Learning to Rank Anomalies: Scalar Performance Criteria and Maximization of Two-Sample Rank Statistics

Myrto Limnios, Nathan Noiry, Stéphan Clémençon

Learning to Rank Anomalies: Scalar Performance Criteria and Maximization of Two-Sample Rank Statistics

Myrto Limnios MYRTO.LIMNIOS@ENS-PARIS-SACLAY.FR

Université Paris-Saclay, ENS Paris-Saclay
CNRS UMR 9010, Centre Borelli, 91190 Gif-sur-Yvette, France

Nathan Noiry NATHAN.NOIRY@TELECOM-PARIS.FR

Stephan Clémençon STEPHAN.CLEMENCON@TELECOM-PARIS.FR

Telecom Paris, LTCI, Institut Polytechnique de Paris
19 place Marguerite Perey, Palaiseau, 91120, France

Editors: Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michał Woźniak and Shuo Wang.

Abstract

The ability to collect and store ever more massive databases has been accompanied by the need to process them efficiently. In many cases, most observations have the same behavior, while a probable small proportion of these observations are abnormal. Detecting the latter, defined as outliers, is one of the major challenges for machine learning applications (*e.g.* in fraud detection or in predictive maintenance). In this paper, we propose a methodology addressing the problem of outlier detection, by learning a data-driven scoring function defined on the feature space which reflects the degree of abnormality of the observations. This scoring function is learnt through a well-designed binary classification problem whose empirical criterion takes the form of a two-sample linear rank statistics on which theoretical results are available. We illustrate our methodology with preliminary encouraging numerical experiments.

Keywords: Anomaly ranking, novelty detection, two-sample linear rank statistics.

1. Introduction

The problem of ranking multivariate data by degree of abnormality, referred to as *anomaly ranking*, is of central importance for a wide variety of applications (*e.g.* fraud detection, fleet monitoring, predictive maintenance). In the standard setup, the 'normal' behavior of the system under study (in the sense of 'not abnormal', without any link to the Gaussian distribution) is described by the (unknown) distribution $F(dx)$ of a generic *r.v.* X, valued in $\mathbb{R}^d$. The goal pursued is to build a scoring function $s : \mathbb{R}^d \to \mathbb{R}_+ \cup \{+\infty\}$ that ranks any observations $x_1, \ldots, x_n$ nearly in the same order as any increasing transform of the density f would do. Ideally, the smaller the score $s(x)$ of an observation x in $\mathbb{R}^d$, the more abnormal it should be considered. In Clémençon and Thomas (2018), a functional criterion, namely a Probability-Measure plot referred to as the *Mass-Volume* curve (the MV curve in abbreviated form), has been proposed to evaluate the anomaly ranking performance of any scoring rule $s(x)$. This performance measure can be viewed as the unsupervised version of the *Receiver Operating Characteristic* (ROC) curve, the gold standard measure to evaluate the accuracy of scoring functions in the bipartite ranking context, see *e.g.*

Clémençon and Vayatis (2009). Beyond this approach, let us highlight that the problem of anomaly detection has also been studied *via* various other modelings. For instance, the works of Bergman and Hoshen (2020) and Steinwart et al. (2005) are based on classification methods, while Liu et al. (2008) build on peeling, Breunig et al. (2000) on local averaging criteria, Frery et al. (2017) on ranking and Schölkopf et al. (2001) on plug-in techniques.
In this paper, we propose a novel two-stage method for detecting and ranking abnormal instances, by means of scalar criteria summarizing the MV curve and extending the area under its curve, when $F(dx)$ has compact support. Briefly, starting from a sample of observations $X_1, \ldots, X_n$, we artificially generate an independent second sample $U_1, \ldots, U_m$ that is used as a proxy for outliers. For theoretical reasons explained in the paper, the agnostic choice consists in sampling the U_i's *i.i.d.* from the uniform law on a subset of $\mathbb{R}^d$, which $F(dx)$'s support is supposedly included in. We then learn to discriminate the X_i's from the U_i's thanks to a scoring function that maximizes two-sample empirical counterparts of the aforementioned criteria, that are in particular robust to imbalanced datasets. The resulting scoring function allows to rank the X_i's by degree of abnormality. This novel class of criteria is based on theoretical guarantees provided by Clémençon et al. (2021) on general classes of two-sample linear rank processes, that incidentally circumvent the difficulty of optimizing the functional MV criterion. Beyond the classical results of statistical learning theory for these processes, Clémençon et al. (2021) obtain theoretical generalization guarantees for their empirical optimizers. The numerical results performed at the end of the paper also provide strong empirical evidence of the relevance of the approach promoted here.

The article is structured as follows. In section 2, the formulation of the (unsupervised) anomaly ranking problem is recalled at length, together with the concept of MV curve. In section 3, the anomaly ranking performance criteria proposed are introduced and their statistical estimation is discussed. Optimization of the statistical counterparts of the criteria introduced to build accurate anomaly scoring functions is also put forward therein. Finally, the relevance of this approach is illustrated by numerical results in section 4.

2. Background and Preliminaries

We start off with recalling the formulation of the (unsupervised) anomaly ranking problem and introducing notations that shall be used here and throughout. By λ is meant the Lebesgue measure on $\mathbb{R}^d$, by $\mathbb{I}\{\mathcal{E}\}$ the indicator function of any event $\mathcal{E}$, while the generalized inverse of any cumulative distribution function $K(t)$ on $\mathbb{R}$ is denoted by $K^{-1}(u) = \inf\{t \in \mathbb{R} : \ K(t) \geq u\}$. We consider a *r.v.* X valued in $\mathbb{R}^d$, $d \geq 1$, with distribution $F(dx) = f(x)\lambda(dx)$, modeling the 'normal' behavior of the system under study. The observations at disposal $X_1, \ldots, X_n$, with $n \geq 1$, are independent copies of X. Based on the X_i's our goal is to learn a ranking rule for deciding among two observations x and x' in $\mathbb{R}^d$ which one is more 'abnormal'. The simplest way of defining a preorder[1] on $\mathbb{R}^d$ consists in transporting the natural order on $\mathbb{R}_+ \cup \{+\infty\}$ onto it through a *scoring function*, *i.e.* a Borel measurable mapping $s : \mathbb{R}^d \to \mathbb{R}_+$: given two observations x and x' in $\mathbb{R}^d$, x is said to be more abnormal

1. A preorder $\preccurlyeq$ on a set $\mathcal{Z}$ is a reflexive and transitive binary relation on $\mathcal{Z}$. It is said to be *total*, when either $z \preccurlyeq z'$ or else $z' \preccurlyeq z$ holds true, for all $(z, z') \in \mathcal{Z}^2$.

according to s than x' when $s(x) \leq s(x')$. The set of all anomaly scoring functions that are integrable with respect to Lebesgue measure is denoted by $\mathcal{S}$. The integrability condition is not restrictive since the preorder induced by any scoring function is invariant under strictly increasing transformation (*i.e.* the scoring function s and its transform $T \circ s$ define the same preorder on $\mathbb{R}^d$ provided that the Borel measurable transform $T : \mathrm{Im}(s) \to \mathbb{R}_+$ is strictly increasing on the image of the *r.v.* $s(X)$, denoted by $\mathrm{Im}(s)$). One wishes to build, from the 'normal' observations only, a scoring function s such that, ideally, the smaller $s(X)$, the more abnormal the observation X. The set of optimal scoring rules in $\mathcal{S}$ should be thus composed of strictly increasing transforms of the density function $f(x)$ that are integrable *w.r.t.* to λ, namely:

$$\mathcal{S}^* = \{T \circ f : \ T : \mathrm{Im}(f) \to \mathbb{R}_+ \text{ strictly increasing}, \int_{\mathbb{R}^d} T \circ f(x)\lambda(dx) < +\infty\} \ . \quad (1)$$

The technical assumptions listed below are required to define a criterion, whose optimal elements coincide with $\mathcal{S}^*$.

H$_1$ The *r.v.* $f(X)$ is continuous, *i.e.* $\forall c \in \mathbb{R}_+$, $\mathbb{P}\{f(X) = c\} = 0$.

H$_2$ The density function $f(x)$ is bounded: $||f||_\infty \stackrel{def}{=} \sup_{x \in \mathbb{R}^d} |f(x)| < +\infty$.

Measuring anomaly scoring accuracy - The MV curve. Consider an arbitrary scoring function $s \in \mathcal{S}$ and denoted by $\Omega_{s,t} = \{x \in \mathcal{X} : \ s(x) \geq t\}$, $t \geq 0$, its level sets. As s is λ-integrable, the measure $\lambda(\Omega_{s,t}) \leq (\int_{u \in \mathbb{R}_+} s(u)du)/t$ is finite for any $t > 0$. Introduced in Clémençon and Thomas (2018), a natural measure of the anomaly ranking performance of any scoring function candidate s is the Probability-Measure plot, referred to as the *Mass-Volume* (MV) curve:

$$t > 0 \mapsto \Big(\mathbb{P}\{s(X) \geq t\}, \ \lambda(\{x \in \mathbb{R}^d : \ s(x) \geq t\})\Big) = (F(\Omega_{s,t}), \ \lambda(\Omega_{s,t})) \ . \quad (2)$$

Connecting points corresponding to possible jumps, this parametric curve can be viewed as the plot of the continuous mapping $\mathrm{MV_s} : \alpha \in (0,1) \mapsto \mathrm{MV_s}(\alpha)$, starting at $(0,0)$ and reaching $(1, \ \lambda\big(\mathrm{supp}(F)\big)$ in the case where the support $\mathrm{supp}(F)$ of the distribution $F(dx)$ is compact, or having the vertical line '$\alpha = 1$' as an asymptote otherwise. A typical MV curve is depicted in Fig. 1.
Let $\alpha \in (0,1)$. Denoting by $F_s(t)$ the cumulative distribution function of the *r.v.* $s(X)$, we have:

$$\mathrm{MV}_s(\alpha) = \lambda\left(\{x \in \mathbb{R}^d : \ s(x) \geq F_s^{-1}(1-\alpha)\}\right), \quad (3)$$

when $F_s \circ F_s^{-1}(\alpha) = \alpha$. This functional criterion is invariant by increasing transform and induces a partial order over the set $\mathcal{S}$. Let $(s_1, s_2) \in \mathcal{S}^2$, the ordering defined by s_1 is said to be more accurate than the one induced by s_2 when:

$$\forall \alpha \in (0,1), \ \ \mathrm{MV_{s_1}}(\alpha) \leq \mathrm{MV_{s_2}}(\alpha) \ .$$

As summarized by the result stated below, the MV curve criterion is adequate to measure the accuracy of scoring functions with respect to anomaly ranking.
It reveals in particular that optimal scoring functions are those whose MV curve is minimum everywhere.

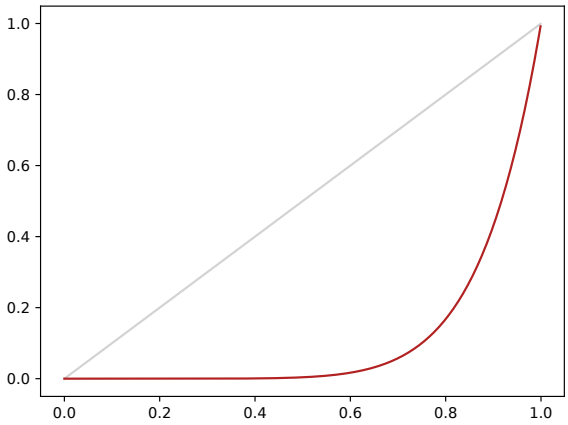


Figure 1: Typical MV curve in red (x-axis:volume, y-axis:mass). In gray, the diagonal $y = x$.

Proposition 1 *(Clémençon and Thomas (2018)) Let the assumptions* $\mathbf{H_1}-\mathbf{H_2}$ *be fulfilled. The elements of the class* $\mathcal{S}^*$ *have the same (convex)* MV *curve and provide the best possible preorder on* $\mathbb{R}^d$ *w.r.t. the* MV *curve criterion:*

$$\forall (s,\alpha) \in \mathcal{S} \times (0,1), \quad \mathrm{MV}^*(\alpha) \leq \mathrm{MV}_s(\alpha) \ , \tag{4}$$

where $\mathrm{MV}^*(\alpha) = \mathrm{MV}_f(\alpha)$ *for all* $\alpha \in (0,1)$.

Equation (4) reveals that the lowest the MV curve (everywhere) of a scoring function $s(x)$, the closer the preorder defined by $s(x)$ is to that induced by $f(x)$. Favorable situations are those where the MV curve increases slowly and rises more rapidly when coming closer to the 'one' value: this correponds to the case where $F(dx)$ is much concentrated around its modes, $s(X)$ takes its highest values near the latter and its lowest values are located in the tail region of the distribution $F(dx)$. Incidentally, observe that the optimal curve MV^* somehow measures the spread of the distribution $F(dx)$ in particular for large values of α *w.r.t.* extremal observations (*e.g.* a light tail behavior corresponds to the situation where $\mathrm{MV}^*(\alpha)$ increases rapidly when approaching 1), whereas it should be examined for small values of α when modes of the underlying distributions are investigated (a flat curve near 0 indicates a high degree of concentration of $F(dx)$ near its modes).

Statistical estimation. In practice, the MV curve of a scoring function $s \in \mathcal{S}$ is generally unknown, just like the distribution $F(dx)$, and it must be estimated. A natural empirical counterpart can be obtained by plotting the stepwise graph of the mapping:

$$\widehat{\mathrm{MV}}_s(\alpha) : \alpha \in (0,1) \mapsto \lambda\left(\left\{x \in \mathbb{R}^d : \ s(x) \geq \widehat{F}_{s,n}^{-1}(1-\alpha)\right\}\right) \ , \tag{5}$$

where $\widehat{F}_{s,n}(t) = (1/n)\sum_{i=1}^{n} \mathbb{I}\{s(X_i) \leq t\}$ denotes the empirical *c.d.f.* of the *r.v.* $s(X)$ and $\widehat{F}_{s,n}^{-1}$ its generalized inverse. In Clémençon and Thomas (2018), for a fixed $s \in \mathcal{S}$, consistency and asymptotic Gaussianity (in sup-norm) of the estimator (5) has been established, together with the asymptotic validity of a smoothed bootstrap procedure to build confidence regions in the MV space. However, depending on the geometry of the superlevel sets of $s(x)$, it can be far from simple to compute the volumes. In the case where F has compact support, included in $[0,1]^d$ say for simplicity, and from now on it is assumed it is the case, they can be estimated by means of Monte-Carlo simulation. Indeed, if one generates a synthetic *i.i.d.* sample $\{U_1, \ldots, U_m\}$, independent from the X_i's and drawn from the uniform distribution on $[0,1]^d$, which we denote by $\mathcal{U}_d$, a natural estimator of the volume $\widehat{\mathrm{MV}}_s(\alpha)$ is:

$$\widetilde{\mathrm{MV}}_s(\alpha) = \frac{1}{m}\sum_{j=1}^{m} \mathbb{I}\{s(U_j) \geq \widehat{F}_{s,n}^{-1}(1-\alpha)\} \ . \tag{6}$$

Minimization of the empirical area under the MV **curve.** Thanks to the MV curve criterion, it is possible to develop a statistical theory for the anomaly scoring problem. From a statistical learning angle, the goal is to build from training data $X_1, \ldots, X_n$ a scoring function with MV curve as close as possible to MV^*. Whereas the closeness between (continuous) curves can be measured in many ways, the L_1-distance offers crucial advantages. Indeed, we have:

$$d_1(s,f) = \int_{\alpha=0}^{1} |\mathrm{MV_s}(\alpha) - \mathrm{MV}^*(\alpha)|\,\mathrm{d}\alpha = \int_{\alpha=0}^{1} \mathrm{MV_s}(\alpha)\mathrm{d}\alpha - \int_{\alpha=0}^{1} \mathrm{MV}^*(\alpha)\mathrm{d}\alpha \ ,$$

Notice that $d_1(s,f)$, $i \in \{1, \infty\}$, is not a distance between the scoring functions s and f but measures the dissimilarity between the preorders they define and that minimizing $d_1(s,f)$ boils down to minimizing the scalar quantity $\int_{\alpha=0}^{1-\varepsilon} \mathrm{MV_s}(\alpha)\mathrm{d}\alpha$, the area under the MV curve. From a practical perspective, one may then learn an anomaly scoring rule by minimizing the empirical quantity:

$$\int_0^1 \widetilde{\mathrm{MV}}_s(\alpha)d\alpha \ .$$

This boils down to maximizing the rank-sum (or Wilcoxon Mann-Whithney) statistic (see Wilcoxon (1945)) given by:

$$\widehat{W}_{n,m}(s) = \sum_{i=1}^{n} \mathrm{Rank}(s(X_i)) \ , \tag{7}$$

where $\mathrm{Rank}(s(X_i))$ is the rank of $s(X_i)$ among the pooled sample $\{s(X_1), \ldots, s(X_n)\} \cup \{s(U_1), \ldots, s(U_m)\}$: $\mathrm{Rank}(s(X_i)) = \sum_{l=1}^{n} \mathbb{I}\{s(X_l) \leq s(X_i)\} + \sum_{j=1}^{m} \mathbb{I}\{s(U_j) \leq s(X_i)\}$. Indeed, just like the empirical area under the ROC curve can be related to the rank-sum statistic, we have:

$$nm\left(1 - \int_0^1 \widetilde{\mathrm{MV}}_s(\alpha)d\alpha\right) + n(n+1)/2 = \widehat{W}_{n,m}(s) \ . \tag{8}$$

In the next section, we introduce more general empirical summaries of the MV curve that are of the form of two-sample rank statistics, just like (7), and propose to solve the anomaly ranking problem through the maximization of the latter.

3. Measuring and Optimizing Anomaly Ranking Performance

In this section, a class of anomaly ranking performance criteria are introduced, which can be estimated by two-sample rank statistics. We also emphasize that a natural approach to anomaly ranking consists in maximizing such empirical scalar criteria.

3.1. Scalar Criteria of Performance and Two-sample Rank Statistics

Here we develop the statistical learning framework we propose for anomaly ranking. Let $p \in (0,1)$, we assume that $N \geq 2$ observations are available: $n = \lfloor pN \rfloor$ 'normal' *i.i.d.* observations $X_1, \ldots, X_n$ taking their values in $[0,1]^d$ for simplicity drawn from $F(dx) = f(x)\lambda(dx)$ and $m = N - n$ *i.i.d.* realizations of the uniform distribution $\mathcal{U}_d$, independent from the X_i's. Hence, p represents the 'theoretical' proportion of 'normal' observations among the pooled sample. Let a class of scoring functions $\mathcal{S}_0 \subset \mathcal{S}$ such that, for all $s(x)$, we consider the mixture distribution $G_s = pF_s + (1-p)\lambda_s$ and its empirical counterpart $\widehat{G}_{s,N}(t) = (1/n)\sum_{i=1}^{n} \mathbb{I}\{s(X_i) \leq t\} + (1/m)\sum_{j=1}^{m} \mathbb{I}\{s(U_i) \leq t\}$. Notice that since $n/N \to p$ as N tends to infinity, the quantity above is a natural estimator of the *c.d.f.* G_s. We refer to the *scored* random samples for $\{s(X_1), \ldots, s(X_n)\}$ and $\{s(U_1), \ldots, s(U_m)\}$. Therefore, motivated by Eq. (8), Definition 2 below provides the class of *W_ϕ-performance criteria* we consider in the subsequent procedure.

Definition 2 *Let $\phi : [0,1] \to \mathbb{R}$ be a nondecreasing function. The 'W_ϕ-ranking performance criterion' with 'score-generating function' $\phi(u)$ based on the mixture cdf $G_s(dt)$ is given by:*

$$W_\phi(s) = \mathbb{E}[(\phi \circ G_s)(s(X))] \ . \tag{9}$$

One can naturally relate this generalized form to the MV curve, justifying this choice of scalar performance criteria as summaries of the MV curve, through the equality:

$$W_\phi(s) = \int_0^1 \phi\left(1 - p\alpha - (1-p)\mathrm{MV}_s(\alpha)\right) d\alpha \ . \tag{10}$$

Equipped with the two random samples, the following Definition 3 provides an empirical counterpart, that generalizes the empirical summaries of the MV curve *via* collections of two-sample linear rank statistics. Precisely, for a given mapping $s(x)$, we allow to weight the sequence of 'normal ranks' *i.e.* the ranks of the scored 'normal' instances among the pooled sample, by means of a *score-generating* function.

Definition 3 (Two-sample linear rank statistics) *Let $\phi : [0,1] \to \mathbb{R}$ be a nondecreasing function. The two-sample linear rank statistics with 'score-generating function' $\phi(u)$ based on the random samples $\{X_1, \ldots, X_n\}$ and $\{U_1, \ldots, U_m\}$ is given by:*

$$\widehat{W}_{n,m}^{\phi}(s) = \sum_{i=1}^{n} \phi\left(\frac{\mathrm{Rank}(s(X_i))}{N+1}\right) \ , \tag{11}$$

where $\mathrm{Rank}(t) = N\widehat{G}_{s,N}(t) = \sum_{i=1}^{n} \mathbb{I}\{s(X_i) \leq t\} + \sum_{j=1}^{m} \mathbb{I}\{s(U_j) \leq t\}$.

Optimality. Briefly, we refer to the comprehensive analysis of the general class of criteria in Clémençon et al. (2021), that establishes the theoretical guarantees for the consistency of the two-stage procedure we detail in the following subsection. Importantly, the set of optimal maximizers of the empirical W_ϕ-criteria coincides with the nondecreasing transforms of the likelihood ratio, just like for the MV curves, as shown thourgh the Eq. (10).
The optimal set $\mathcal{S}^*$ derived in Eq. (1) underlines the implicit characterization that inherits an outlier: the lower the scalar score is and the likelier anomalous the observation can be considered. Also, the notion of distance induced by the rank-based criteria is in fact directly related to the distribution of the 'normal' sample compared to the Uniform one.

Choosing ϕ. As foreshadowed above, the choice of the score-generating function is an asset of this class of criteria as it provides a flexibility *w.r.t.* the weighting of the area under the MV curve. Indeed, its minimization directly implies the maximization of the W_ϕ-criterion (see Eq. (10)), recalling the nondecreasing variation of $\phi(u)$. Therefore, one can hope to recover at best the MV* curve by the right choice of $\phi(u)$, especially when the initial sample is noisy. Additionally, when going back to the problem of learning to rank the (possible abnormal) instances, it is an advantage to weight the ranks accordingly.
First, we recall the simplest uniform weighting of each 'normal' rank with $\phi(u) = u$. It parenthetically yields to Eq. (8), of continuous version: $W(s) = p/2 + (1-p)(1 - \int_0^1 \mathrm{MV}_s(\alpha)\mathrm{d}\alpha)$, where the area under the MV curve is clearly computed. Other functions were introduced in the literature related to classic univariate two-sample rank statistics. Figure 2 gathers classical nondecreasing score-generating functions broadly used for two-sample statistical tests (refer to Hájek (1962)).

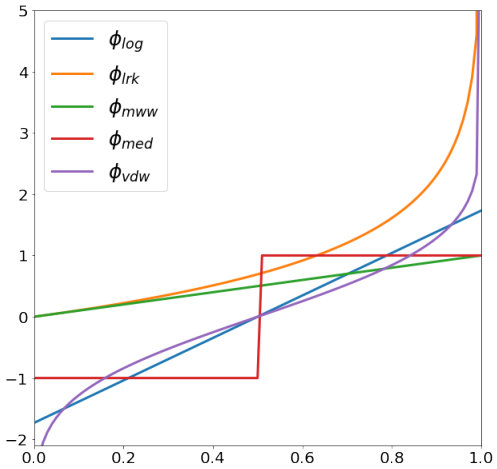


Figure 2: Curves of two-sample score-generating functions with the associated statistical test: Logistic test $\phi_{log}(u) = 2\sqrt{3}(u-1/2)$ in blue, Logrank test $\phi_{lrk}(u) = -\log(1-x)$ in orange, Mann-Whitney-Wilcoxon test $\phi_{mww}(u) = u$ in green, Median test $\phi_{med}(u) = \mathrm{sgn}(u-1/2)$ in red, Van der Waerden test $\phi_{vdw}(u) = \Phi^{-1}(u)$ in purple, Φ being the normal quantile function.

3.2. The Two-Stage Procedure

In this paragraph, we detail the two-stage procedure, where we assume that both the framework and assumptions detailed in the previous subsection are adopted. We define the test sample as the set of *i.i.d.* random variables $\{X_1^t, \ldots, X_{n_t}^t\}$, with $n_t \in \mathbb{N}^*$, *a priori* drawn from $F(dx)$. The goal pursued is to distinguish among the test sample, the instances the most likelier to be anomalous. In particular, we propose a first step (1.) that outputs an optimal ranking rule $\hat{s}_{n,m}(x)$, in the sense of the maximization of the rank statistics of Eq. (3). Then, in the second step (2.) and equipped with this rule, the instances of the test sample are optimally ranked by increasing order of similarity *w.r.t.* the X's. We also choose to watch a number of $n_{lowest} \in \mathbb{N}^*$ worst ranked instances *i.e.* of lowest empirical score. The procedure is detailed in the following Fig. 3. By means of the recalled theoretical guarantees proved in Clémençon et al. (2021), it results to the asymptotic consistency of step (1.) as well as its nonasymptotic consistency with high probability, under some technical assumptions.

Anomaly Ranking Procedure

Consider the 'normal' *i.i.d.* random sample $\{X_1, \ldots, X_n\}$ of unknown probability probability function $F(dt)$ defined on a feature space included $[0,1]^d$. Consider a new random sample $\{X_1^t, \ldots, X_{n_t}^t\}$ *a priori* drawn from $F(dt)$, $n_t \in \mathbb{N}^*$, independent of the X's. Let $\mathcal{S}_0 \subset \mathcal{S}$ a class of scoring functions and ϕ a score-generating function. Set $n_{lowest} \in \mathbb{N}^*$.

1. Maximizing the W_ϕ-ranking performance criterion.

(i) Generate a *i.i.d.* random sample $\{U_1, \ldots, U_m\}$ from $\mathcal{U}_d(dt)$, independent of $\{X_1, \ldots, X_n\}$,

(ii) Output the optimal empirical scoring rule $\hat{s}_{n,m} = \arg\max_{s \in \mathcal{S}_0} \widehat{W}_{n,m}^{\phi}(s)$ based on the two samples $\{X_1, \ldots, X_n\}$ and $\{U_1, \ldots, U_m\}$.

2. Ranking anomalies.

(i) Compute the empirical scores of the test sample $\{\hat{s}_{n,m}(X_1^t), \ldots, \hat{s}_{n,m}(X_{n_t}^t)\}$,

(ii) Define as anomalous the n_{lowest} observations of lowest empirical scores among the sequence $\hat{s}_{n,m}(X_i^t)$, $i \leq n_t$.

Result. Output the set of anomalous observations with their corresponding rank.

Figure 3: Two-stage procedure for learning to rank anomalies.

4. Numerical Experiments

In this section, we illustrate the procedure promoted along the paper through numerical experiments on imbalanced synthetic data. As these experiments are mainly here to support our methodology, we propose for the step (1.) to learn the empirical maximizer $\hat{s}_{n,m}$ by

means of a regularized classification algorithm. At a technical level, we would ideally like to replace usual loss criterion such as the BCE (Binary Cross-Entropy) loss by our tailored objective W_ϕ. Unfortunately, the latter is not smooth and of highly correlated terms, which results in many challenges regarding its optimization. In order to incorporate W_ϕ and still keeping good performances, we *(i)* use a regularized proxy of it and *(ii)* incorporate the regularized criterion in a penalization term. The second point allows to drive the learning with a usual BCE loss, which asymptotically amounts to estimate the conditional probability $\mathbb{P}(y = 1 \,|\, X)$, while considering W_ϕ.

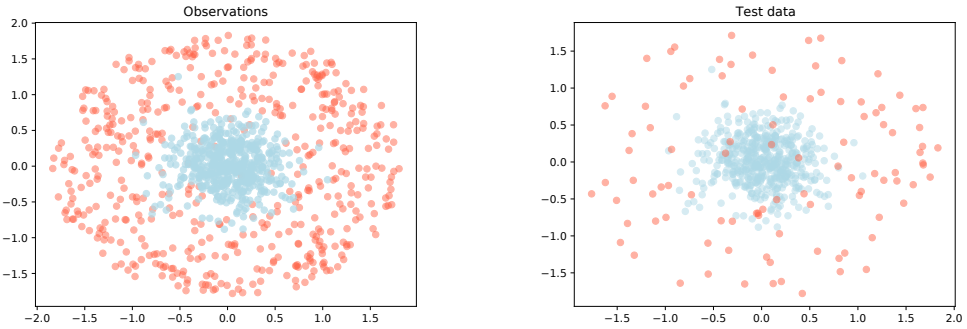
Data generating process. We generated the 'positive' sample by *i.i.d.* Gaussian variables $X_1, \ldots, X_n$, $n = 1000$, in dimension $d = 2$, centered and with covariance matrix $0.1 \times I_2$ (where I_2 is the identity matrix). We chose the Gaussian law for its attractive structure and in particular for its symmetry, it can be a reasonable choice in many situations where the data at hand are indeed well structured. We then sampled the 'negative' sequence of *i.i.d.* *r.v.* $U'_1, \ldots, U'_m$, $m = 500$, from the following radial law, expressed in terms of its density in polar coordinates:

$$\mathrm{RadLaw}_{\alpha,\beta} : (v, r) \in \mathbb{S}^{d-1} \times (0, 1) \mapsto \frac{1}{\mathrm{Area}(\mathbb{S}^{d-1})} dv \times \frac{1}{B(\alpha, \beta)} r^{\alpha-1}(1 - r)^{\beta-1} dr \ ,$$

where $\alpha, \beta > 0$ are two tunable parameters, $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d, \ \|x\| = 1\}$ is the unit sphere, and where $B(\alpha, \beta) = \int_0^1 r^{\alpha-1}(1-r)^{\beta-1} dr$. In other words, v is uniformly sampled in the unit sphere and r has Beta law with parameters α and β. Notice that $\alpha = \beta = 1$ corresponds to the Uniform law and that, when $\beta = 1$, the law puts more mass around 1 as $\alpha > 1$ increases. In our experiment, we choose $\alpha = 3$ and $\beta = 1$. Denoting by $\mathrm{rad} = \max_{1 \leq i \leq n} ||X_i||$, we finally obtained m 'synthetic outliers' $U_1, \ldots, U_m$ defined by $U_i = (\mathrm{rad} + \varepsilon) \times U'_i$, with $\varepsilon = 0.01$. To simplify the notations, we denote by $\mathbf{Z}_{train}$ the concatenation of the X_i's and the U_i's. We also denote by $\mathbf{y}_{train}$ the labels, where we choose to assign the label 1 (*resp.* 0) to the 'positive' (*resp.* 'negative') sample. Figure 4 illustrates both data generating processes. For the test set, we generated similarly a sequence of $n_t = 400$ *i.i.d.* Gaussian *r.v.* $X^t_1, \ldots, X^t_{n_t}$ from the same Gaussian law as the 'positive' sample, and a *i.i.d.* random sequence $U^t_1, \ldots, U^t_{m_t}$, $m_t = 100$, drawn from the law $\mathrm{RadLaw}_{\alpha_t,\beta_t}$, with $\alpha_t = 2.$ and $\beta_t = 1.$, dilated by a factor $(\mathrm{rad} + \varepsilon)$.

Metrics. Once the algorithm that learns a (renormalized) optimal scoring function $\hat{s}_{n,m} : \mathbb{R}^d \to (0, 1)$ has been trained (*i.e.* step (1.)), we score the test data with $\hat{s}_{n,m}$ and compute the proportion of true outliers among the n_{lowest} points having lowest scores (*i.e.* step (2.)). We let n_{lowest} varies in $\{25, 50, 75, 100\}$. Formally, if $\xi_1 \preccurlyeq \cdots \preccurlyeq \xi_{n_t+m_t}$ denote the points X^t_i and U^t_i sorted by scores, *i.e.* the ordered sequence based on $\hat{s}_{n,m}(\mathbf{Z}_{test,1}), \cdots, \hat{s}_{n,m}(\mathbf{Z}_{test,n_t+m_t})$, we compute the following accuracy:

$$\mathrm{Acc}_{n_{lowest}} = \frac{1}{n_{lowest}} \sum_{i=1}^{n_{lowest}} \mathbb{I}\{\xi_i \in \{U^t_1, \ldots, U^t_{m_t}\}\} \ . \tag{12}$$



(a) Train data. $(n,\ m) = (1000,\ 500)$. (b) Test data. $(n_t,\ m_t) = (400,\ 100)$.

Figure 4: Data visualization for the two generating processes. The Gaussian observations are represented in blue. The 'synthetic outliers' samples drawn from the radial law are represented in red. The left figure (a) corresponds to the train dataset, the right (b) to the test dataset.

Neural Network. We trained a neural network MLP composed of one hidden layer of size $2 \times d$, a ReLu activation function and whose last layer is a Sigmoid function, computing the desired score. For each $n_{epoch} = 30$ epochs, we use the following training scheme:

1. Each sample of $(\mathbf{Z}_{train},\ \mathbf{y}_{train})$ is individually passed through the network, the BCE loss is computed[2] and a backpropagation step is performed,

2. At the end of each epoch, the whole batch of the training dataset $(\mathbf{Z}_{train},\ \mathbf{y}_{train})$ is passed through the network and we computed the Binary Cross Entropy loss, denoted by BCE, and the following proxy of W_ϕ:

$$\widehat{W}^{\phi}_{n,m} = \sum_{i=1}^{n} \phi\left(\frac{(n+m) \times \text{MLP}(X_i) + 1}{n+m+1}\right) .$$

   In our experiments, we choose $\phi(u) = u$ and $\phi_{u_0}(u) = u\mathbb{I}\{u \geq u_0\}$ with $u_0 = 0.7$, as defined in section 3.1. We then compute the regularized loss $\text{BCE} - \lambda\widehat{W}^{\phi}_{n,m}$, where λ is a hyperparameter in $\{0, 0.01, 0.1, 1, 10\}$.

The training procedure of the Neural Net is summarized in the Algorithm 1.

Repetitions. We repeat $B = 100$ times the procedure, each time computing the accuracy metric defined above.

Visualization and results. In this section, we only display the results obtained with $\phi(u) = u$ since they are very similar to the one obtained with $\phi(u) = u\mathbb{I}\{u \geq u_0\}$. This is probably due to the very simple framework adopted for the data generating process and further investigations would be of interest.

2. Remember it is given by $-y \ln \hat{y} - (1-y)\ln(1-\hat{y})$, where $\hat{y} = \text{MLP}(X)$.

Algorithm 1: Training of the Neural Network

Data: $(\mathbf{Z}_{train}, \mathbf{y}_{train})$.
Input: Network MLP, number of epochs n_{epoch}, penalization strength λ.
Result: Trained network.

for $n = 0, \ldots,\ n_{epoch}$ **do**
  for $X, y \in \mathbf{Z}_{train}, \mathbf{y}_{train}$ **do**
    compute $\hat{y} = \text{MLP}$;
    compute $BCE = BCE(\hat{y}, y)$, backpropagate and zero_grad ;
  end
  compute $\hat{\mathbf{y}} = \text{MLP}(\mathbf{Z}_{train})$;
  compute $\text{BCE} = \text{BCE}(\hat{\mathbf{y}}, \mathbf{y})$ and $\widehat{W}^{\phi}_{n,m}$;
  compute the regularized loss $\text{BCE} - \lambda \widehat{W}^{\phi}_{n,m}$, backpropagate and zero_grad ;
end

For the first learning loop, we saved the evolution of the BCE losses, for all values of λ, computed at each epoch together with the W_ϕ proxy and the accuracy metric for $n_{lowest} = 75$. As displayed in Figure 5, one can see that the incorporation of the empirical W_ϕ criterion in the penalization term improves the performances for a well chosen parameter λ. For instance, $\lambda \in \{1, 10\}$ output the best results in this setting.

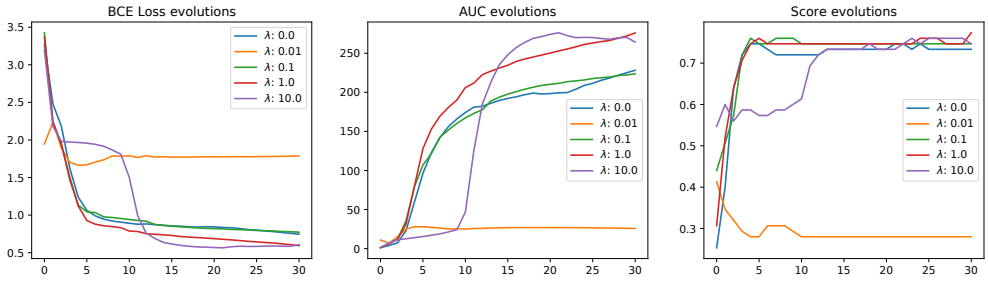


Figure 5: Evolutions of the BCE loss, the AUC proxy and the accuracy for $n_{lowest} = 75$ in function of the epochs, for $\phi(u) = u$ and all values of the hyperparameter $\lambda \in \{0, 0.01, 0.1, 1, 10\}$.

At the end of the training, we select the network having the highest empirical W_ϕ score, which here corresponds to choosing $\lambda = 1$. We then score the initial observations $X_1, \ldots, X_n$ and display in Figure 6 the points with an intensity varying from red to blue as the score increases from 0 to 1. The fact that the red points are on the sides of the dataset empirically validates our methodology. We represent in Fig. 7 the averaged mass volume curve together with standard deviation computed for $\lambda = 1$ over $B = 50$ repetitions. Table 1 gathers the results averaged over $B = 50$ repetitions. Notice that these results support the soundness of our approach. Indeed, the area under the MV curve is minimized and the proportion of detected outliers is high even when n_{lowest} increases.

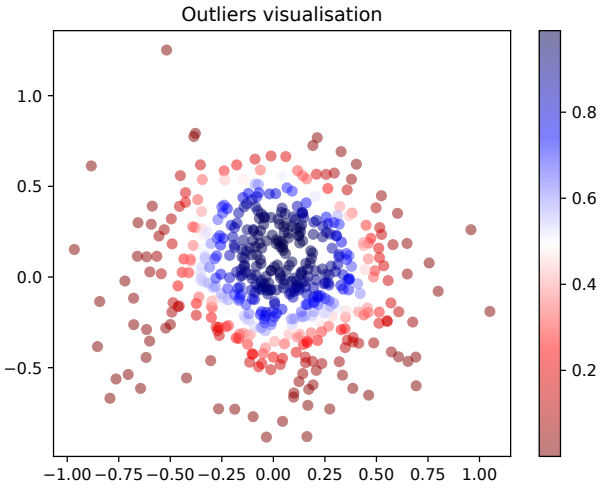
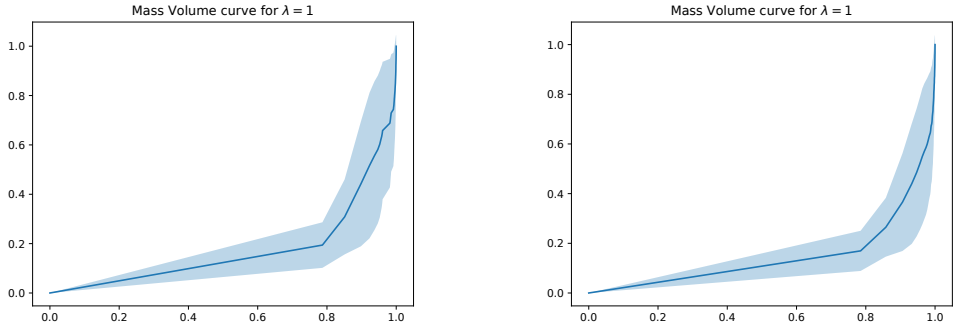


Figure 6: A heatmap of the scores for $\phi(u) = u$.

n_{lowest}	25	50	75	100
$\text{Acc}_{n_{lowest}}$	0.91 ± 0.13	0.84 ± 0.15	0.74 ± 0.15	0.64 ± 0.13

Table 1: Tabular view of the empirical accuracy $+-$ its standard deviation, when n_{lowest} varies in $\{25, 50, 75, 100\}$, with $\lambda = 1$.



(a) $\lambda = 1$ and $\phi(u) = u$.

(b) $\lambda = 1$ and $\phi(u) = u\mathbb{I}\{u \geq u_0\}$.

Figure 7: Empirical Mass-Volume curves.

5. Conclusion

In this paper, we promoted a binary classification approach to the problem of learning to rank anomalies. We established a clear theoretical link between these two machine learning tasks through the study of the mass-volume curve. In particular, our procedure is robust with respect to imbalanced datasets through the choice of the parameter p that is chosen

initially in practice. Previous results (see Clémençon et al. (2021)) support the effectiveness of our methodology. Moreover, we illustrate our method with numerical experiments of synthetic data.

Acknowledgments

We thank Yannick Guyonvarch for his insightful comments. Moreover, we are greatly indebted to the chair DSAIDIS of Telecom Paris and to the Région Ile-de-France for the support.

References

L. Bergman and Y. Hoshen. Classification-Based Anomaly Detection for General Data. *arXiv:2005.02359*, 2020.

M.M. Breunig, H.P. Kriegel, R.T. Ng, and J. Sander. Lof: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pages 93–104, 2000.

S. Clémençon and N. Vayatis. Tree-based ranking methods. *IEEE Transactions on Information Theory*, 55(9):4316–4336, 2009.

S. Clémençon, M. Limnios, and N. Vayatis. Concentration Inequalities for Two-Sample Rank Processes with Application to Bipartite Ranking. *arXiv:2104.02943*, 2021.

S. Clémençon and A. Thomas. Mass volume curves and anomaly ranking. *Electronic Journal of Statistics*, 12(2):2806 – 2872, 2018.

J. Frery, A. Habrard, M. Sebban, O. Caelen, and L. He-Guelton. Efficient top rank optimization with gradient boosting for supervised anomaly detection. In *European Conference on Machine Learning & Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD'17)*, 2017.

J. Hájek. Asymptotically most powerful rank-order tests. *The Annals of Mathematical Statistics*, 33(3):112–1147, 09 1962.

F.T. Liu, K.M. Ting, and Z.H. Zhou. Isolation forest. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on Data Mining*, pages 413–422, 2008.

B. Schölkopf, J. Platt, A. J. Shawe-Taylor, J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7), 2001.

I. Steinwart, D. Hush, and C. Scovel. A classification framework for anomaly detection. *Journal of Machine Learning Research*, 6(8):211–232, 2005.

F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1:80–83, 1945.