Deep Anomaly Detection

Joel Mbouwe

DataLab GBIS/CDO

19 Juillet 2020

Plan

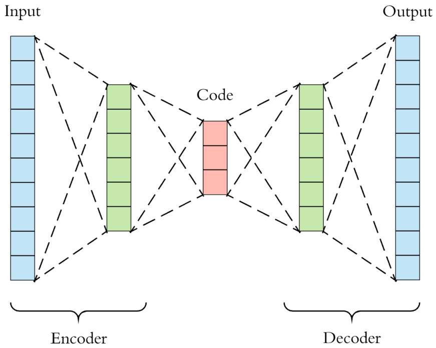
1. Introduction
2. Deep Learning techniques for anomaly detection
   - AutoEncoder
   - Deep Support Vector Data Descriptor
   - REPEN
   - Deviation Network
3. Application

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 consists precisely in making a state of the art of deep learning techniques for anomaly detection.

AutoEncoder for Anomaly Detection
Approach

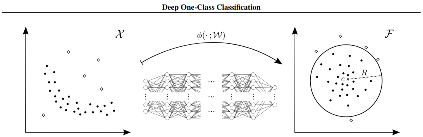
- Model for learning a low dimensional representation of the data
- Encoder for dimension reduction and the decoder for the reconstruction of the data
- The learning process is done by minimizing the reconstruction error : $\|\hat{X} - X\|_2$
- We except high reconstruction error for abnormal data points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



Deep SVDD

It is a deep learning method for anomaly detection where the goal is to learn a representation by mapping the data as close as possible to a defined center.

- The model is forced to extract the common features 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



2 configurations :

One-class

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

Soft boundary

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

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 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 method. Given a sample x_i, the approach defines its outlierness as follows : $r_i = \frac{1}{m}\sum_1^m nn\_dist(x_i|S_j)$ where $S_j \subset X$ 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 representation of data 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.

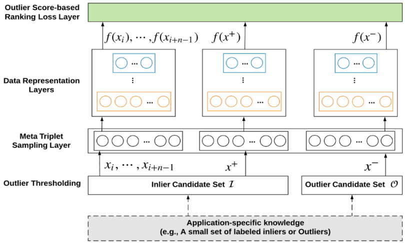


Figure 1: The Proposed RAMODO Framework. RAMODO learns a representation function $f(\cdot)$ to map D-dimensional input objects into a M-dimensional space, with $M \ll D$.

REPEN

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

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

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

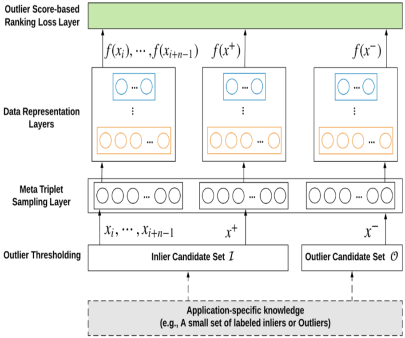


Figure 1: The Proposed RAMODO Framework. RAMODO learns a representation function $f(\cdot)$ to map D-dimensional input objects into a M-dimensional space, with $M \ll D$.

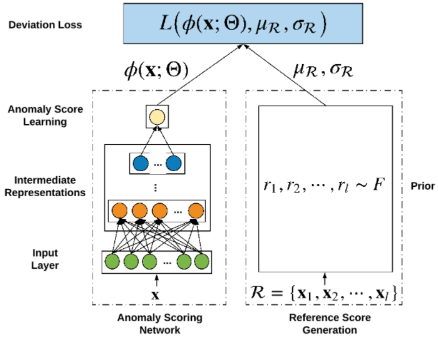
Deviation Network

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 a reference score 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 $r_1, r_2, \ldots, r_l$ 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 learns 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**
    $\mathcal{B} \leftarrow$ Randomly sample b data objects with a half of objects from $\mathcal{K}$ and another half from $\mathcal{U}$
    Randomly sample l anomaly scores from $N(\mu, \sigma^2)$
    Compute $\mu_{\mathcal{R}}$ and $\sigma_{\mathcal{R}}$ of the l anomaly scores: $\{r_1, r_2, \cdots, r_l\}$
    $loss \leftarrow \frac{1}{b}\sum_{\mathbf{x}\in\mathcal{B}} L\big(\phi(\mathbf{x};\Theta), \mu_{\mathcal{R}}, \sigma_{\mathcal{R}}\big)$
    Perform a gradient descent step w.r.t. the parameters in Θ
  end for
end for
return ϕ

Application

Thank you