Deep Anomaly Detection

Joel Mbouwe

DataLab GBIS/CDO

19 Juillet 2020

Plan

Introduction

- An anomaly is « *an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism.* » Hawkings.

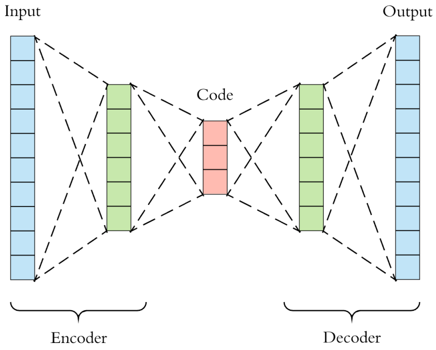
- An anomaly detection model is a model that learns how to characterize the normality of the data and how far samples deviate from that normality.

- Part of my internship consisted precisely in making a state of the art of deep learning techniques for anomaly detection.

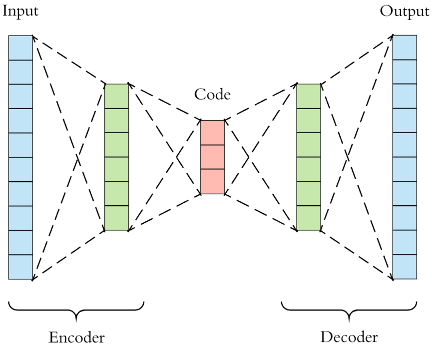
AutoEncoder for Anomaly Detection

Approach

- Model for learning a low dimensional representation of data
- Encoder for dimension reduction and the decoder for the reconstruction of the data
- The learning process is done by minimizing the reconstruction error : $\|\hat{X} - X\|_2$
- We except high reconstruction error for abnormal data points since the model is forced to capture only the essentials characteristic of the data.



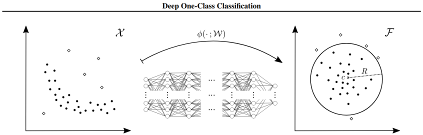
AutoEncoder for Anomaly Detection
Illustration



Deep SVDD

It is a deep learning method for anomaly detection where the goal is to train a neural network while minimizing the volume of a hyper-sphere that encloses the network representations of the data.

- The model is forced to extract the common features of the data that enables the contraction.
- The Outlier score is defined as the distance to the center $\|\phi(\boldsymbol{x}_i, \mathcal{W}) - \boldsymbol{c}\|^2$
- Point of attention : No bias and upper bounded activation functions in the network otherwise the model will map the data to the center



Objective function :

$$\min_{\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n} \|\phi(\boldsymbol{x}_i, \mathcal{W}) - \boldsymbol{c}\|^2 + \frac{\lambda}{2}\sum_{\ell=1}^{L} \left\|\boldsymbol{W}^{\ell}\right\|_F^2$$

$$\min_{\mathcal{W},R} R^2 + \frac{1}{n}\sum_{i=1}^{n} max\{0, \|\phi(\boldsymbol{x}_i, \mathcal{W}) - \boldsymbol{c}\|^2 - R^2\} + \frac{\lambda}{2}\sum_{\ell=1}^{L} \left\|\boldsymbol{W}^{\ell}\right\|_F^2$$

Distance based anomaly detection models

K-nearest neighbors

Anomaly score is modeled by the mean distance between a sample and its K nearest neighbors

- Not adapted for high dimensional data and it is time consuming
- Not suited for group outlier detection since it will require a high value of K

Least Similar Nearest Neighbor : Lesinn

It is a random distance based outlier detection. Given a data object x_i,the approach defines its outlierness as follows : $r_i = \frac{1}{m} \sum_1^m nn\_dist(x_i|S_j)$ where S_j X is a random data subsample, m is the ensemble size, and nn_dist returns the nearest neighbor distance of x_i in S_j

- Faster approach that can leverage better results than KNN and it is robust to group anomalies

REPEN

Framework to learn low-dimensional representations and uses a distance-based outlier detection approach to learn a set of features that can discriminate normal and abnormal data.
More formally, given a distance-based outliers function ϕ (KNN, Lesinn etc.) the goal is to learn a mapping function f such that $\phi(f(x_{abnormal})) > \phi(f(x_{normal}))$

- Either ϕ is applied on the original data to obtain sets of inlier and outlier candidates or there is a small set of labeled anomalies.
- Each batch point is composed of a triplet $(query, x_+, x_-)$. The sampling is done by fitting a probability distribution depending on the score obtained previously.

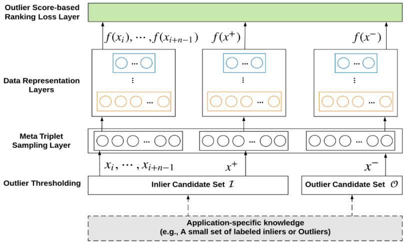


Figure 1: The Proposed RAMODO Framework. RAMODO learns a representation function $f(\cdot)$ to map D-dimensional input objects into a M-dimensional space, with $M \ll D$.

REPEN

- The goal is to learn representation for which the pseudo outlier x has a larger nearest neighbor distance in Q than the pseudo inlier x^+

$$\mathcal{L} = \max\Big[0, c + \text{nn\_dist}\left(f_\Theta\left(\mathbf{x}^+\right) \mid f_\Theta(Q)\right) - \text{nn\_dist}\left(f_\Theta\left(\mathbf{x}^-\right) \mid f_\Theta(Q)\right)\Big]$$

- Inference : The anomality score of a sample x is defined by $\phi(f_\Theta(x))$

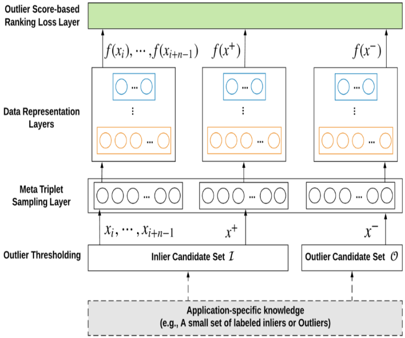


Figure 1: The Proposed RAMODO Framework. RAMODO learns a representation function $f(\cdot)$ to map D-dimensional input objects into a M-dimensional space, with $M \ll D$.

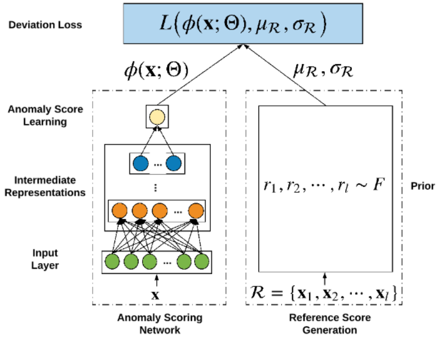
Deviation Network

A model that directly learns an anomaly score function ϕ_Θ such that

$$\phi_\Theta(x_{abnormal}) > \phi_\Theta(x_{normal})$$

The learning phase is guided in a way that the scores of anomalies statistically significantly deviate from μ_R while at the same time having the scores of normal objects as close as possible to μ_R.

- As in the previous method, sets of candidate inliers and outliers are needed (distance-based approaches or labeled data)
- A reference score generator (learned or defined by a prior probability) is used to generate another scalar score termed as reference score, which is defined as the mean of the anomaly scores $r_1, r_2, \dots, r_l$ for a set of l randomly selected normal objects, denoted as μ_R.



Deviation Network

- The deviation to the reference score of a sample x : $\text{dev}(\mathbf{x}) = \frac{\phi(\mathbf{x};\Theta) - \mu_{\mathcal{R}}}{\sigma_{\mathcal{R}}}$

$$\mathcal{L} = (1-y)|\,\text{dev}(\mathbf{x})| \\ + y \max(0, a - \text{dev}(\mathbf{x}))$$

  with $y = 1$ for candidate outliers and $y = 0$ for inliers

- The loss forces the normal objects cluster around F in terms of their anomaly scores but pushes anomalies statistically far away F, thus the intermediate representation learning to discriminate normal objects from anomalies.

Algorithm 1 *Training DevNet*

Input: $\mathcal{X} \in \mathbb{R}^D$ - training data objects, i.e., $\mathcal{X} = \mathcal{U} \cup \mathcal{K}$ and $\emptyset = \mathcal{U} \cap \mathcal{K}$
Output: $\phi : \mathcal{X} \mapsto \mathbb{R}$ - an anomaly scoring network

```
Randomly initialize Θ
for i = 1 to n_epochs do
    for j = 1 to n_batches do
        B ← Randomly sample b data objects with a half of objects from
            K and another half from U
        Randomly sample l anomaly scores from N(μ, σ²)
        Compute μ_R and σ_R of the l anomaly scores: {r_1, r_2, ··· , r_l}
        loss ← (1/b) Σ_{x∈B} L(φ(x;Θ), μ_R, σ_R)
        Perform a gradient descent step w.r.t. the parameters in Θ
    end for
end for
return φ
```

Application

The End