Deep Anomaly Detection

Joel Mbouwe

DataScience GBIS/CDO

17 Juillet 2020

Plan

Introduction

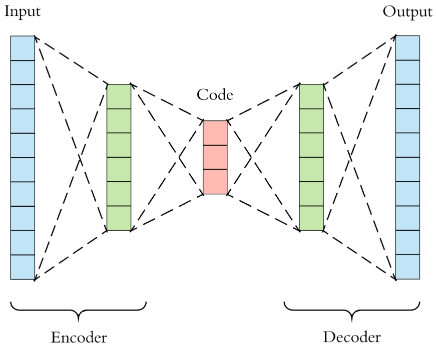
- An anomaly is « *an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism.* » Hawkings.

- An anomaly detection model is a model that learns how to characterize the normality of the data and estimates 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

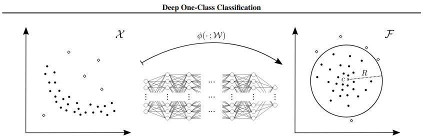
- Model for learning a low dimensional representation of the data
- Encoder for dimension reduction and the decoder for the reconstruction of the data
- The learning process is done by minimizing the reconstruction error : $\|\hat{X} - X\|_2$
- We except high reconstruction error for abnormal data points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



Deep SVDD

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

- The model is forced to extract the common features that enables the proximity to the center..
- The anomaly score is defined as the distance to the center $\|\phi(\boldsymbol{x}_i, \mathcal{W}) - \boldsymbol{c}\|^2$
- Point of attention : No bias, upper bounded and minored by something other than zero 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

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

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. The outlierness of a sample x_i is : $r_i = \frac{1}{m} \sum_1^m nn\_dist(x_i|S_j)$ where $S_j \subset X$ is a random data subsample of fixed size, m the number of subsample.

- Faster approach which is more robust to group anomalies

REPEN

Framework to learn low-dimensional representation of data such that given a distance-based outliers function ϕ (KNN, Lesinn etc.) the learned mapping function f satisfies $\phi(f(x_{abnormal})) > \phi(f(x_{normal}))$

- 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.

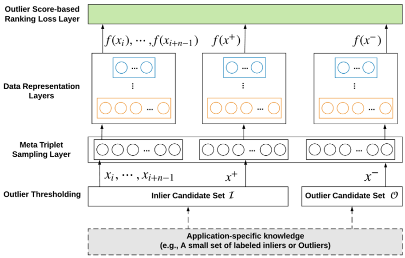


Figure 1: The Proposed RAMODO Framework. RAMODO learns a representation function $f(\cdot)$ to map D-dimensional input objects into a M-dimensional space, with $M \ll D$.

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^+

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

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

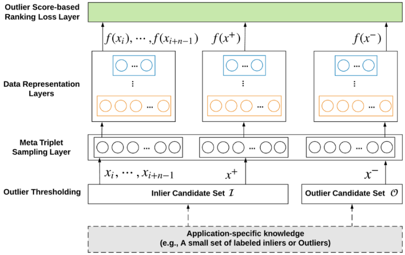


Figure 1: The Proposed RAMODO Framework. RAMODO learns a representation function $f(\cdot)$ to map D-dimensional input objects into a M-dimensional space, with $M \ll D$.

Deviation Network

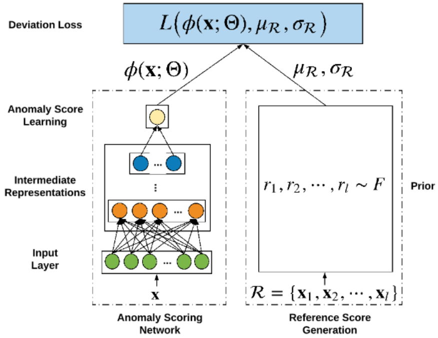
A semi-supervised model that directly learns an anomaly score function ϕ_Θ such that

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

The learning phase is guided in a way that the scores of anomalies 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.

- A very small set of labeled anomalies that provide some prior knowledge of anomalies (We tried using a distance-based method to obtain pseud-outliers)

- A reference score generator (learned or defined by a prior probability) is used to generate a reference score μ_R defined as the mean of the anomaly scores $r_1, r_2, \ldots, r_l$ for a set of l randomly selected normal objects.



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 the reference score in terms of their anomaly scores but pushes anomalies far away from μ_R, 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 $\mathcal{N}(\mu, \sigma^2)$
    Compute $\mu_{\mathcal{R}}$ and $\sigma_{\mathcal{R}}$ of the l anomaly scores: $\{r_1, r_2, \cdots, r_l\}$
    $loss \leftarrow \frac{1}{b}\sum_{\mathbf{x}\in\mathcal{B}} L(\phi(\mathbf{x};\Theta), \mu_{\mathcal{R}}, \sigma_{\mathcal{R}})$
    Perform a gradient descent step w.r.t. the parameters in Θ
  end for
end for
return ϕ

Application

- Synthetic data composed of a mixture of two gaussian distributions with anomalies between the axis of the Gaussians
- Tests are mostly in an unsupervised configuration and we will dig into the normality model learned by the approaches
- Both cases when the training data is contaminated (includes real anomalies) or not will be investigated
- The performance measure will be the Area Under The Curve Precision-Recall

Thank you