Anomaly Ranking in a High Dimensional Space: The UNSUPERVISED TREERANK Algorithm

S. Clémençon, N. Baskiotis, and N. Vayatis

Abstract Ranking unsupervised data in a multivariate feature space $\mathscr{X} \subset \mathbb{R}^d$, $d \geq 1$ by degree of abnormality is of crucial importance in many applications (e.g., fraud surveillance, monitoring of complex systems/infrastructures such as energy networks or aircraft engines, system management in data centers). However, the learning aspect of unsupervised ranking has only received attention in the machine-learning community in the past few years. The Mass-Volume (MV) curve has been recently introduced in order to evaluate the performance of any scoring function $s : \mathscr{X} \to \mathbb{R}$ with regard to its ability to rank unlabeled data. It is expected that relevant scoring functions will induce a preorder similar to that induced by the density function $f(x)$ of the (supposedly continuous) probability distribution of the statistical population under study. As far as we know, there is no efficient algorithm to build a scoring function from (unlabeled) training data with nearly optimal MV curve when the dimension d of the feature space is high. It is the major purpose of this chapter to introduce such an algorithm which we call the UNSUPERVISED TREERANK algorithm. Beyond its description and the statistical analysis of its performance, numerical experiments are exhibited in order to provide empirical evidence of its accuracy.

S. Clémençon (✉)
Institut Mines Telecom, LTCI UMR, Telecom ParisTech & CNRS No. 5141,
46 rue Barrault, 75013 Paris, France
e-mail: stephan.clemencon@telecom-paristech.fr

N. Baskiotis
Université Paris, 6 Pierre et Marie Curie, LIP6 UMR CNRS No. 7606,
Place Jussieu, 75005 Paris, France
e-mail: nicolas.baskiotis@lip6.fr

N. Vayatis
ENS Cachan, CMLA, UMR CNRS No. 8536, Cachan, France
e-mail: nicolas.vayatis@cmla.ens-cachan.fr

M.E. Celebi, K. Aydin (eds.), *Unsupervised Learning Algorithms*,
DOI 10.1007/978-3-319-24211-8_2

1 Introduction

The issue of ranking multivariate data by degree of abnormality, which shall be referred to as *anomaly ranking* throughout the present chapter, is essential for a wide variety of applications, ranging from fraud surveillance to distributed fleet monitoring through system management in data centers (see, e.g., [1–3]). Anomaly/novelty detection has been the subject of a good deal of attention in the machine-learning literature these past ten years (see [4–8] or [9] among others). In most of these papers, unsupervised learning methods are proposed to decide whether an observation x in the feature space $\mathscr{X}$ lies among the $\alpha\%$ the most abnormal observations $\Omega_\alpha^* \subset \mathscr{X}$. However, the interest in applications is sometimes to learn how to rank all possible observations $x_1, \ldots, x_n$ by degree of abnormality. As pointed out in [10], the anomaly ranking problem can be viewed as a *continuum* of "imbricated" anomaly detection problems, requiring to recover from training data the collection of nested sets $\{\Omega_\alpha^*;\ \alpha \in (0,1)\}$. Indeed, in contrast to anomaly detection rules, an anomaly ranking function would permit to prioritize the action check-list and allows for a progressive examination of the situations predicted as most abnormal. Although ad-hoc procedures (e.g., [11]) for ranking "outliers" among a statistical population, either fully data-driven or based on prior expertise, have been proposed in support of various problems (e.g., predictive maintenance, fraud surveillance), the issue of measuring the performance of orderings induced by scoring functions has only been considered in [10], where a Probability-Measure plot, termed Mass-Volume curve (MV curve in abbreviated form), has been introduced for this specific purpose. In [12], when the feature space $\mathscr{X}$ is supposed to be compact, the connection between MV curve minimization and the (supervised) bipartite ranking related to the discrimination between the distribution of the data observed and the uniform distribution on $\mathscr{X}$ has been highlighted, offering the possibility, in theory at least, to turn bipartite ranking algorithms into unsupervised versions based on a pooled dataset formed by the original observations plus a sample of simulated data uniformly distributed on $\mathscr{X}$. However, one faces computational difficulties when trying to implement such an approach when the dimension of the feature space increases: a simulated data sample of "reasonable" size would be unable to fill a high dimensional space with enough points, consequently compromising the supervised learning procedure. Building on the connection between anomaly ranking and supervised bipartite ranking, the goal of this chapter is to introduce a novel algorithm for anomaly ranking, referred to as UNSUPERVISED TREERANK. It can be viewed as an extension of the TREERANK algorithm introduced in [13, 14] when the LEAFRANK splitting procedure is a cost-sensitive version of the popular CART algorithm with axis parallel splits, in the specific case where one of the two distributions is known and coincides with the uniform distribution $\mathscr{U}_{\mathscr{X}}(dx)$ on $\mathscr{X}$. In this case, the actual computation of the true cell volume (or $\mathscr{U}_{\mathscr{X}}$-measure) along the recursive partitioning is straightforward and permits to avoid the simulation stage.

The chapter is organized as follows. Key notions involved in the formulation of the *anomaly ranking* problem (including the form of the decision rules, the criterion to measure accuracy and the related optimal elements), as well as its connection with the (supervised) *bipartite ranking* problem, are recalled in Sect. 2. The UNSUPERVISED TREERANK algorithm to optimize the MV curve is described in Sect. 4. Numerical results empirically assessing the performance of the method promoted are displayed in Sect. 5. Finally, several concluding remarks are collected in Sect. 6.

2 Anomaly Ranking: Background and Preliminaries

We start off with introducing notations that shall be used throughout the chapter and recalling key notions related to the anomaly ranking problem. Throughout the chapter, the Lebesgue measure on $\mathbb{R}^d$ is denoted by λ, the indicator function of any event $\mathscr{E}$ by $\mathbb{I}\{\mathscr{E}\}$ and the generalized inverse of any cumulative distribution function $K(t)$ on $\mathbb{R}$ by $K^{-1}(u) = \inf\{t \in \mathbb{R} : \ K(t) \geq u\}$.

2.1 A Scoring Approach to Anomaly Ranking

We consider a generic random variable X taking its values in a measurable space $\mathscr{X}$, assumed to be a subset of the Euclidean space $\mathbb{R}^d$, $d \geq 1$, for simplicity. We denote by $F(dx)$ its supposedly continuous probability distribution and by $f(x)$ the related density function. The observations $X_1, \ldots, X_n$, with $n \geq 1$, are modeled as independent copies of the r.v. X. The most convenient way of defining a preorder on the feature space $\mathscr{X}$ is undoubtedly to transport the natural order on the real half-line $\mathbb{R}_+$ onto it by means of a *scoring function*, i.e., a Borel function $s : \mathscr{X} \to \mathbb{R}_+$: given two observations x and x' in $\mathscr{X}$, x will be said as more abnormal according to s than x' when $s(x) \leq s(x')$. We denote by $\mathscr{S}$ the set of all scoring functions on $\mathscr{X}$ that are integrable with respect to Lebesgue measure on $\mathscr{X}$. Notice incidentally that this condition is not restrictive insofar as the preorder induced by any scoring function is invariant by strictly increasing transformation (i.e., the scoring function s and its transform $T \circ s$ define the same preorder on the feature space $\mathscr{X}$ provided that the Borel transformation $T : \mathrm{Im}(s) \to \mathbb{R}_+$ is strictly increasing on the image of the r.v. $s(X)$, which is denoted by $\mathrm{Im}(s)$). An explanation for this integrability constraint is given in the next section.

Informally, the goal pursued is to build from the training dataset $\{X_1, \ldots, X_n\}$ a scoring function s such that, ideally, the smaller $s(X)$, the more abnormal the observation X. Stated this way, the set of optimal scoring rules in $\mathscr{S}$ is the set of strictly increasing transforms of the density function $f(x)$ that are integrable w.r.t. to Lebesgue measure

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

The next section describes a criterion, whose optimal elements coincides with $\mathscr{S}^*$. It is of *functional* nature (i.e., it takes its values in a path space), which was expected, insofar as we previously pointed out that the anomaly scoring problem could be viewed as a continuum of (nested) anomaly detection problems.

The technical assumptions listed below are involved in the subsequent analysis.

H₁ The r.v. $f(X)$ is continuous, i.e., $\forall c \in \mathbb{R}_+, \mathbb{P}\{f(X) = c\} = 0$.

H₂ The density function $f(x)$ is bounded: $||f||_\infty \stackrel{def}{=} \sup_{x \in \mathscr{X}} |f(x)| < +\infty$.

2.2 Measuring Scoring Accuracy: The Mass-Volume Curve

Let $s \in \mathscr{S}$ be an arbitrary scoring function. Its level sets are denoted by $\Omega_{s,t} = \{x \in \mathscr{X} : \ s(x) \geq t\}$ for all $t \geq 0$. Observe that, since s is supposed to be λ-integrable, the measure $\lambda(\Omega_{s,t}) \leq (\int_{u \in \mathbb{R}_+} s(u)du)/t$ is finite for any $t > 0$. A natural measure of the anomaly ranking performance of a given scoring function $s \in \mathscr{S}$ has been introduced in [10]. It is the Probability-Measure plot, termed Mass-Volume curve (MV curve in short), given by:

$$t > 0 \mapsto (\mathbb{P}\{s(X) \geq t\}, \ \lambda(\{x \in \mathscr{X} : \ s(x) \geq t\})) = (F(\Omega_{s,t}), \ \lambda(\Omega_{s,t})) \, . \quad (1)$$

Connecting points corresponding to possible jumps, this parametric curve can be seen as the plot of a 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. See Fig. 1 for a typical MV curve.

Let $\alpha \in (0, 1)$. Denoting by $F_s(t)$ the cumulative distribution function of the r.v. $s(X)$, we have:

$$\mathrm{MV}_s(\alpha) = \lambda\left(\{x \in \mathscr{X} : \ s(x) \geq F_s^{-1}(1 - \alpha)\}\right), \quad (2)$$

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 c. Let $(s_1, s_2) \in \mathscr{S}^2$, 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. Indeed, we recall that, under the Assumptions **H₁**–**H₂**, the set

$$\Omega_\alpha^* \stackrel{def}{=} \{x \in \mathscr{X} : \ f(x) \geq Q^*(\alpha)\}$$

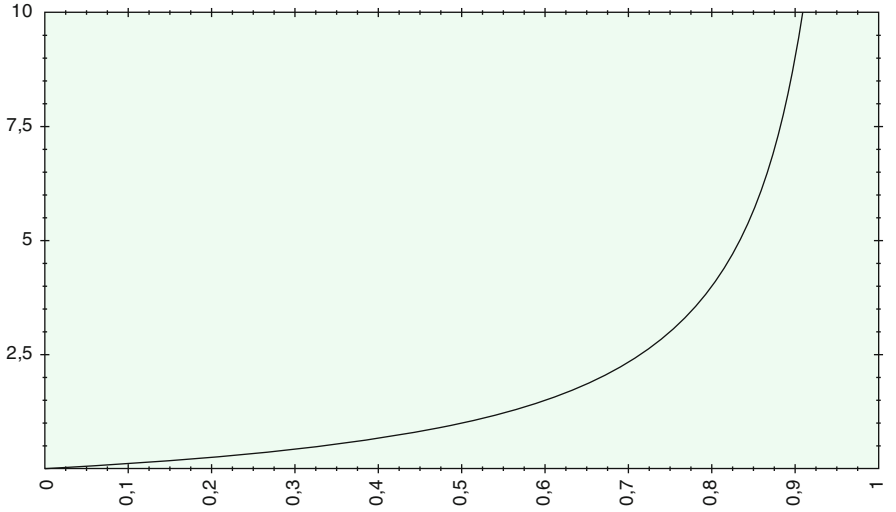

Fig. 1 A typical MV curve (x-axis: mass, y-axis: volume)

is the unique solution of the *minimum volume set* problem

$$\min_{\Omega \in \mathscr{B}(X)} \lambda(\Omega) \text{subjectto} F(\Omega) \geq \alpha, \tag{3}$$

where $\mathscr{B}(\mathscr{X})$ denotes the ensemble made of all Borel subsets of $\mathscr{X}$ and $Q^*(\alpha) = F_f^{-1}(1-\alpha)$ denotes the quantile of level $1-\alpha$ of the r.v. $f(X)$. For small values of the mass level α, minimum volume sets are expected to contain the modes of the distribution, whereas their complementary sets correspond to "abnormal observations" when considering large values of α. Refer to [4, 15] for an account of minimum volume set theory and to [16] for related statistical learning results. This implies in particular that optimal scoring functions are those whose MV curves are minimum everywhere, as shown by the following summary of some crucial properties of the optimal MV curve, which plots the (minimum) volume $\lambda(\Omega_\alpha^*)$ against the mass $F(\Omega_\alpha^*) = \alpha$.

Proposition 1 ([10]). *Let the assumptions* $\mathbf{H_1}$–$\mathbf{H_2}$ *be fulfilled.*

(i) *The elements of the class* $\mathscr{S}^*$ *have the same* MV *curve and provide the best possible preorder on* $\mathscr{X}$ *in regard to the* MV *curve criterion:*

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

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

(ii) *In addition, we have:* $\forall (s, \alpha) \in \mathscr{S} \times (0,1)$,

$$0 \leq \mathrm{MV}_s(\alpha) - \mathrm{MV}^*(\alpha) \leq \lambda\left(\Omega_\alpha^* \Delta \Omega_{s,Q(s,\alpha)}\right), \tag{5}$$

where Δ denotes the symmetric difference operation between two sets, and $Q(s,\alpha)$ denotes the quantile of level $1-\alpha$ of the r.v. $s(X)$.

(iii) *The optimal* MV *curve is convex. It is also differentiable provided that the density $f(x)$ is differentiable with a gradient taking non zero values on the boundary $\partial\Omega_\alpha^* = \{x \in \mathscr{X} : f(x) = Q^*(\alpha)\}$.*

Equation (5) reveals that the lowest the MV curve (everywhere) of a scoring function $s(x)$, the closer the preorder defined by $s(x)$ 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 corresponds 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 in particular that the optimal curve MV^* somehow measures the spread of the distribution $F(dx)$: attention should be paid to the curve for large values of α when focus is on 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).

Given this performance criterion, it becomes possible to develop a statistical theory for the anomaly scoring problem. In a statistical estimation perspective, 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^*. Of course, closeness between (continuous) curves can be measured in many ways. The L_p-distances, $1 \leq p \leq +\infty$ can be used for this purpose for instance. Fix $\epsilon \in (0,1)$ and consider in particular the losses related to the L_1-distance and to the sup-norm:

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

$$d_\infty(s,f) = \sup_{\alpha\in[0,1-\epsilon]} |\mathrm{MV}_s(\alpha) - \mathrm{MV}^*(\alpha)| = \sup_{\alpha\in[0,1-\epsilon]} \big(\mathrm{MV}_s(\alpha) - \mathrm{MV}^*(\alpha)\big).$$

We point out that $d_i(s,f)$, $i \in \{1, \infty\}$, is not a distance between the scoring functions s and f but measures the dissimilarity between the preorders they induce. In addition, notice that minimizing $d_1(s,f)$ boils down to minimizing the scalar quantity $\int_{\alpha=0}^{1-\epsilon} \mathrm{MV}_s(\alpha) d\alpha$, the area under the MV curve over $[0,\ 1-\epsilon]$.

In practice, the MV curve of a scoring function $s \in \mathscr{S}$ 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

$$\alpha \in (0,1) \mapsto \widehat{\mathrm{MV}}_s(\alpha) \stackrel{def}{=} \lambda\left(\left\{x \in \mathscr{X} :\ s(x) \geq \hat{F}_s^{-1}(1-\alpha)\right\}\right), \tag{6}$$

where $\hat{F}_s(t) = (1/n)\sum_{i=1}^n \mathbb{I}\{s(X_i) \leq t\}$ denotes the empirical cdf of the r.v. $s(X)$ and $\hat{F}_s^{-1}$ its generalized inverse. In [10], for a fixed $s \in \mathscr{S}$, consistency and

asymptotic Gaussianity (in sup-norm) of the estimator (6) has been established, together with the asymptotic validity of a smoothed bootstrap procedure to build confidence regions in the MV space.

Remark 1 (Volume Estimation). We point out that the exact computation of (empirical) level sets of a scoring function can be numerically unfeasible and approximation schemes (*e.g., Monte Carlo procedures*) must be implemented in practice for this purpose. However, for scoring functions that are constant on the cells of a partition of the feature space $\mathscr{X}$ expressible as union of hypercubes (such as those produced by the UNSUPERVISED TREERANK algorithm, as shall be seen below), such approximation/estimation methods, which are extremely challenging in high dimension, can be avoided.

3 Turning Anomaly Ranking into Bipartite Ranking

We now explain in detail the connection between anomaly ranking and (supervised) bipartite ranking. Motivated by a wide variety of applications ranging from the design of diagnosis support tools in medicine to supervised anomaly detection in signal processing through credit-risk screening in finance, learning to rank data with a feedback provided by binary labels has recently received much attention in the machine-learning literature, see, e.g., [17–19]. Several performance criteria have been considered in order to formulate the issue of ranking observations in an order as close as possible to that induced by the ordinal output variable as a M-estimation problem: the (area under the) receiver operator characteristic curve (ROC curve in short) and the precision-recall curve, the normalized discounted cumulative gain criterion. Many practical ranking algorithms, supported by sound theoretical results extending the probabilistic theory of pattern recognition, have been introduced in the literature, see, e.g., [13, 20–22].

3.1 Bipartite Ranking and ROC *Analysis*

Let $G(dx)$ and $H(dx)$ be two probability distributions on the feature space $\mathscr{X}$, supposedly absolutely continuous with respect to each other. The ROC curve of any scoring function $s(x)$ is then defined as the Probability–Probability plot $t > 0 \mapsto (1 - H_s(t), 1 - G_s(t))$, where $H_s(dt)$ and $G_s(dt)$ respectively denote the images of the distributions H and G by the mapping $s : \mathscr{X} \to \mathbb{R}_+$. Connecting by convention possible jumps by line segments, the ROC curve of the scoring function $s(x)$ can always be viewed as the plot of a continuous mapping $\mathrm{ROC}_s : \alpha \in (0, 1) \mapsto \mathrm{ROC}_s(\alpha)$. It starts at $(0, 0)$ and ends at $(1, 1)$. At any point $\alpha \in (0, 1)$ such that $H_s \circ H_s^{-1}(\alpha) = \alpha$, we have: $\mathrm{ROC}_s(\alpha) = 1 - G_s \circ H_s^{-1}(1 - \alpha)$. The curve ROC_s measures the capacity of s to discriminate between distributions H

and G. It coincides with the first diagonal when $H_s = G_s$. Observe also that the *stochastically larger* than H_s the distribution G_s, the closer to the left upper corner of the ROC space the curve ROC_s. One may refer to [23] for an account of ROC analysis and its applications.

The concept of ROC curve induces a partial order on $\mathscr{S}$. A scoring function s_1 is more accurate than s_2 iff: $\forall \alpha \in (0,1)$, $\mathrm{ROC}_{s_1}(\alpha) \geq \mathrm{ROC}_{s_2}(\alpha)$. A Neyman-Pearson argument shows that the optimal ROC curve, denoted by ROC^*, is that of the likelihood ratio statistic $\phi(x) = dG/dH(x)$. It dominates any other ROC curve everywhere: $\forall (s,\alpha) \in \mathscr{S} \times (0,1)$, $\mathrm{ROC}_s(\alpha) \leq \mathrm{ROC}^*(\alpha)$. The set $\mathscr{S}^*_{H,G} = \{T \circ \phi,\ T : \mathrm{Im}\phi(X) \to \mathbb{R}_+ \mathrm{strictlyincreasing}\}$ is the set of optimal scoring functions regarding the bipartite problem considered.

The goal of bipartite ranking is to build a scoring function with a ROC curve as high as possible, based on two independent *labeled* datasets: $(X_1^-, \ldots, X_m^-)$ and $(X_1^+, \ldots, X_q^+)$ made of independent realizations of H and G, respectively, with $m,\ q \geq 1$. Assigning the label $Y = +1$ to observations drawn from $G(dx)$ and label $Y = -1$ to those drawn from $H(dx)$, the objective can be also expressed as to rank/score any pooled set of observations (in absence of label information) so that, ideally, the higher the score of an observation X, the likelier its (hidden) label Y is positive.

The accuracy of any $s \in \mathscr{S}$ can be measured by:

$$D_p(s, s^*) = ||\mathrm{ROC}_s - \mathrm{ROC}^*||_p, \tag{7}$$

where $s^* \in \mathscr{S}^*_{H,G}$ and $p \in [1, +\infty]$. Observe that, in the case $p = 1$, one may write $D_1(s, s^*) = \mathrm{AUC}^* - \mathrm{AUC}(s)$, where $\mathrm{AUC}(s) = \int_{\alpha=0}^1 \mathrm{ROC}_s(\alpha) d\alpha$ is the *Area Under the* ROC *Curve* (AUC in short) and $\mathrm{AUC}^* = \mathrm{AUC}(\phi)$ is the maximum AUC. Hence, minimizing $D_1(s, s^*)$ boils down to maximizing the ROC summary $\mathrm{AUC}(s)$. The popularity of this quantity arises from the fact it can be interpreted, in a probabilistic manner, as the *rate of concording pairs*

$$\mathrm{AUC}(s) = \mathbb{P}\left\{s(X) < s(X')\right\} + \frac{1}{2}\mathbb{P}\left\{s(X) = s(X')\right\}, \tag{8}$$

where X and X' denote independent r.v.'s defined on the same probability space, drawn from H and G, respectively. An empirical counterpart of $\mathrm{AUC}(s)$ can be straightforwardly derived from (8), paving the way for the implementation of "empirical risk minimization" strategies, see [18].

The algorithms proposed to optimize the AUC criterion or surrogate performance measures are too numerous to be listed in an exhaustive manner. Among methods well-documented in the literature, one may mention in particular the TreeRank method and its variants (see [13, 14, 24]), which relies on recursive AUC maximization, the RankBoost algorithm, which implements a boosting approach tailored for the ranking problem (see [20]), the SVMrank algorithm originally designed for ordinal regression (see [25]) and the RankRLS procedure proposed in [22].

3.2 A Bipartite View of Anomaly Ranking

With the notations of Sect. 3.1, we take $H(dx)$ as the uniform distribution $U(dx)$ on $[0,1]^d$ and $G(dx)$ as $F(dx)$, the distribution of interest in the *anomaly ranking* problem. It follows immediately from the definitions and properties recalled in Sect. 2 that, for any scoring function $s \in \mathscr{S}$, the curves MV_s and ROC_s are symmetrical with respect to the first diagonal of the unit square $[0,1]^2$. Hence, as stated in the next result, solving the anomaly ranking problem related to distribution $F(dx)$ is equivalent to solving the bipartite ranking problem related to the pair (U, F).

Theorem 1. *Suppose that assumptions* $\mathbf{H_1}$, $\mathbf{H_2}$ *hold true. Let* $U(dx)$ *be the uniform distribution on* $[0,1]^d$. *For any* $(s,\alpha) \in \mathscr{S} \times (0,1)$, *we have:* $\mathrm{ROC}_s^{-1}(\alpha) = \mathrm{MV}_s(\alpha)$. *We also have* $\mathscr{S}^* = \mathscr{S}^*_{U,F}$, *and*

$$\forall (s, s^*) \in \mathscr{S} \times \mathscr{S}^*, \quad D_p(s, s^*) = d_p(s, s^*),$$

for $1 \leq p \leq +\infty$. *In particular, we have:* $\forall s \in \mathscr{S}$,

$$1 - \int_{\alpha=0}^{1} \mathrm{MV}_s(\alpha) d\alpha = \mathbb{P}\{s(W) < s(X)\} + \frac{1}{2}\mathbb{P}\{s(W) < s(X)\},$$

where W *and* X *are independent r.v.'s, drawn from* $U(dx)$ *and* $F(dx)$ *respectively.*

The proof is straightforward, it suffices to observe that $\phi = dG/dH = f$ in this context. Details are left to the reader. Incidentally, we point out that, under the assumptions listed above, the minimal area under the MV curve may be thus interpreted as a measure of dissimilarity between the distribution $F(dx)$ and the uniform distribution on $[0,1]^d$. The closer $\int_0^1 \mathrm{MV}^*(\alpha)d\alpha$ to $1/2$, the more similar to $U(dx)$ the distribution $F(dx)$.

Remark 2 (On the Support Assumption). In general, the support of $F(dx)$ is unknown, just like the distribution itself. However, the argument above remains valid in the case where $\mathrm{Supp}(F(dx)) \subset [0,1]^d$. The sole difference lies in the fact that the curve MV^* then ends at the point of mass-axis coordinate equal to one and volume-axis coordinate $\lambda(\mathrm{Supp}(F)) \leq 1$, the corresponding curve ROC^* exhibiting a plateau: it reaches the value one from the false positive rate $\lambda(\mathrm{Supp}(F))$. We point out that, when no information about the support is available, one may always carry out the analysis for the conditional distribution given $X \in \mathscr{C}$, $F\,|_{\mathscr{C}}\,(dx)$, where $\mathscr{C}$ denotes any compact set containing the observations $X_1, \ldots, X_n$. Observe in addition that, when $\mathscr{C} = \{x \in \mathscr{X} : f(x) > t\}$ for $t > 0$, the optimal MV curve related to $F\,|_{\mathscr{C}}\,(dx)$ coincides with that related to $F(dx)$ on $[0, F(\mathscr{C})]$, as can be immediately seen with the change of parameter $t \to t/\int_{\mathscr{C}} f(x)dx$ in (1).

3.3 Extending Bipartite Methods via Uniform Sampling

Now that the connection between anomaly ranking and bipartite ranking has been highlighted, we show how to exploit it to extend efficient algorithms proposed in the supervised framework to MV curve minimization. Learning procedures are based on a training i.i.d. sample $X_1, \ldots, X_n$, distributed according to the unknown probability measure $F(dx)$ with compact support, included in $[0, 1]^d$ say.

One may extend the use of any bipartite ranking algorithm $\mathscr{A}$ to the unsupervised context by simulating extra data, uniformly distributed on the unit hypercube, as follows.

ONE-CLASS SCORING VIA UNIFORM SAMPLING

Input: *unlabeled data sample* $\{X_1, \ldots, X_n\}$, *bipartite ranking algorithm* $\mathscr{A}$, $m \geq 1$

1. *Sample additional data* $X_1^-, \ldots, X_m^-$, *uniformly distributed over* $[0, 1]^d$.
2. *Assign a "negative" label to the sample* $\mathscr{D}_m^- = \{X_1^-, \ldots, X_m^-\}$ *and a "positive" label to the original data* $\mathscr{D}_n^+ = \{X_1, \ldots, X_n\}$.
3. *Run algorithm* $\mathscr{A}$ *based on the bipartite statistical population* $\mathscr{D}_m^- \cup \mathscr{D}_n^+$, *producing the anomaly scoring function* $s(x)$.

Except the choice of the algorithm $\mathscr{A}$ and the selection of its hyperparameters, the sole tuning parameter which must be set is the size m of the uniformly distributed sample. In practice, it should be chosen as large as possible, depending on the current computational constraints. From a practical perspective, it should be noticed that the computational cost of the sampling stage is reduced. Indeed, the d components of a r.v. uniformly distributed on the hypercube $[0, 1]^d$ being independent and uniformly distributed according to the uniform distribution on the unit interval, the "negative" sample can be thus generated by means of pseudo-random number generators (PRNG's), involving no complex simulation algorithm. Furthermore, uniform distributions on any (Borel) subset of $[0, 1]^d$ can be naturally simulated in a quite similar fashion, with an additional conditioning step.

We point out that, in the context of density estimation, a similar sampling technique for transforming this unsupervised problem into one of supervised function approximation is discussed in Sect. 14.2.4 in [26], where it is used in particular to build *generalized association rules*. This idea is also exploited in [7] for anomaly detection, see also [27]. In this respect, it should be mentioned that a variety of techniques, including that proposed in [6] where the SVM machinery has been extended to the unsupervised framework and now referred to as ONE CLASS SVM, have been proposed to recover the set Ω_α^* for a target mass level $\alpha \in (0, 1)$, fixed in advance. Therefore, even if the estimates produced of are of the form $\{x \in \mathscr{X} : \hat{f}(x) > t_\alpha\}$ and one could consider using the decision function $\hat{f}(x)$ as scoring function, one should keep in mind that there is no statistical guarantee

that the ensembles $\{x \in \mathscr{X} : \hat{f}(x) > t\}$ are good estimates of density level sets for $t \neq t_\alpha$. This explains the poor performance of such a "plug-in" approach observed in practice.

4 The UNSUPERVISED TREERANK Algorithm

The major drawback of the simulation-based approach described above arises from the fact that sampling a "dense" dataset representative of the uniform distribution rapidly becomes unfeasible in practice as the dimension d of the space increases. In high dimension, uniformly distributed datasets of "reasonable" sizes will inevitably fill extremely sparsely the feature space, jeopardizing the supervised learning procedure. In contrast, the algorithm promoted in this section is not confronted with this difficulty, since it completely avoids simulation of uniform data. Like other recursive tree building methods, it is implemented in two steps: growing first a possibly overfitted decision tree model and selecting next a submodel with highest expected generalization ability by pruning the original tree.

4.1 Anomaly Ranking Trees

In supervised and unsupervised learning problems, decision trees are undoubtedly among the most popular techniques. One of the main reasons arises from the fact that they straightforwardly offer a visual model summary, taking the form of an easily interpretable binary tree graph, refer to [28, 29] or [30] for instance. Indeed, predictions can be generally described by means of a hierarchical combination of elementary rules of the type "$X^{(j)} \leq \kappa$" or "$X^{(j)} > \kappa$," comparing the value taken by a (quantitative) component of the input vector X (the *split variable*) to a certain threshold (the *split value*). In contrast to (supervised) learning problems such as classification or regression, which are of local nature, predictive rules for a global problem such as *anomaly ranking* cannot be described by a (tree-structured) partition of the feature space: cells (corresponding to the terminal leaves of the binary decision tree) must be ordered so as to define a scoring function. Hence, we define an *anomaly ranking trees* as a binary tree equipped with a "left-to-right" orientation, defining a tree-structured collection of anomaly scoring functions, as depicted by Fig. 2. The root node of an anomaly ranking tree $\mathscr{T}_J$ of depth $J \geq 0$ represents the whole feature space $\mathscr{X}$: $\mathscr{C}_{0,0} = \mathscr{X}$, while each internal node (j, k) with $j < J$ and $k \in \{0, \ldots, 2^j - 1\}$ corresponds to a subset $\mathscr{C}_{j,k} \subset \mathscr{X}$, whose left and right siblings respectively correspond to disjoint subsets $\mathscr{C}_{j+1,2k}$ and $\mathscr{C}_{j+1,2k+1}$ such that $\mathscr{C}_{j,k} = \mathscr{C}_{j+1,2k} \cup \mathscr{C}_{j+1,2k+1}$. Equipped with the left-to-right orientation, any subtree $\mathscr{T} \subset \mathscr{T}_J$ defines a preorder on $\mathscr{X}$: elements lying in the same terminal cell of $\mathscr{T}$ being equally ranked. The anomaly scoring function related to the oriented tree $\mathscr{T}$ can be written as:

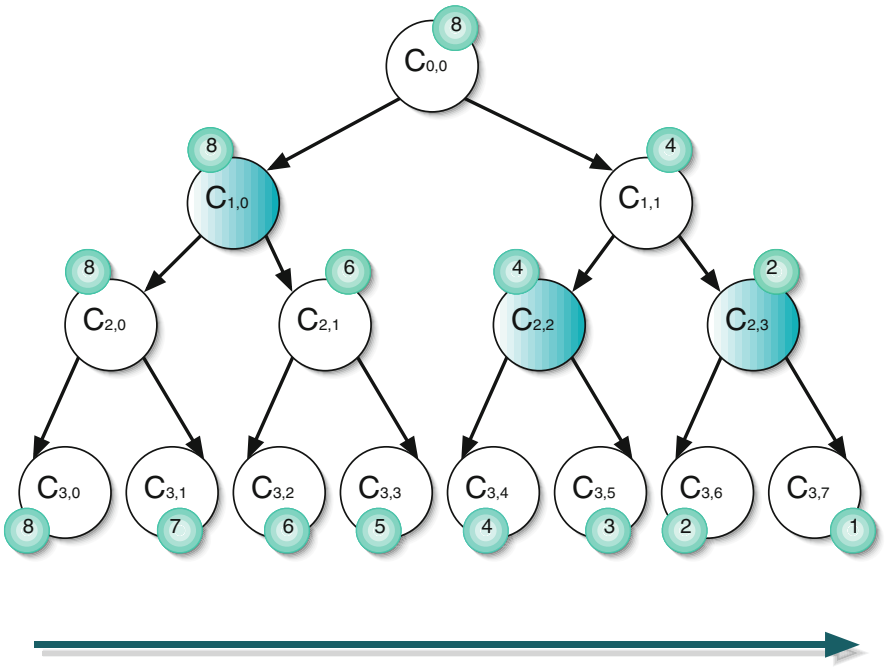

Fig. 2 An anomaly scoring function described by an oriented binary subtree $\mathscr{T}$. For any element $x \in \mathscr{X}$, the quantity $s_{\mathscr{T}}(x)$ can be computed very fast in a top-down manner using the heap structure: starting from the initial value 2^J at the root node, at each internal node $\mathscr{C}_{j,k}$, the score remains unchanged if x moves down to the *left* sibling and one subtracts $2^{J-(j+1)}$ from it if x moves down to the *right*

$$s_{\mathscr{T}}(x) = \sum_{\mathscr{C}_{j,k}\text{: terminal leaf of } \mathscr{T}} 2^J \left(1 - \frac{k}{2^j}\right) \cdot \mathbb{I}\{x \in \mathscr{C}_{j,k}\}. \tag{9}$$

Suppose that the feature space $\mathscr{X}$ is compact for simplicity. Then, observe that the MV curve of the anomaly scoring function $s_{\mathscr{T}}(x)$ is the piecewise linear curve connecting the knots:

$$(0,0) \text{ and } \left(\sum_{l=0}^{k} F(\mathscr{C}_{j,l}),\ \sum_{l=0}^{k} \lambda(\mathscr{C}_{j,l})\right) \text{ for all terminal leaf } \mathscr{C}_{j,k} \text{ of } \mathscr{T}.$$

A statistical version can be computed by replacing the $F(\mathscr{C}_{m,l})$'s by their empirical counterpart. However, as pointed out in Remark 1, the evaluation of the volume $\lambda(\mathscr{C}_{m,l})$ can be problematic unless the geometry of the cells forming the partition is appropriate (e.g., union of hypercubes).

4.2 The Algorithm: Growing the Anomaly Ranking Tree

The TREERANK algorithm, a bipartite ranking technique optimizing the ROC curve in a recursive fashion, has been introduced in [13] and its properties have been investigated in [14] at length. Its output consists of a tree-structured scoring rule (9) with a ROC curve proved to be nearly optimal under mild assumptions. The growing stage is performed as follows. At the root, one starts with a constant scoring function $s_1(x) = \mathbb{I}\{x \in \mathscr{C}_{0,0}\} \equiv 1$ and after $m = 2^j + k$ iterations, $0 \leq k < 2^j$, the current scoring function is

$$s_m(x) = \sum_{l=0}^{2k-1} (m-l) \cdot \mathbb{I}\{x \in \mathscr{C}_{j+1,l}\} + \sum_{l=k}^{2^j-1} (m-k-l) \cdot \mathbb{I}\{x \in \mathscr{C}_{j,l}\}$$

and the cell $\mathscr{C}_{j,k}$ is split in order to form a refined version of the scoring function,

$$s_{m+1}(x) = \sum_{l=0}^{2k} (m-l) \cdot \mathbb{I}\{x \in \mathscr{C}_{j+1,l}\} + \sum_{l=k+1}^{2^j-1} (m-k-l) \cdot \mathbb{I}\{x \in \mathscr{C}_{j,l}\}$$

namely, with maximum (empirical) AUC. Therefore, it happens that this problem boils down to solve a cost-sensitive binary classification problem on the set $\mathscr{C}_{j,k}$, see Sect. 3.3 in [14]. Indeed, one may write the AUC increment as

$$\mathrm{AUC}(s_{m+1}) - \mathrm{AUC}(s_m) = \frac{1}{2} H(\mathscr{C}_{j,k}) G(\mathscr{C}_{j,k}) \times (1 - \Lambda(\mathscr{C}_{j+1,2k} \mid \mathscr{C}_{j,k})),$$

where

$$\Lambda(\mathscr{C}_{j+1,2k} \mid \mathscr{C}_{j,k}) \stackrel{def}{=} G(\mathscr{C}_{j,k} \setminus \mathscr{C}_{j+1,2k}) / G(\mathscr{C}_{j,k}) + H(\mathscr{C}_{j+1,2k}) / H(\mathscr{C}_{j,k}).$$

Setting $p = G(\mathscr{C}_{j,k}) / (H(\mathscr{C}_{j,k}) + G(\mathscr{C}_{j,k}))$, the crucial point of the TREERANK approach is that the quantity $2p(1-p)\Lambda(\mathscr{C}_{j+1,2k} \mid \mathscr{C}_{j,k})$ can be interpreted as the cost-sensitive error of a classifier on $\mathscr{C}_{j,k}$ predicting positive label on $\mathscr{C}_{j+1,2k}$ and negative label on $\mathscr{C}_{j,k} \setminus \mathscr{C}_{j+1,2k}$ with cost p (respectively, $1-p$) assigned to the error consisting in predicting label $+1$ given $Y = -1$ (resp., label -1 given $Y = +1$), balancing thus the two types of error. Hence, at each iteration of the ranking tree growing stage, the TREERANK algorithm calls a *cost-sensitive* binary classification algorithm, termed LEAFRANK, in order to solve a statistical version of the problem above (replacing the theoretical probabilities involved by their empirical counterparts) and split $\mathscr{C}_{j,k}$ into $\mathscr{C}_{j+1,2k}$ and $\mathscr{C}_{j+1,2k+1}$. As described at length in [14], one may use cost-sensitive versions of celebrated binary classification algorithms such as CART or SVM for instance as LEAFRANK procedure, the performance depending on their ability to capture the geometry of the level sets of the likelihood ratio $dG/dH(x)$. The growing stage, which can be interpreted as a statistical version of an adaptive

piecewise linear interpolation technique for approximating the optimal ROC curve, is generally followed by a pruning procedure, where children of a same parent node are recursively merged in order to produce a ranking subtree that maximizes an estimate of the AUC criterion, based on cross validation usually, see Sect. 4 in [14]. Under appropriate hypothese, consistency results and rate bounds for the TREERANK method (in the sup norm sense and for the AUC deficit both at the same time) are established in [13, 14]. A detailed experimental study can be found in [31].

As explained in Sect. 3.2, in the anomaly ranking context, the "negative distribution" is $H(dx) = U(dx)$ while $F(dx)$ plays the role of the "positive" distribution. Therefore, in the situation where LEAFRANK is chosen as a cost-sensitive version of the CART algorithm with axis parallel splits (see [28]), all the cells $\mathscr{C}_{j,k}$ can be expressed as union of hypercubes. The exact computation of the volume $U(\mathscr{C}_{j,k})$ is then elementarily feasible, as a function of the threshold values involved in the decision tree describing the split and of the volume of the parent node, as described below and illustrated by Fig. 3. We call this splitting rule UNSUPERVISED LEAFRANK. Suppose that the training data consist of i.i.d. observations $X_1, \ldots, X_n$, copies of the generic r.v. $X = (X^{(1)}, \ldots, X^{(d)})$ with probability distribution $F(dx)$ and taking its values in a compact feature space that can be expressed as union of hypercubes $\prod_{l=1}^{d} [a_{0,0}^{(l)},\ b_{0,0}^{(l)}] \subset \mathbb{R}^d$ with $-\infty < a_{0,0} < b_{0,0} < +\infty$. Denoting by $\hat{F}_n = (1/n)\sum_{i=1}^{n} \delta_{X_i}$ the empirical distribution of the training observations, the UNSUPERVISED LEAFRANK algorithm is implemented in three steps as follows.

UNSUPERVISED LEAFRANK

- **Input.** Maximal depth $D \geq 1$ of the tree depicting the splitting rule.
- **Initialization.** Start from the root node $\mathscr{C}_{0,0}$, suppose to be the union of a finite number of disjoint hypercubes $[a_{0,0}^{(1)}, b_{0,0}^{(1)}] \times \cdots \times [a_{0,0}^{(d)}, b_{0,0}^{(d)}]$. The volume of the root cell $\lambda_{0,0} = \lambda(\mathscr{C}_{0,0})$ is obtained by summing the volumes $\prod_{i=1}^{d} (b_{0,0}^{(i)} - a_{0,0}^{(i)})$ of the hypercubes forming it.
- **Iterations** For $j = 1, \ldots, D$ and for $k = 0, \ldots, 2^j - 1$:

  1. In a greedy fashion, compute

$$\mathscr{C}_{j+1,2k} = \arg\max_{\mathscr{C}} \left\{ \frac{\lambda(\mathscr{C})}{\lambda(\mathscr{C}_{j,k})} - \frac{\hat{F}_n(\mathscr{C})}{\hat{F}_n(\mathscr{C}_{j,k})} \right\},$$

  over subsets $\mathscr{C}$ of $\mathscr{C}_{j,k} = \prod_{i=1}^{d} [a_{j,k}^{(i)}, b_{j,k}^{(i)}]$ of the form $\mathscr{C}_{j,k} \cap \{X^{(l)} \leq s_l\}$ or $\mathscr{C}_{j,k} \cap \{X^{(l)} > s_l\}$ for pairs "split variable—split value" $(X^{(l)}, s_l)$ with $l \in \{1, \ldots, d\}$ and $s_l \in \{X_i^{(l)} :\ i = 1, \ldots, n\}$. Stop if the maximum is not strictly positive and set $\mathscr{C}_{j+1,2k} = \mathscr{C}_{j,k}$, $\mathscr{C}_{j+1,2k+1} = \emptyset$.

2. If $\mathscr{C}_{j+1,2k} = \mathscr{C}_{j,k} \cap \{X^{(l)} \leq s_l\}$, then set

$$a_{j,k}^{(i)} = a_{j+1,2k}^{(i)} = a_{j+1,2k+1}^{(i)} \text{and} b_{j,k}^{(i)} = b_{j+1,2k}^{(i)} = b_{j+1,2k+1}^{(i)} \text{for} i \neq l$$
$$s_l = a_{j+1,2k+1}^{(l)} = b_{j+1,2k}^{(l)} \text{and} b_{j,k}^{(l)} = b_{j+1,2k+1}^{(l)} \text{and} a_{j,k}^{(l)} = a_{j+1,2k}^{(l)}$$

And if $\mathscr{C}_{j+1,2k} = \mathscr{C}_{j,k} \cap \{X^{(l)} > s_l\}$, then set

$$a_{j,k}^{(i)} = a_{j+1,2k}^{(i)} = a_{j+1,2k+1}^{(i)} \text{and} b_{j,k}^{(i)} = b_{j+1,2k}^{(i)} = b_{j+1,2k+1}^{(i)} \text{for} i \neq l$$
$$s_l = a_{j+1,2k}^{(l)} = b_{j+1,2k+1}^{(l)} \text{and} b_{j,k}^{(l)} = b_{j+1,2k}^{(l)} \text{and} a_{j,k}^{(l)} = a_{j+1,2k+1}^{(l)}$$

3. If $X^{(l)}$ denotes the split variable, compute $\lambda_{j+1,2k} \stackrel{def}{=} \lambda(\mathscr{C}_{j+1,2k})$ by summing over the disjoint hypercubes forming $\mathscr{C}_{j,k}$ the quantities

$$\lambda_{j,k} \times \frac{b_{j+1,2k}^{(l)} - a_{j+1,2k}^{(l)}}{b_{j,k}^{(l)} - a_{j,k}^{(l)}}.$$

Then, set $\lambda_{j+1,2k+1} \stackrel{def}{=} \lambda(\mathscr{C}_{j+1,2k+1}) = \lambda_{j,k} - \lambda_{j+1,2k}$.

- **Output** the left node $\mathscr{L} = \cup_k \mathscr{C}_{D,2k}$ and the right node $\mathscr{R} = C_{0,0} \setminus \mathscr{L}$. Compute $\lambda(\mathscr{L}) = \sum_k \lambda_{2^D,2k}$ as well as $\lambda(\mathscr{R}) = \lambda_{0,0} - \lambda(\mathscr{L})$.

Hence, only empirical counterparts of the quantities $F(\mathscr{C})$ for subset $\mathscr{C} \subset [0,1]^d$ candidates, $\hat{F}_n(\mathscr{C}) = (1/n) \sum_{i=1}^n \mathbb{I}\{X \in \mathscr{C}\}$, are required to estimate the cost-sensitive classification error and implement the splitting stage (AUC maximization). Hence, this approach does not require to sample any additional data, in contrast to that proposed in Sect. 3.3. This is a key advantage in practice, in contrast to "simulation-based" approaches: for high values of the dimension d, data are expected to lie very sparsely in $[0,1]^d$ and can be then very easily separated from those obtained by sampling a "reasonable" number of uniform observations, leading bipartite ranking algorithms to overfit. Similarly to the supervised case, the UNSUPERVISED TREERANK algorithm corresponds to a statistical version of an adaptive piecewise linear interpolation scheme of the optimal MV curve, see [13]. The growing stage is implemented by calling recursively the UNSUPERVISED LEAFRANK procedure, as follows. Again, rather than translating and rescaling the input vector, we assume that the compact feature space is $[0,1]^d$.

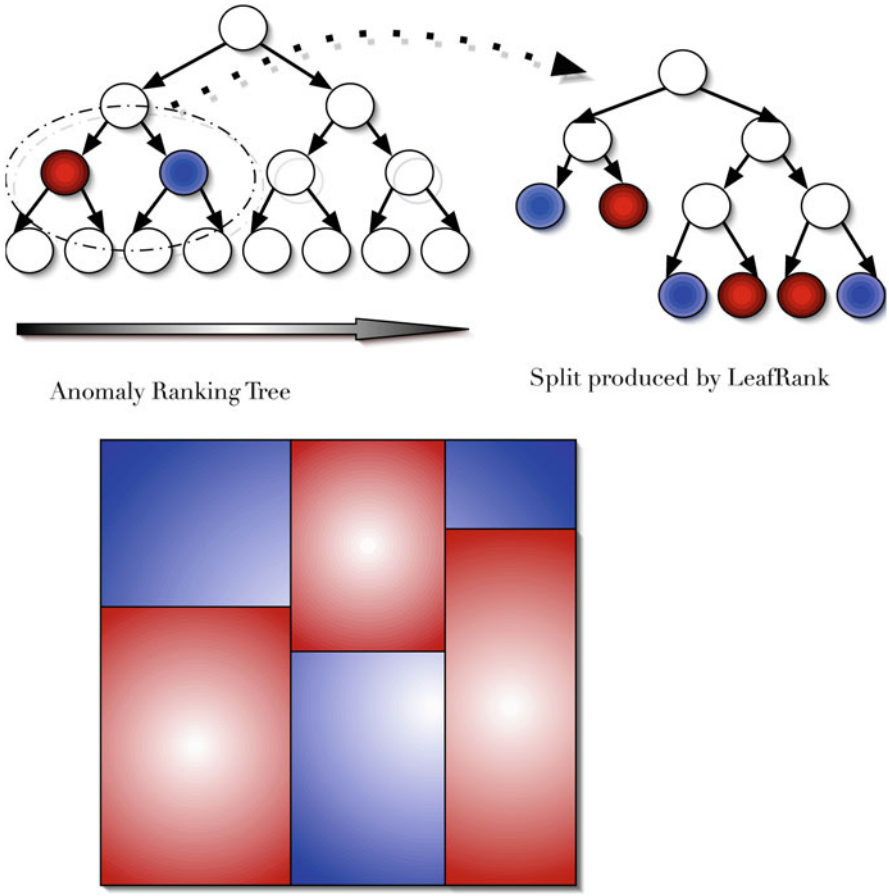

Fig. 3 Anomaly ranking tree produced by the UNSUPERVISED TREERANK algorithm. Ideally, the cells at the *bottom right* form the tail region, while those at the *bottom left* correspond to the modes of the underlying distribution

UNSUPERVISED TREERANK—GROWING STAGE

- **Input.** Maximal depth $D \geq 1$ of the anomaly ranking tree, maximal depth D_0 related to the implementation of UNSUPERVISED LEAFRANK. Training data $\mathscr{D}_n = \{X_1, \ldots, X_n\}$
- **Initialization.** Start from the root node $\mathscr{C}_{0,0} = [0,1]^d$. Set $\lambda_{0,0} = \lambda(\mathscr{C}_{0,0}) = 1$.
- **Iterations** For $j = 1, \ldots, D$ and for $k = 0, \ldots, 2^j - 1$:

1. Considering the input space $\mathscr{C}_{j,k}$, its volume $\lambda_{j,k}$ and the sample composed of observations of the original sample $\mathscr{D}_n$, run the UNSUPERVISED LEAFRANK procedure to split the cell $\mathscr{C}_{j,k}$, yielding the left and right siblings $\mathscr{L}$ and $\mathscr{R}$ with volumes respectively given by $\lambda(\mathscr{L})$ and $\lambda(\mathscr{R})$.
2. Set $\mathscr{C}_{j+1,2k} = \mathscr{L}$, $\lambda_{j+1,2k} = \lambda(\mathscr{L})$, $\mathscr{C}_{j+1,2k+1} = \mathscr{R}$ and $\lambda_{j+1,2k+1} = \lambda(\mathscr{R})$

- **Output** the anomaly ranking tree $\mathscr{T}_D = \{\mathscr{C}_{j,k} = \ 0 \leq j \leq D,\ 0 \leq k < 2^j\}$.

4.3 *Pruning the Anomaly Ranking Tree: Model Selection*

Recursive partitioning techniques fragment the data, splitting rules becoming more and more unstable as the depth increases. The second stage of the UNSUPERVISED TREERANK algorithm consists in choosing an anomaly ranking subtree of that produced by the growing stage. In general, the growing stage is followed by a pruning procedure, where children of a same parent node are recursively merged, until the root of the original tree is reached. The goal pursued here is to select a ranking subtree that minimizes a certain estimator of the area under the MV curve criterion, the area under the empirical MV curve being of course a much too optimistic estimate.

Like for the CART algorithm in classification/regression or for the original TREERANK method (see Sect. 4 in [14]), a natural approach consists in using cross validation, as follows. The idea is to penalize the area under the empirical MV curve of a scoring function related to any anomaly ranking subtree candidate $\mathscr{T} \subset \mathscr{T}_D$, denoted by $\widehat{\mathrm{AMV}}(\mathscr{T})$, by the number of terminal leaves, denoted by $|\mathscr{T}|$, in a linear fashion, yielding the *complexity-penalized empirical area under the* MV *curve* criterion:

$$\widehat{\mathrm{CPAMV}}_{\nu}(\mathscr{T}) \stackrel{def}{=} \widehat{\mathrm{AMV}}(\mathscr{T}) + \nu \cdot |\mathscr{T}|, \tag{10}$$

where $\nu > 0$ is a parameter tuning the trade-off between "goodness-of-fit" and "complexity" (as measured by $|\mathscr{T}|$). For each $\nu > 0$, define

$$\mathscr{T}_{\nu}^{*} = \operatorname*{arg\,min}_{\mathscr{T} \subset \mathscr{T}_D} \widehat{\mathrm{CPAMV}}_{\nu}(\mathscr{T}).$$

In practice, the $\mathscr{T}_{\nu}^{*}$'s are determined using *weakest link pruning*, i.e., by successively merging leaves of a same parent so as to produce the smallest increase in $\widehat{\mathrm{AMV}}(\mathscr{T})$ in a bottom-up manner, producing a decreasing sequence of anomaly ranking subtrees $\mathscr{T}_D = \mathscr{T}^{(2^D)} \subset \mathscr{T}^{(2^D-1)} \supset \cdots \supset \mathscr{T}^{(1)}$, denoting by $\mathscr{T}^{(1)}$ the anomaly ranking tree reduced to the root. The "best" model $\mathscr{T}_{\nu*}^{*}$ is then picked among this finite collection by cross validation.

Interpretation From a practical angle, a crucial advantage of the approach describes above lies in the interpretability of the anomaly ranking rules produced. In contrast to alternative techniques, they can be summarized by means of a left-to-right oriented binary tree graphic: observations are all the more considered as abnormal as they are located in terminal leaves at the right of the *anomaly ranking tree*. An arrow at the bottom of the tree indicates the direction in which the density decreases. Each splitting rule possibly involves the combination of elementary threshold rules of the type "$X^{(k)} > \kappa$" or "$X^{(k)} \leq \kappa$" with $\kappa \in \mathbb{R}$ in a hierarchical manner. In addition, it is also possible to rank the $X^{(k)}$'s depending on their *relative importance*, measured through the empirical *volume under the* MV *curve* decrease induced by splits involving $X^{(k)}$ as *split variable*, just like in the supervised setup, see Sect. 5.1 in [14] for further details. This permits to identify the variables which have most relevance to detect anomalies.

5 Numerical Experiments

We now illustrate the advantages of the UNSUPERVISED TREERANK algorithm by means of numerical experiments, based on unlabeled synthetic/real datasets.

Mixture of 2-d Gaussian Distributions We first display results based on a two-dimensional toy example to compare the performance of the algorithm described in the previous section with that of a simulation-based approach, namely the TREERANK algorithm fed with the original (positive) sample plus a (negative) sample of i.i.d. data uniformly distributed over a square containing the data to be ranked, as proposed in Sect. 3.3. For comparison purpose, the values of the complexity tuning parameters are the same in both cases (maximal depth in LEAFRANK/TREERANK), the positive and negative samples have the same size: $n = m = 2500$.

Figure 4 respectively depicts the scoring function learnt with UNSUPERVISED TREERANK and Fig. 5 the related MV curves computed via fivefold cross validation. The latter must be compared with its analogue, Fig. 6, corresponding to the performance of the scoring function learnt by means of the simulated dataset. As the results summarized in Table 1 show, the UNSUPERVISED TREERANK algorithm outperforms its supervised version based on a simulated dataset, even in a small-dimension setting.

Real Dataset We have also implemented the UNSUPERVISED TREERANK algorithm to rank the input data of the breast cancer database ($n = 569$), lying in a feature space of dimension $d = 32$, see https://archive.ics.uci.edu/ml/datasets/. The MV curves produced by UNSUPERVISED TREERANK via fivefold cross validation are presented in Fig. 7: the mean of the area under the MV curve criterion is 0.037, while its standard deviation is 0.005. In contrast, in such a high dimensional space, simulation-based approaches completely fail: even if one increases drastically the

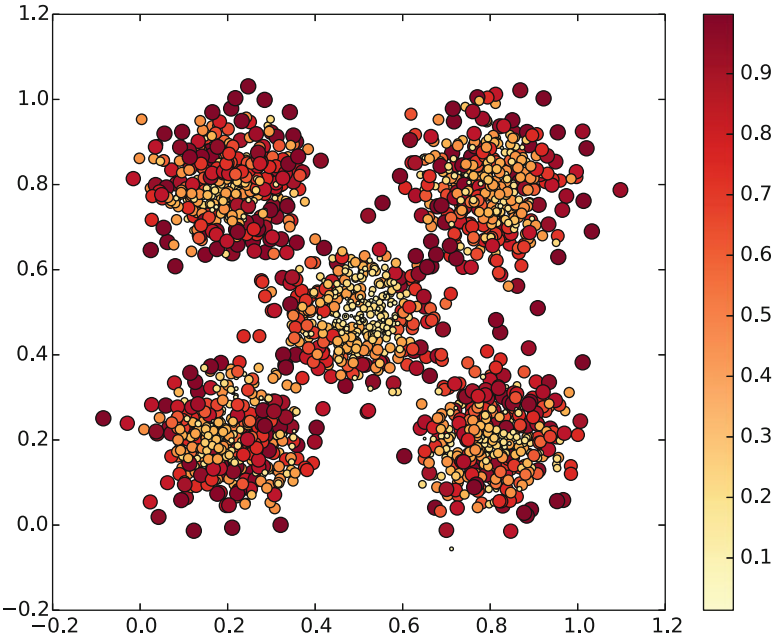

Fig. 4 Synthetic 2-d dataset: scoring function learnt by UNSUPERVISED TREERANK with 7 as ranking tree depth and 7 as LEAFRANK depth

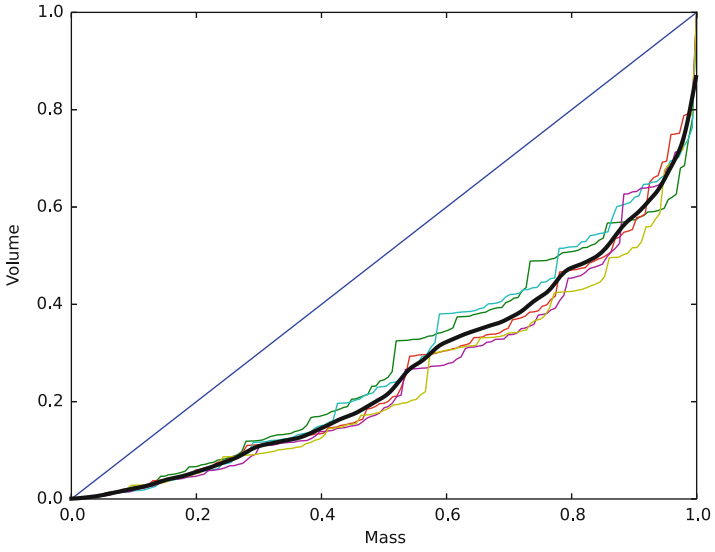

Fig. 5 Synthetic 2-d dataset: MV curves related to the scoring functions obtained via UNSUPERVISED TREERANK through fivefold cross validation

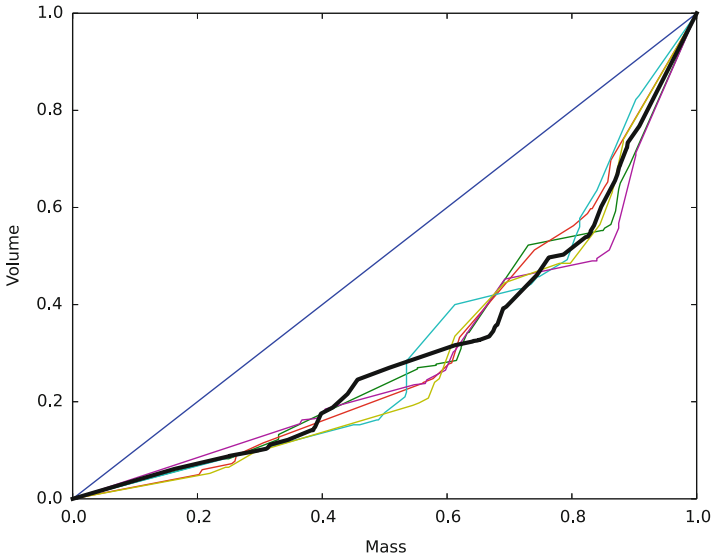

Fig. 6 Synthetic 2-d dataset: MV curves related to the scoring functions obtained via TREERANK and simulation through fivefold cross validation

Table 1 Area under the MV curve estimates: mean and standard error via fivefold cross validation

Method	Mean	Standard error
UNSUPERVISED TREERANK	0.265	0.016
TREERANK and simulation	0.306	0.0075

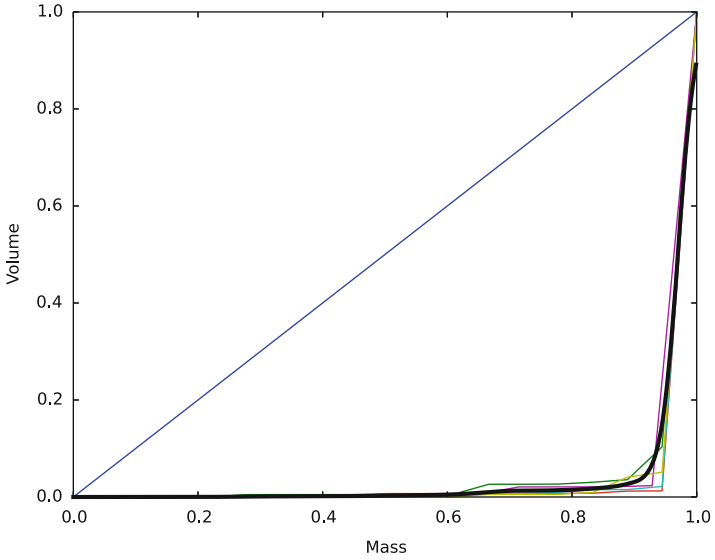

Fig. 7 Real dataset: MV curves related to the scoring functions obtained via UNSUPERVISED TREERANK in a 32-dimensional feature space through fivefold cross validation

number of simulated "negative" instances, supervised technique generally overfit and yield a null area under the MV curve for each replication of the cross validation scheme.

6 Conclusion

In this chapter, we have presented a novel algorithm for unsupervised anomaly ranking, cast as minimization of the Mass-Volume curve criterion, producing scoring rules described by oriented binary trees, referred to as *anomaly ranking trees*. It is called UNSUPERVISED TREERANK. Although it relies on the connection between anomaly ranking and supervised bipartite ranking, its major advantage consists in the fact that it does not involve any sampling stage: since the cells output by this recursive partitioning method can be expressed as unions of hypercubes, the exact computation of their volume is straightforward (avoiding the use of any Monte Carlo sampling scheme). In contrast to alternative techniques, this approach is thus quite tailored to large-scale and high dimensional (unlabeled) data, as empirically confirmed by experimental results. Though it is very promising, there is still room for improvement. Just like for supervised bipartite ranking (see [24]), combining aggregation and randomization, the main ingredients of *ensemble learning* techniques, could lead to dramatically improve stability and accuracy of anomaly ranking tree models both at the same time, while maintaining most of their advantages (e.g., scalability, interpretability). This will be the subject of future research.

References

1. Provost, F., Fawcett, T.: Adaptive fraud detection. Data Min. Knowl. Disc. **1**, 291–316 (1997)
2. Martin, R., Gorinevsky, D., Matthews, B.: Aircraft anomaly detection using performance models trained on fleet data. In: Proceedings of the 2012 Conference on Intelligent Data Understanding (2012)
3. Viswanathan, K., Choudur, L., Talwar, V., Wang, C., Macdonald, G., Satterfield, W.: Ranking anomalies in data centers. In: James, R.D. (ed.) Network Operations and System Management, pp. 79–87. IEEE, New York (2012)
4. Polonik, W.: Minimum volume sets and generalized quantile processes. Stoch. Process. Their Appl. **69**(1), 1–24 (1997)
5. Scott, C., Nowak, R.: Learning Minimum Volume Sets. J. Mach. Learn. Res. **7**, 665–704 (2006)
6. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A., Williamson, R.: Estimating the support of a high-dimensional distribution. Neural Comput. **13**(7), 1443–1471 (2001)
7. Steinwart, I., Hush, D., Scovel, C.: A classification framework for anomaly detection. J. Mach. Learn. Res. **6**, 211–232 (2005)
8. Vert, R., Vert, J.-P.: Consistency and convergence rates of one-class svms and related algorithms. J. Mach. Learn. Res. **7**, 817–854 (2006)
9. Park, C., Huang, J.Z., Ding, Y.: A computable plug-in estimator of minimum volume sets for novelty detection. Oper. Res. **58**(5), 1469–1480 (2010)

10. Clémençon, S., Jakubowicz, J.: Scoring anomalies: a M-estimation formulation. In: Proceedings of AISTATS, JMLR W&CP, vol. 31 (2013)
11. Han, J., Jin, W., Tung, A., Wang, W.: Ranking Outliers Using Symmetric Neighborhood Relationship. Lecture Notes in Computer Science, vol. 3918, pp. 148–188. Springer, Berlin (2006)
12. Clémençon, S., Robbiano, S.: Anomaly ranking as supervised bipartite ranking. In: Jebara, T., Xing, E.P. (eds.) Proceedings of the 31st International Conference on Machine Learning (ICML-14), pp. 343–351 (2014)
13. Clémençon, S., Vayatis, N.: Tree-based ranking methods. IEEE Trans. Inf. Theory **55**(9), 4316–4336 (2009)
14. Clémençon, S., Depecker, M., Vayatis, N.: Adaptive partitioning schemes for bipartite ranking. Mach. Learn. **43**(1), 31–69 (2011)
15. Einmahl, J.H.J., Mason, D.M.: Generalized quantile process. Ann. Stat. **20**, 1062–1078 (1992)
16. Scott, C., Nowak, R.: Learning minimum volume sets. J. Mach. Learn. Res. **7**, 665–704 (2006)
17. Duchi, J., Mackey, L., Jordan, M.: On the consistency of ranking algorithms. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10) (2010)
18. Clémençon, S., Lugosi, G., Vayatis, N.: Ranking and empirical risk minimization of U-statistics. Ann. Stat. **36**(2), 844–874 (2008)
19. Agarwal, S., Graepel, T., Herbrich, R., Har-Peled, S., Roth, D.: Generalization bounds for the area under the ROC curve. J. Mach. Learn. Res. **6**, 393–425 (2005)
20. Freund, Y., Iyer, R., Schapire, R., Singer, Y.: An efficient boosting algorithm for combining preferences. J. Mach. Learn. Res. **4**, 933–969 (2003)
21. Rakotomamonjy, A.: Optimizing area under Roc curve with SVMs. In: Proceedings of the First Workshop on ROC Analysis in AI (2004)
22. Pahikkala, T., Tsivtsivadze, E., Airola, A., Boberg, J., Salakoski, T.: Learning to rank with pairwise regularized least-squares. In: Proceedings of SIGIR, pp. 27–33 (2007)
23. Fawcett, T.: An introduction to ROC analysis. Lett. Pattern Recogn. **27**(8), 861–874 (2006)
24. Clémençon, S., Depecker, M., Vayatis, N.: Ranking forests. J. Mach. Learn. Res. **14**, 39–73 (2013)
25. Herbrich, R., Graepel, T., Obermayer, K.: Large margin rank boundaries for ordinal regression. In: Advances in Large Margin Classifiers, pp. 115–132. MIT Press, Cambridge (2000)
26. Friedman, J., Hastie, T., Tibshirani, R.: The Elements of Statistical Learning. Springer, New York (2009)
27. Scott, C., Davenport, M.: Regression level set estimation via cost-sensitive classification. IEEE Trans. Signal Process. **55**(6), 2752–2757 (2007)
28. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees. Wadsworth and Brooks, Belmont (1984)
29. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, Los Altos (1993)
30. Rokach, L., Maimon, O.: Data-Mining with Decision Trees: Theory and Applications, 2nd edn. Series in Machine Perception and Artificial Intelligence. World Scientific, Singapore (2014)
31. Clémençon, S., Depecker, M., Vayatis, N.: An empirical comparison of learning algorithms for nonparametric scoring: the treerank algorithm and other methods. Pattern Anal. Appl. **16**(4), 475–496 (2013)