Beyond Individual Input for Deep Anomaly Detection on Tabular Data

Hugo Thimonier, Fabrice Popineau, Arpad Rimmel and Bich-Liên Doân

Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire Interdisciplinaire des Sciences du Numérique, 91190, Gif-sur-Yvette, France.

ICML International Conference On Machine Learning

Paper GitHub

Context

- **Semi-supervised anomaly detection** (AD) is a **good alternative** to standard supervised models when there is **extreme imbalancing** between classes.
- General AD methods offer **good performance on unstructured data**.
- Current best performing AD methods for tabular data **take into account its particular structure.**
- Recent works on **deep learning for tabular data** have highlighted that leveraging **both inter-feature and inter-sample relations** may foster performance.

Method

Mask-reconstruction

Train a model ϕ_θ to reconstruct the **masked features of normal samples**.

- Sample vector $\mathbf{x} \in \mathbb{R}^d$, binary mask vector $\mathbf{m} \in \mathbb{R}^d$.
- $\mathbf{x}^m, \mathbf{x}^o \in \mathbb{R}^d$ represent respectively the masked and unmasked entries of sample $\mathbf{x}$

$$\mathbf{x}^m = \mathbf{m} \odot \mathbf{x}$$
$$\mathbf{x}^o = (\mathbf{1}_d - \mathbf{m}) \odot \mathbf{x}$$

- The **training objective** consists in minimizing

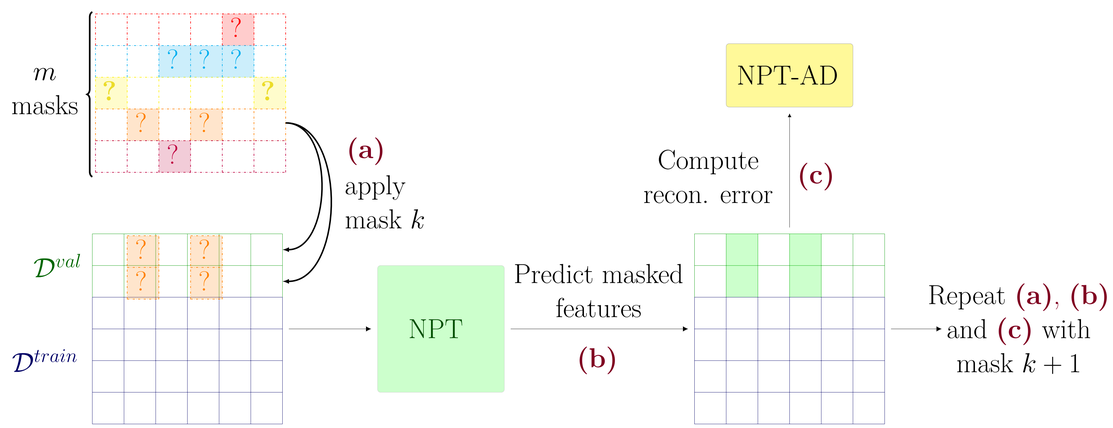
$$\min_{\theta \in \Theta} \sum_{\mathbf{x} \in \mathcal{D}_{train}} d(\mathbf{x}^m, \phi_\theta(\mathbf{x}^o)),$$

where $\phi_\theta(\mathbf{x}^o) \in \mathbb{R}^d$ denotes the reconstructed masked features of $\mathbf{x}$ by the model, and $d(.,.)$ a distance measure.

Non-Parametric Transformer

- We rely on **Non-Parametric Transformers** (NPT) as our **core model** ϕ_θ.
- NPT enables **leveraging both inter-feature and inter-sample relations**.

NPT-AD



NPT-AD Inference Pipeline

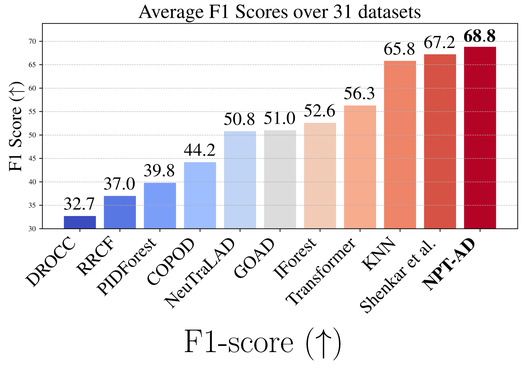
(a) Mask j is applied to each validation sample. We construct a matrix $\mathbf{X}$ **composed of the masked validation samples and the whole *unmasked* training set.**

(b) We feed $\mathbf{X}$ to the Non-Parametric Transformer (NPT), which tries to **reconstruct the masked features** for each validation sample

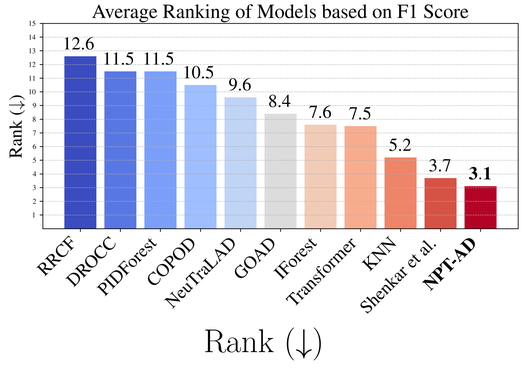
(c) We compute the **reconstruction error** that we later aggregate in the NPT-AD score

Experiment

- We evaluate our method on a **benchmark of 31 tabular datasets**.
- We compare to both **deep and non-deep AD methods** and observe that we obtain **SOTA performance**

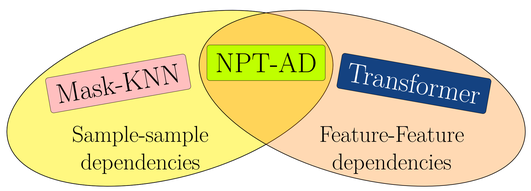


F1-score (↑)



Rank (↓)

Is combining dependencies useful ?



- Mask-KNN: mask reconstruction **only** using **sample-sample dependencies**.
- Transformer: mask reconstruction **only** using **feature-feature dependencies**.

	Transformer	Mask-KNN	NPT-AD
F1	57.4	57.5	**68.8**
AUROC	83.0	84.5	**89.8**

Combining dependencies **boosts AD performances**!

Robustness to Data Contamination

- What happens when the training set contains anomalies?
- NPT-AD's performance deteriorate **starting from 5% contamination share**.