomaly detection can be used are in bank fraud, structural defect in some buildings, some medical errors, etc. This project revolves around the use of isolation forest for anomaly detection of a fraud in a credit card transaction using h2o.ai and scikit. Fraud detection mainly refers to detection of criminal activities which occur in many organizations such as banks, insurance companies, stock exchange market, etc. The archetypal method of anomaly detection procedures is to keep up a usage profile for each client and monitor the profiles to notice any deviations. The isolation forest has been not been a much explored method in the domain of anomaly detection. H2o.ai and scikit both are used in machine learning and artificial intelligence algorithms. Both of them can be used from different environments like r, python, etc. In this project we have used python as our environment to implement the anomaly detection technique. Also we will take the help of two performance measures namely – AUC and AUCPR for getting an estimate about the quality of the score in the results.

INTRODUCTION

Anomaly detection is the problem of finding odd patterns in a dataset that do not show regular behaviour. These odd things are generally referred as noise, outliers, exceptions and deviations. All these terms are used interchangeably. Anomaly detection has
a lot of practicality in applications. Some of the applications include fraud detection in credit card, surveillance of enemy activities by military, intrusion detection in cyber- security, etc. The significance of anomaly detection is thanks to the very fact that anomalies in information square measure rendered into essential and sensitive information in varied application domains. As an example, anomalous patterns in traffic in a very computer network may mean that a hacked laptop is causing out sensitive information to an unauthorized destination. Anomalous image might indicate presence of malignant tumor. Anomalies in mastercard dealings information may indicate MasterCard fraud or theft of identity or abnormal readings from a spacecraft device may signify a fault in some part of the spacecraft. This method can be applied in three ways which are supervised, semi-supervised and unsupervised. In a supervised learning, the person should have the prior information about the observations and data items whether they are genuine or anomalous and then this information is used during the training of the dataset. Here building a predictive model is a typical approach in such cases. After building the model an unseen data is compared with the model and then the class is determined where that particular data belongs to. In semi-supervised learning only the information about genuine data is known. There is no idea about the anomalous data and therefore only the genuine dataset is used for training. They are more widely used than the supervised techniques. Here the model is build using the genuine dataset and then the given dataset is compared against this model. In unsupervised inferences from dataset are drawn without labelled responses. This is the most widely accepted out of all three methods. This method makes an assumption that normal instances are more frequent than the exceptions in a test data. if this assumption is wrong then this method has a high false alarm rate.

Other algorithms used for anomaly detection identify anomalies with the help of profiling normal data points but isolation tree is an ensemble method. It creates a tree like structure which helps to make decisions. Partitions are created here by first randomly selecting the different features and then splitting is performed between the maximum and minimum values of the features present. Then the concept that the outliers or anomalies are less frequent than the other regular points is used. These anomalies can be detected near the root of the tree and then can be further analysed.
Then in this project we will use two frameworks –h20.ai and scikit to implement isolation forest and will compare the results of both these libraries in the python environment.

LITERATURE SURVEY

1. Anomaly-based network intrusion detection: Techniques, systems and challenges

P. Garcı´a-Teodoroa, J. Dı´az-Verdejoa, G. Macia´-Erna´ndeza, E. Va´zquezb

The Internet associated computer networks are exposed to an increasing range of security threats. With new kinds of attacks showing frequently, developing versatile and accommodative security orienting approaches may be a severe challenge. during this context, anomaly-based network intrusion detection techniques area unit a valuable technology to safeguard target systems and networks against malicious activities. However, despite the range of such strategies delineated in recent years, security tools integrating anomaly detection characterstics are simply beginning to be possible, and several other necessary issues stay to be resolved. This paper begins with a review of the foremost popularly known anomaly-based intrusion detection practices. Then, accessible platforms, systems below development and analysis comes within the space area unit conferred. Finally, we have a tendency to define the most challenges to be controlled for the wide scale preparation of anomaly-based intrusion detectors, with special stress on assessment problems.

2. Network Anomaly Detection by Cascading K-Means Clustering and C4.5 Decision Tree algorithm

Amuthan Prabakar Muniyandia, R. Rajeswarib, R. Rajaramc

International Conference on Communication Technology and System Design 2011

In this paper, authors propose an anomaly detection methodology using "K-Means + C4.5", a way to cascade k-Means cluster and therefore the C4.5 call tree methods for classifying abnormal and real activities in a computer network. The k-Means cluster methodology is initial accustomed to partition the coaching instances into k clusters using Euclidean distance similarity. On every cluster, representing a density region of traditional or anomaly instances, we build decision trees using C4.5 call tree rule. the decision tree on every cluster refines the choice boundaries by learning the subgroups inside the cluster. to get a final conclusion we tend to exploit the results derived from the decision tree on every cluster.

3. An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection

Shih-Wei Lin, Kuo-Ching Ying, Chou-Yuan Lee, Zne-Jung Lee

Published in Appl. Soft Comput. 2012 DOI:10.1016/j.asoc.2012.05.004

In this paper shih-Wei Lin et al. proposes By analysing the knowledge from using KDD'99 dataset, DT and SA will acquire decision rules for brand new attacks and
can improve accuracy of classification. additionally, the simplest parameter settings for the DT and SVM area unit mechanically adjusted by SA. The planned algorithmic program outperforms alternative existing approaches. Simulation results demonstrate that the planned algorithmic program is no-hit in detection of anomaly intrusion detection.

4. A survey of network anomaly detection techniques

Mohiuddin Ahmed, Abdun Naser Mahmood, Jiankun Hu

Published in Journal of Network and Computer Applications DOI: 10.1016/j.jnca.2015.11.016

This paper anomaly detection is a vital information analysis task that is helpful for distinctive the network intrusions. This paper presents an in-depth analysis of 4 major classes of anomaly detection

techniques that embrace classification, applied math, scientific theory and agglomeration. The paper conjointly discusses analysis challenges with the datasets used for network intrusion detection.

5. Isolation Forest

Fei Tony Liu, Kai Ming Ting, Zhi-Hua Zhou
Published in:Proceeding ICDM '08 Proceedings of the 2008 Eighth IEEE International Conference on Data Mining Pages 413-422
The paper proposes a fundamentally different model-based method that isolates anomalies instead of profiles normal points. The notion of isolation has not been explored very much in current literature iForest take the advantage of sub-sampling to an extent that's not possible in existing ways, making a rule that encompasses a linear time quality with a lesser constant and lesser memory demand. The paper performs an empirical evaluation that shows iForest performs better in comparison to ORCA which is a near-linear time complexity distance-based method, LOF and Random Forests in terms of AUC and processing time, and especially in huge data sets. These also works very well with problems having a large number of irrelevant attributes, and in circumstances in which the training set does not have any anomalies.

METHODOLOGY:

There are multiple approaches to an unattended anomaly detection downside that try and exploit the variations between the properties of common and distinctive observations. the thought behind the Isolation Forest is as follows.

1. Dataset

The project detects the anomaly by detecting the credit card fraud transaction recorded this included all the transaction with a time frame and all the values associated with transaction including amount of transaction. The dataset contains around 500 fraudulent activities and 290000 genuine transaction recorded. The data set is been taken from GitHub and Kaggle.

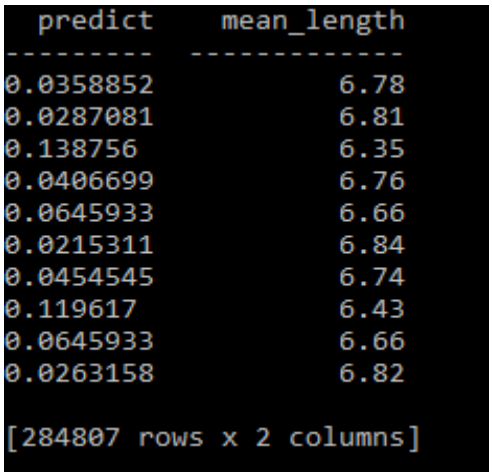
2. Anomaly Detection

2.1 We begin by building multiple decision trees such the trees isolate the observations in their leaves. Ideally, every leaf of the tree isolates precisely one observation from your information set. The trees are being split every which way. we have a tendency to assume that if one observation is comparable to others in our information set, it'll take a lot of random splits to absolutely isolate this observation, as against segregating an outlier from other data points.
2.2 For all the data points which has values considerably totally different from the opposite observations, every which way finding the split isolating it shouldn't be too exhausting. As we have a tendency to build multiple isolation trees, therefore the isolation forest, for every observation we are able to calculate the common variety of splits across all the trees that segregate the observation. the most common type of split among all the found are then taken into consideration, wherever the less splits the observation wants, the more likely it's to be abnormal.

```
import h2o
from sklearn.metrics import roc_curve, precision_recall_curve, auc
import matplotlib.pyplot as plt
import numpy as np
from tqdm import tqdm_notebook
h2o.init()

df = h2o.import_file("creditcard.csv")
seed = 12345
ntrees = 100
isoforest = h2o.estimators.H2OIsolationForestEstimator(
    ntrees=ntrees, seed=seed)
isoforest.train(x=df.col_names[0:31], training_frame=df)
predictions = isoforest.predict(df)
```

Fig. 1. The initial code for importing the dataset and starting the h20 instance and giving initial predictions .



predict	mean_length
0.0358852	6.78
0.0287081	6.81
0.138756	6.35
0.0406699	6.76
0.0645933	6.66
0.0215311	6.84
0.0454545	6.74
0.119617	6.43
0.0645933	6.66
0.0263158	6.82

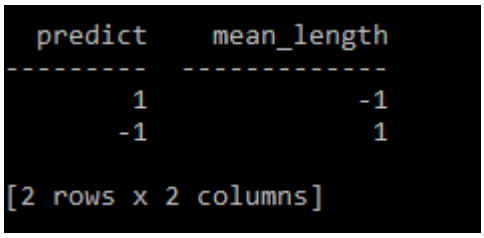
[284807 rows x 2 columns]

Fig2. Intial normalized predicted length and the mean length for the multiple decision trees created.

3 Inspecting Predictions

On print the results of the h20 frame containing the predictions: we forecast showing a normalized incongruity score, and mean length showing the shared variation of splits across all trees to separate the remark.

These 2 columns ought to have the inversely proportional because of the property satisfied the less random splits you would like to isolate the observation, the lot of abnormalities. We are able to simply ensure.



predict	mean_length
1	-1
-1	1

[2 rows x 2 columns]

Fig 2: The correlation matrix between predicted and mean length column

Table 1: Input variables

INPUT VARIABLE	CATEGORY
Time	Numerical
V1,V2,..V28	Numerical
Amount	Numerical

Table 2: Output variable

OUTPUT VARIABLE	CATEGORY
Class	Binary

4 Predicting Anomalies using Quantiles

As we tend to formulated this downside in an unsupervised fashion, however will we go from the common variety of splits / anomaly score to the particular predictions? employing a threshold! If we've got a thought regarding the relative variety of outliers in our dataset, we are able to notice the corresponding quantile price of the score and use it as a limiting value for the predictions made by the h20 frame created by us. we are able to use the edge to predict the abnormal category

Probs	predictQuantiles	mean_lengthQuantiles
0.95	0.138756	6.99

[1 row x 3 columns]

Fig3. Probablity of predicting quantiles and mean length quantiles for the dataset

predict	mean_length	predicted_class	class
0.0358852	6.78	0	0
0.0287081	6.81	0	0
0.138756	6.35	0	0
0.0406699	6.76	0	0
0.0645933	6.66	0	0
0.0215311	6.84	0	0
0.0454545	6.74	0	0
0.119617	6.43	0	0
0.0645933	6.66	0	0
0.0263158	6.82	0	0

[284807 rows x 4 columns]

Fig4. Results of the predicted class with the actual class of the dataset

5 Evaluation

Since the isolation forest is an unsubstantiated technique, it is sensible to possess a glance at the classification metrics that aren't captivated with the prediction threshold and provides an estimate of the standard of evaluation. 2 such metrics square measure space underneath the Receiver operative graphical record (AUC) and space underneath the Precision-Recall Curve (AUCPR).

AUC may be a metric evaluating however well a binary classification model of such type distinguishes true positives from false positives and in such case the right AUC score is a maximum of one; the minimum admissible score of an arbitrary estimation is 0.5.

AUCPR may be a metric evaluating the exactness recall trade-off of a binary classification exploitation totally different thresholds of the continual prediction score. The right AUCPR score is 1; the baseline score is that the relative count of the positive category.

For extremely unbalanced information, AUCPR is suggested over AUC because the AUCPR is a lot of sensitive to True positives, False positives and False negatives, whereas not caring concerning True negatives, that in great quantity typically overshadow the result of different metrics.

RESULT

Looking at the results of the twenty runs, we are able to see that the binary compound isolation forest implementation on the average scores equally to the scikit-learn implementation in each FTO and AUCPR. The large advantage of binary compound is that the ability to simply rescale too many nodes and work seamlessly with Apache Spark exploitation drinking water. This enables you to method extraordinarily giant datasets, which could be crucial within the transactional information setting.

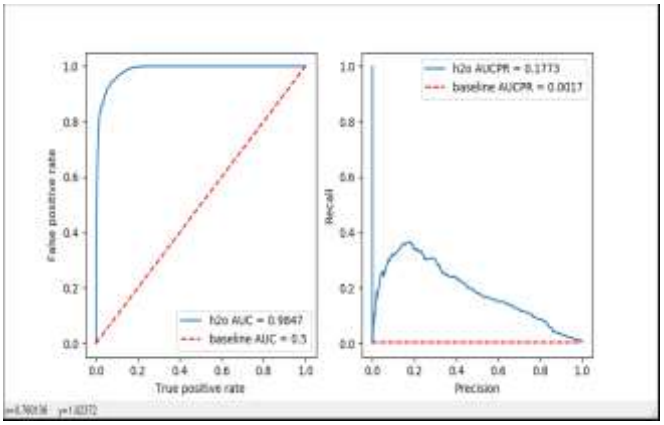


Fig5. This shows the baseline AUCPR and h20 AUCPR for true positive vs true negative and for precision vs recall for the dataset.

CONCLUSION

In this project, we propose a methodology that shows incredible potential in identifying anomalies and pure inliers by making a decision tree for each data point. We showed that application of proposed method in different scenarios such as to make recommendation for potential outliers for further investigation with high precision and to create training sets for novelty detection algorithms. We also claimed for a better evaluation metric and showed that area under the precision recall curve is a more improved method than area under the ROC curve. Lastly, we showed the efficiency of our method in a fraud detection dataset.

REFERENCES

[1] Books for Python, Learning Python by David Ascher Mark Lutz , Edition 1999.

[2] Python Tutorials point, https://www.w3schools.com/python/

[3] Jianshu Sun **,** Yuansheng Lou, Feng Ye(2017). **Research on Anomaly Pattern Detection in Hydrological Time Series**, Web information Systems and Applications Conference(WISA) **IEEE Conferences**.

[4] A. Asuncion and D. Newman. UCI machine learning repository, 2007

[5] Anomaly-based network intrusion detection: Techniques, systems and challenges by P. Garcı´a-Teodoroa, *, J. Dı´az-Verdejoa , G. Macia´-Ferna´ndeza , E. Va´zquezb, 2008

[6] Anomaly Detection: A Survey by Varun Chandola, Arindam Banerjee, and Vipin Kumar August 15, 2007

[7] Anomaly Detection in Crowded Scenes by Vijay Mahadevan Weixin Li Viral Bhalodia Nuno Vasconcelos,2010

[8] Isolation-based Anomaly Detection by Fei Tony Liu and Kai Ming Ting,2010

[9] Anomaly Based Network Intrusion Detection with Unsupervised Outlier Detection by Jiong Zhang and Mohammad Zulkernine,2011

[10] Fast Anomaly Detection for Streaming Data by Swee Chuan Tan ,Kai Ming Ting and Tony Fei Liu,2008