Outlier Detection

Raghav Singh-21074022

Ujjawal Modi-21074032

Project Supervisor - Dr. Bidyut Kumar Patra

Department of Computer Science and Engineering
Indian Institute of Technology (BHU)
Varanasi, India

May 8, 2024

Introduction

- What is an outlier?
  - An outlier is an abnormal data point which is greatly different from the rest of the data.
  - Arise due to malicious actions, system failures, intentional fraud, etc.
- Why is outlier detection important?
  - Anomalies in credit card transactions could signify fraudulent use of credit cards.
  - Anomalous spot in an astronomy image could indicate the discovery of a new star.
  - An unusual computer network traffic pattern could stand for an unauthorised access.

Literature Survey

Paper Title and Author	Work Done	Shortcomings
Anomaly Detection with Robust Deep Autoencoders: Chong Zhou, Randy C. Paffenroth	Performed anomaly detection using Robust Deep Autoencoders, in the absence of availability of a clean noise free dataset. Introduced an anomaly regularising penalty using vector norms.	Cost function is not convex and thus will not always converge to a global optima. Very high computational complexity. Sensitive to hyperparameters.
Unsupervised Anomaly Detection With LSTM Neural Networks: Tolga Ergen and Suleyman Serdar Kozat	Obtained fixed-length representation of variable-length sequence using LSTM-based structure. Found a decision function for anomaly detectors based on OC-SVM and SVDD algorithms.	Mainly used for time-series data. High computational complexity because of joint training of LSTM and OC-SVM.
Deep Semisupervised Anomaly Detection: Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, Marius Kloft	Introduced Deep SAD, a generalization of the unsupervised Deep SVDD method to the semi-supervised AD setting. Introduced an information-theoretic framework based on lower entropy in normal data than anomalous data.	Model requires access to a small pool of labeled samples, e.g. a subset verified by some domain expert as being normal or anomalous.
Deep Anomaly Detection Using Geometric Transformations: Izhak Golan, Ran El-Yaniv	Trained a multi-class neural classifier over a self-labeled dataset created from the normal instances and their transformed versions, obtained by applying numerous geometric transformations.	Can only be used for image data, cannot be generalised to general anomaly detection. Sensitive to specific geometric transformations. Requires dataset with all normal samples during training.

Literature Survey

Paper Title and Author	Work Done	Shortcomings
Classification-Based Anomaly Detection for General Data: Liron Bergman, Yedid Hoshen	Semi supervised approach. Transforms the training data into M subspaces, learning a feature space. From the learned features, the distance from the cluster center is used for detection.	Requires a dataset containing only normal instances for training.
ECOD: Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions: Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, George H. Chen	Estimates the distribution of input data by computing the empirical cumulative distribution for each dimension. Tail probabilities are estimated per dimension and aggregated to compute an outlier score.	Considers dimensions to be independent of each other. Cannot handle multimodal distributions for which an outlier could be in neither left nor right tails.
A Novel Outlier Detection Method for Multivariate Data: Yahya Almardeny, Noureddine Boujnah and Frances Cleary	Decomposed the full attribute space into 3D subspaces and rotated the 3D vectors about the geometric median, using Rodrigues rotation formula, to construct the overall outlying score.	Very high time complexity so cannot be used for large datasets with many features. Cannot be used for univariate or bivariate data.
Deep Anomaly Detection with Deviation Networks: Guansong Pang, Chunhua Shen and Anton van den Hengel	Used a neural anomaly score learner to assign the data points an anomaly score. Leveraged labeled anomalies and a Gaussian prior to optimise anomaly scores using a Z-Score-based deviation loss to detect anomalies.	Requires prior knowledge of some anomalies in the dataset, which is not available in unsupervised learning. Assumes Gaussian distribution for the data to estimate anomalies.

Local Outlier Factor (LOF)

- Returns an outlier factor for each object, which is the degree of being outlying.
- Depends on isolation of object with respect to the surrounding neighborhood.
- The reachability distance of object p with respect to object o is defined as

$$\text{reach-dist}_k(p, o) = \max\{\text{distance}_k(o), d(p, o)\}$$

- The local reachable density (lrd) of p is defined as

$$\text{lrd}_k(p) = 1/\left(\frac{\sum_{o \in N_\text{k}(p)} \text{reach-dist}_k(p, o)}{|N_\text{k}(p)|}\right)$$

  inverse of average reachability distance of p based of k-nearest neighbours.
- The local outlier factor(LOF) of p is defined as

$$\text{LOF}_k(p) = \frac{\sum_{o \in N_\text{k}(p)} \frac{\text{lrd}_k(o)}{\text{lrd}_k(p)}}{|N_\text{k}(p)|}$$

- The lower p's lrd is, and the higher the lrd of p's k-nearest neighbors are, the higher is the LOF value of p.
- For most objects p in a cluster, the LOF of p is approximately equal to 1.

Isolation Forest

- Explicitly isolates outliers instead of profiling normal instances.
- Isolation forest consists of a collection of isolation trees.
- Each isolation tree successively partitions points using a random attribute and random threshold.
- Anomalies are isolated close to root of trees, and thus have a shorter average path length in the forest.
- For a data point x, average path length $E(h(x))$ over all possible trees is estimated as the mean path length in the forest.
- The outlier score for an instance x in a dataset of n instances is defined as

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}$$

  wherein the average path lengths are normalised using $c(n)$,
- $c(n)$ is the average path length of unsuccessful search in BSTs, and is calculated as

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n}$$

- $H(i)$ is the harmonic number estimated as $ln(i) + 0.5772156649$ (Euler's constant).

GAN based approach

- Approaches the problem as a binary-classification issue.
- Assumes that the entire dataset contains only normal instances.
- Generates informative potential outlier that occurs close to the real data.
- Optimization process of a GAN can be written as

$$\min_{\theta_g} \max_{\theta_d} V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}} [\log D(x)] + \mathbb{E}_{z \sim p_z} [\log(1 - D(G(z)))].$$

- With single generator all generated outliers will occur inside or close to a part of real data as the training is progressed.
- Use multiple generator, each generator will generate outliers close to part of data.

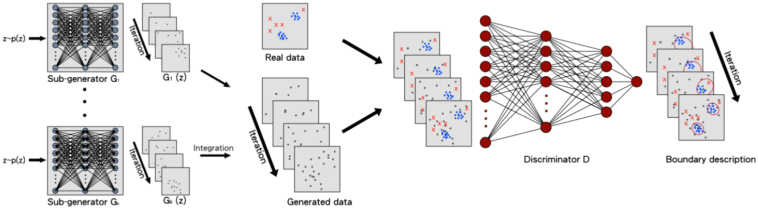


Figure 1: Process of detecting outliers. Blue dots represents inliers and Red cross represents outliers.

One Class Neural Networks based approach

- One-Class SVM (OC-SVM) is used for unsupervised anomaly detection.
- One-Class SVM considers all instances of dataset as positive and learns a hyperplane around the data.
- The parameters of the boundary are determined using the following optimisation problem:

$$\min_{w,r} \frac{1}{2}\|w\|_2^2 + \frac{1}{\nu} \cdot \frac{1}{N} \sum_{n=1}^{N} \max\left(0, r - \langle w, \Phi\left(\mathbf{X}_{n:}\right)\rangle\right) - r$$

W and b are the normal vector and bias for the boundary, N is the number of instances, Φ is the kernel map and parameter ν is the fraction of points allowed to cross the hyperplane.

- One-Class Neural Network uses a simple feed forward network with one hidden layer having activation $g()$ and one output node.
- OCNN modifies the objective of OC-SVM as follows:

$$\min_{w,V,r} \frac{1}{2}\|w\|_2^2 + \frac{1}{2}\|V\|_F^2 + \frac{1}{\nu} \cdot \frac{1}{N} \sum_{n=1}^{N} \max\left(0, r - \langle w, g\left(V\mathbf{X}_{n:}\right)\rangle\right) - r$$

where V is the weight matrix from hidden to output layer.

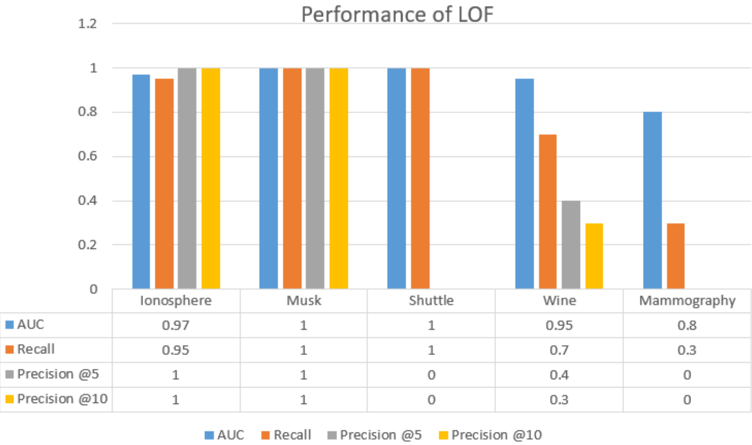
- w and V are optimised using standard backpropagation. Optimisation of r is a quantile selection problem.

Dataset Description

Dataset	No. of instances	Dimension	Outlier count	Outlier percentage(%)
Ionosphere	255	33	35	5.85
Musk	2975	166	10	0.33
Shuttle	45636	9	50	0.109
Wine	4918	12	20	0.406
Mammography	10973	6	50	0.455

Table 1: Dataset description.

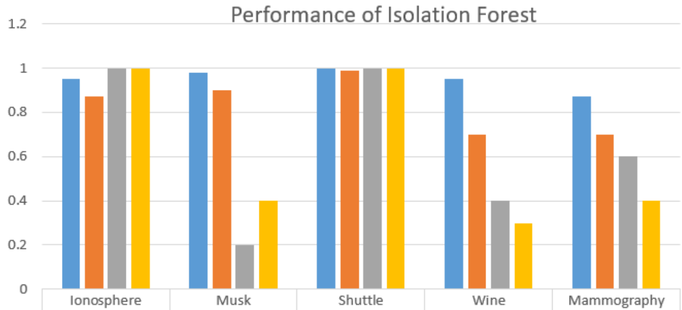
Results



	Ionosphere	Musk	Shuttle	Wine	Mammography
AUC	0.97	1	1	0.95	0.8
Recall	0.95	1	1	0.7	0.3
Precision @5	1	1	0	0.4	0
Precision @10	1	1	0	0.3	0

Figure 2: LOF Performance

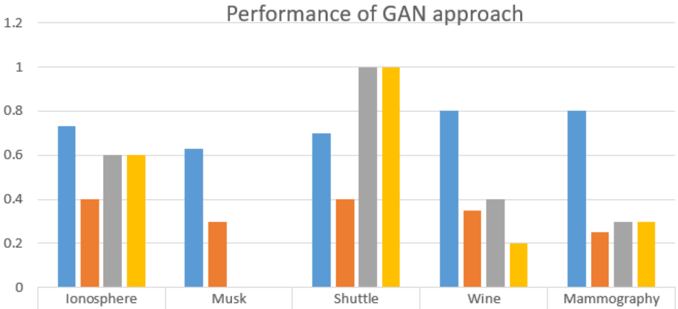
Results



	Ionosphere	Musk	Shuttle	Wine	Mammography
AUC	0.95	0.98	1	0.95	0.87
Recall	0.87	0.9	0.99	0.7	0.7
Precision @5	1	0.2	1	0.4	0.6
Precision @10	1	0.4	1	0.3	0.4

Figure 3: Isolation Forest Performance

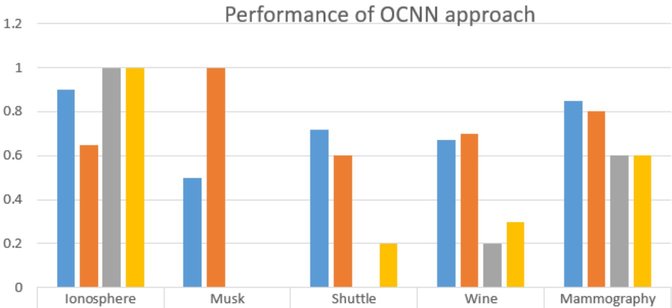
Results



	Ionosphere	Musk	Shuttle	Wine	Mammography
AUC	0.73	0.63	0.7	0.8	0.8
Recall	0.4	0.3	0.4	0.35	0.25
Precision @5	0.6	0	1	0.4	0.3
Precision @10	0.6	0	1	0.2	0.3

Figure 4: GAN Performance

Results



	Ionosphere	Musk	Shuttle	Wine	Mammography
AUC	0.9	0.5	0.72	0.67	0.85
Recall	0.65	1	0.6	0.7	0.8
Precision @5	1	0	0	0.2	0.6
Precision @10	1	0	0.2	0.3	0.6

AUC Recall Precision @5 Precision @10

Figure 5: OCNN Performance

Conclusion and Future Commitment

- Several existing approaches for outlier detection have been studied and implemented.
- Deep learning approaches have been compared with state of the art methods.
- Deep learning results are promising, and have great scope of improvement.
- Our next step will be to come up with a novel method for outlier detection.
- We plan on using deep learning approaches using GANs and autoencoders.
- Clustering based approaches shall also be explored.
- Domain specific outlier detection methods, such as on image and time series data can also be explored.