Anomaly Ranking as Supervised Bipartite Ranking

Stéphan Clémençon STEPHAN.CLEMENCON@TELECOM-PARISTECH.FR
LTCI UMR Telecom ParisTech/CNRS No. 5141, 46 rue Barrault, 75634 Paris Cedex, France

Sylvain Robbiano SYLVAIN.ROBBIANO@GMAIL.COM
LTCI UMR Telecom ParisTech/CNRS No. 5141, 46 rue Barrault, 75634 Paris Cedex, France
CIMFAV-Facultad de Ingeniera, Universidad de Valparaso, Valparaso, Chile

Abstract

The Mass Volume (MV) curve is a visual tool to evaluate the performance of a scoring function with regard to its capacity to rank data in the same order as the underlying density function. *Anomaly ranking* refers to the unsupervised learning task which consists in building a scoring function, based on unlabeled data, with a MV curve as low as possible at any point. In this paper, it is proved that, in the case where the data generating probability distribution has compact support, anomaly ranking is equivalent to (supervised) bipartite ranking, where the goal is to discriminate between the underlying probability distribution and the uniform distribution with same support. In this situation, the MV curve can be then seen as a simple transform of the corresponding ROC curve. Exploiting this view, we then show how to use bipartite ranking algorithms, possibly combined with random sampling, to solve the MV curve minimization problem. Numerical experiments based on a variety of bipartite ranking algorithms well-documented in the literature are displayed in order to illustrate the relevance of our approach.

1. Introduction

Motivated by a great range of applications such as the design of search engines in information retrieval, credit-risk screening in finance or supervised anomaly detection in signal processing, the problem of learning how to rank data with ordinal labels has been the subject of a good deal of attention in machine-learning these last few years, see (Duchi et al., 2010; Clémençon et al., 2008; Agarwal et al., 2005) among others. A wide variety of criteria have been considered so as to cast the task of ranking observations in an order as close as possible to that induced by the ordinal output variable as a M-estimation problem, including the (area under the) receiver operator characteristic curve (ROC curve in short) and the precision-recall curve in the bipartite situation (*i.e.* when the output variable is binary), the normalized discounted cumulative gain criterion or the ROC manifold in the general multipartite framework. Many practical ranking algorithms, supported by sound theoretical results extending the probabilistic theory of pattern recognition, are now documented in the literature, see (Freund et al., 2003; Clémençon & Vayatis, 2009; Rakotomamonjy, 2004; Pahikkala et al., 2007) for instance. However, in many applications, which can be referred to as *unsupervised anomaly/novelty detection* and comprise the monitoring of complex systems such as the functioning of aircraft engines, system management in data centers (*cf* (Viswanathan et al., 2012)), network intrusion surveillance or fraud detection, it would be desirable as well to rank multivariate data, so that top ranked observations should be ideally the likeliest "outliers", in absence of any output variable indicating the degree of "abnormality". Throughout the article, this problem shall be termed *anomaly ranking* or *anomaly scoring*, insofar as the most natural way to define a preorder on a general feature space $\mathcal{X} \subset \mathbb{R}^{\mathsf{d}}$, with $\mathsf{d} \geq 1$, is to transport the natural order on the real half-line by means of a (measurable) scoring function $\mathsf{s} : \mathcal{X} \to \mathbb{R}_+$. The ideal way of ordering all the elements of the feature space naturally corresponds to the reverse of the order induced by the (generally unknown) underlying density function. Because density estimation should be avoided in a high dimensional setup, due to the curse of dimensionality phenomenon, a perfor-

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

mance criterion of functional nature, just like the ROC curve in the supervised framework, has been recently introduced in (Clémençon & Jakubowicz, 2013), which permits to evaluate the accuracy of any ranking rule, as regards the objective pursued. It has been termed the *Mass Volume curve* (MV curve) and the lower the MV curve of a scoring function, the more accurate the ranking it induces. But, in contrast to the supervised framework, no algorithm is readily available to build a scoring function with a nearly optimal MV curve from (unlabeled) training data. It should also be recalled that *minimum volume set* estimation techniques (Scott & Nowak, 2006), originally designed to solve unsupervised anomaly detection problems, are not appropriate for anomaly ranking, cast as MV curve minimization, since they consist in recovering a single density level set, corresponding thus to single point of the optimal MV curve (with the target mass level as abcissa).

It is the major purpose of this paper to highlight the connection between (supervised) bipartite ranking and anomaly ranking. Indeed, it is shown in the present article that, in the case where the underlying probability distribution $F(dx)$ has compact support, included in $[0,1]^d$ say, the MV curve of any scoring function $s(x)$ and its ROC curve with regard to the bipartite ranking problem, where the "positive distribution" is the probability measure $F(dx)$ which the observations are drawn according to and the "negative distribution" is uniform on $[0,1]^d$, are symmetrical w.r.t. the first diagonal. Hence, minimization of the MV curve boils down to maximizing the corresponding ROC curve, which task can be achieved by various algorithms based on two independent data samples, one drawn from $F(dx)$ and the other drawn from the uniform distribution. Building on this crucial observation, we first propose to "hijack" bipartite ranking algorithms by implementing them with the originally unlabeled sample as "negative sample" and a simulated "positive sample" made of i.i.d. data uniformly distributed on $[0,1]^d$. Incidentally, we point out that sampling data from a reference measure in order to reveal properties of the density under study by means of supervised techniques is well-known folklore in applied statistics, *generalized association rules* for instance are precisely based on this approach (see Chapter 14 in (Friedman et al., 2009)) and (Steinwart et al., 2005) proposed a method, involving the simulation of uniformly distributed data too, to turn (unsupervised) anomaly detection into a supervised binary classification problem. We next explain how to extend the TreeRank approach, originally introduced in (Clémençon & Vayatis, 2009) to the unsupervised framework. Beyond the fact that it may produce interpretable ranking rules, visualizable by means of an oriented tree graphic, in contrast to other bipartite ranking algorithms, its implementation in the unsupervised context does not require to sample any extra data uniformly distributed over $[0,1]^d$. Numerical experiments have been also carried out in order to illustrate the performance of various bipartite ranking algorithms for anomaly ranking.

The rest of the article is structured as follows. In section 2, notations are first set out and key notions of the *anomaly ranking* problem are recalled, together with basic concepts of *bipartite ranking* and ROC analysis. Section 3 explains at length the connection between the supervised and unsupervised ranking problems in the compact support case, while section 4 shows how to extend the use of certain bipartite ranking algorithms to anomaly ranking from a practical perspective. Finally, numerical results based on synthetic/real datasets are displayed in section 5 for illustration purpose.

2. Background and Preliminaries

Here we essentially describe the issue of anomaly ranking and briefly recall the related key concepts of MV curve analysis. We also set out the notations that shall be needed throughout the paper.

2.1. The Statistical Framework

In the *anomaly ranking* problem, the goal pursued is to learn how to order observations by degree of "abnormality", on the basis of a training sample $X_1, \ldots, X_n$ made of i.i.d. copies of a random variable X, taking its values in a (possibly high-dimensional) space $\mathcal{X} \subset \mathbb{R}^d$ with $d \geq 1$ and distributed according to a continuous probability measure $F(dx) = f(x)dx$. The preorder on $\mathcal{X}$ is defined by means of a (measurable) scoring function $s : \mathcal{X} \to \mathbb{R}_+$: $\forall (x, x') \in \mathcal{X}^2$, $x \preceq_s x' \Leftrightarrow s(x) \leq s(x')$. We denote by $\mathcal{S}$ the set of all scoring functions on $\mathcal{X}$ integrable w.r.t. Lebesgue measure on $\mathcal{X}$ (see the next subsection for the explanation of the integrability constraint). Given the nature of the problem, the optimal ranking is naturally that determined by the density function $f(x)$. The set $\mathcal{S}^* \subset \mathcal{S}$ of optimal scoring functions is made of (λ-integrable) nonnegative strictly increasing transforms of the density function.

We point out that the nature of the problem considered is very different from that of (nonparametric) *density estimation*: there is no need to estimate the local values taken by the density function, only the preorder on $\mathcal{X}$ it induces is of interest here. We also empha-

size that in many applications, the very purpose of unsupervised anomaly detection is not to assign a label "normal" vs. "abnormal" to any new observation, compared to the vast majority of the data previously observed (*i.e.* the training data), but to rank any new set of observations (*i.e.* test data) by degree of abnormality. The form of the output, an ordered list namely, may greatly facilitates the work of human operators. For instance, in the context of Distributed Fleet Monitoring (DFM) for Flight Operational Quality Assurance (FOQA) programs, an anomaly scoring function (taking flight data and aircraft features as input variables in particular) could permit to set priorities and help optimize the work of FOQA analysts who do not have time to look at the data for hundreds of thousands of flights: depending on operational constraints, the 10 most abnormal flights will be examined first, then the next 10 flights, *etc.*. Finally, we underline that the framework we develop in this paper is fully nonparametric. In particular, no parametric assumption on the tail behavior of the underlying multivariate distribution, permitting to rank new observations according the corresponding p-values, is made.

The following assumptions shall be involved in the subsequent analysis.

$\mathbf{H_1}$ The random variable $f(X)$ is continuous.

$\mathbf{H_2}$ The density f is bounded: $\sup_{x\in\mathcal{X}} f(x) < +\infty$.

Here we denote by λ the Lebesgue measure on $\mathcal{X}$, by $\mathbb{I}\{\mathcal{E}\}$ the indicator function of any event $\mathcal{E}$. The generalized inverse function of any cdf $K(t)$ on $\mathbb{R}$ is denoted by $K^{-1}(u) = \inf\{t \in \mathbb{R} : K(t) \geq u\}$.

2.2. Mass Volume Curves

In (Clémençon & Jakubowicz, 2013), a natural way of measuring the ranking performance of a given scoring function $s \in \mathcal{S}$ in the unsupervised setting has been introduced (see Definition 2 therein). It consists of plotting its Mass Volume (MV) curve, namely the Probability Measure plot:

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

With $\Omega_{s,t} = \{x \in \mathcal{X} :\ s(x) \geq t\}$ for any $t \geq 0$, the MV curve is the parametrized curve $t > 0 \mapsto (F(\Omega_{s,t}),\ \lambda(\Omega_{s,t}))$. 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 $t > 0$. Connecting points corresponding to possible jumps of the parametric curve, the curve can be seen as the plot of a continuous mapping $MV_s : \alpha \in (0,1) \mapsto MV_s(\alpha)$, starting at $(0,0)$. Denoting by $F_s(t)$ the cdf of the r.v. $s(X)$, in the case where $F_s \circ F_s^{-1}(\alpha) = \alpha$, we have:

$$MV_s(\alpha) = \lambda\left(\{x \in \mathcal{X} :\ s(x) \geq F_s^{-1}(1-\alpha)\}\right). \quad (2)$$

Let $\alpha \in (0,1)$. Under the Assumptions $\mathbf{H_1} - \mathbf{H_2}$, the set $\Omega_\alpha^* = \{x \in \mathcal{X} :\ f(x) \geq F_f^{-1}(1-\alpha)\}$ is the unique solution of the *minimum volume set* problem

$$\min_{\Omega\in\mathcal{B}(X)} \lambda(\Omega) \text{ subject to } F(\Omega) \geq \alpha, \quad (3)$$

where $\mathcal{B}(\mathcal{X})$ denotes the ensemble made of all borelian subsets of $\mathcal{X}$. 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 (Einmahl & Mason, 1992; Polonik, 1997) for an account of minimum volume set theory and to (Vert & Vert, 2006; Scott & Nowak, 2006) for related statistical learning results. As stated in Proposition 3 of (Clémençon & Jakubowicz, 2013), this implies that the MV curve of the scoring function $f(x)$, which plots the (minimum) volume $\lambda(\Omega_\alpha^*)$ against the mass $F(\Omega_\alpha^*) = \alpha$ and which shall be denoted by MV^*, is dominated by any other MV curve everywhere:

$$\forall s \in \mathcal{S},\ \forall \alpha \in (0,1),\ \ MV^*(\alpha) \leq MV_s(\alpha). \quad (4)$$

It is noteworthy that MV^* is a convex function and MV curves are invariant under strictly increasing transforms. A list of properties of MV curves is given in Proposition 5 in (Clémençon & Jakubowicz, 2013). A typical MV curve is depicted in Fig. 5.1. When the distribution $F(dx)$ is much concentrated around its modes and exhibits a light tail behavior, the optimal MV curve increases very slowly and rises near 1. Of course, MV curves are generally unknown in practice, just like the distribution $F(dx)$, and must be estimated by their empirical counterparts, see Theorem 8 in (Clémençon & Jakubowicz, 2013) for more details on statistical estimation of MV curves.

A partial order on $\mathcal{S}$. The concept of MV curve induces a partial order on the set of scoring functions $\mathcal{S}$. Let $(s_1, s_2) \in \mathcal{S}^2$, we will say that the scoring function s_1 is more accurate than s_2 if and only if its MV curve is below of s_2 everywhere, *i.e.* $\forall \alpha \in (0,1)$, $MV_{s_1}(\alpha) \leq MV_{s_2}(\alpha)$: for any fixed mass level, s_1 defines a subset of smaller volume. Equipped with this functional performance criterion, the optimal scoring functions $s \in \mathcal{S}$ are the elements of $\mathcal{S}^* = \{T \circ f :\ T : Im f(X) \to \mathbb{R}_+ \text{ strictly increasing}\}$, where $Im f(X)$ denotes the image of the r.v. $f(X)$. The closer the MV curve of a scoring function candidate

to MV^*, the more accurate the ranking it defines. Of course, there a many ways of quantifying closeness in the MV space. One could naturally consider L_p distances, $1 \leq p \leq +\infty$, as in (Clémençon & Jakubowicz, 2013).

In this paper, focus is on the situation where the assumption below is fulfilled.

H_3 The probability distribution $F(dx)$ has compact support, equal to $[0,1]^d$ say.

In this case, the MV curve of any scoring function $s \in \mathcal{S}$ ends at $(1,1)$. Let $p \in [1,+\infty]$, the performance of any $s \in \mathcal{S}$ can be measured through the quantity

$$d_p(s,s^*) = \|MV_s - MV^*\|_p, \tag{5}$$

where $\|.\|_p$ denotes the L_p-norm on $[0,1]$ and $s^* \in \mathcal{S}^*$. The goal of *anomaly ranking* can be then stated in a quantitative manner. Based on a training dataset $\{X_1, \ldots, X_n\}$, the objective is to build a scoring function $s \in \mathcal{S}$ such that $d_p(s,s^*)$ is as small as possible with overwhelming probability. Notice finally that, since we have the decomposition $d_1(s,s^*) = \int_{\alpha=0}^{1} MV_s(\alpha)d\alpha - \int_{\alpha=0}^{1} MV^*(\alpha)d\alpha$ (see Eq. (4)), anomaly ranking reduces to minimization of the area under the MV curve in the case $p = 1$.

Anomaly ranking versus anomaly detection. Anomaly ranking is very different from (unsupervised) anomaly detection, cast as minimum volume set estimation. A mass level $\alpha \in (0,1)$ being preliminarily fixed, the latter consists in recovering from data a specific density level set Ω^*_α, while the former aims at building a scoring function $s(x)$ whose collection of level sets $\{\Omega_{s,t}\}_{t>0}$ nearly corresponds to that of the underlying density $f(x)$, $\{\Omega^*_\alpha\}_{\alpha\in(0,1)}$ (*i.e.* an increasing transform of $f(x)$). Hence, anomaly ranking should be viewed as a continuum of anomaly detection problems: finding the observations forming the top 1% the most abnormal, then those forming the top 2%, *etc.*

2.3. Ranking Bipartite and ROC Analysis

Consider now two probability distributions on the space $\mathcal{X}$, $G(dx)$ and $H(dx)$, absolutely continuous with respect to each other. The ROC curve of any scoring function $s(x)$ is defined as the *PP*-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 $s : \mathcal{X} \to \mathbb{R}_+$. 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 $ROC_s : \alpha \in (0,1) \mapsto ROC_s(\alpha)$. 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: $ROC_s(\alpha) = 1 - G_s \circ H_s^{-1}(1-\alpha)$. 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 (Fawcett, 2006) for an account of ROC analysis and its applications.

The notion of ROC curve defines a partial order on $\mathcal{S}$. A scoring function s_1 is more accurate than s_2 iff: $\forall \alpha \in (0,1)$, $ROC_{s_1}(\alpha) \geq ROC_{s_2}(\alpha)$. 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: $\forall (s,\alpha) \in \mathcal{S} \times (0,1)$, $ROC_s(\alpha) \leq ROC^*(\alpha)$. The set $\mathcal{S}^*_{H,G} = \{T \circ \phi,\ T : Im\phi(X) \to \mathbb{R}_+ \text{ strictly increasing}\}$ 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 $s \in \mathcal{S}$ can be measured by:

$$D_p(s,s^*) = \|ROC_s - ROC^*\|_p, \tag{6}$$

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

$$AUC(s) = \mathbb{P}\{s(X) < s(X')\} + \frac{1}{2}\mathbb{P}\{s(X) = s(X')\}, \tag{7}$$

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 $AUC(s)$ can be straightforwardly derived from (7), paving the way for the implementation of "empirical risk minimization" strategies, see (Clémençon et al., 2008).

The algorithms proposed to optimize the AUC criterion or surrogate performance measures are too numerous to be listed in an exhaustive manner. In the present article, due to space limitations, we restrict our attention to the extension of the following algorithms to the *anomaly ranking* problem: 1) the TreeRank method and its variants (see (Clémençon & Vayatis, 2009; Clémençon et al., 2011; 2013)), which relies on recursive AUC maximization, see subsection 4.2, 2) the RankBoost algorithm, which implements a boosting approach tailored for the ranking problem (see (Freund et al., 2003)), 3) the SVMrank algorithm originally designed for ordinal regression (see (Herbrich et al., 2000)) and 4) the RankRLS procedure proposed in (Pahikkala et al., 2007).

3. Anomaly Ranking: a Bipartite View

With the notations of subsection 2.3, 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 section 2 that, for any scoring function $s \in \mathcal{S}$, the curves MV_s and ROC_s are symmetrical with respect to the first diagnonal 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* $\mathbf{H_1}$, $\mathbf{H_2}$ *and* $\mathbf{H_3}$ *hold true. Let* $U(dx)$ *be the uniform distribution on* $[0,1]^d$. *For any* $(s,\alpha) \in \mathcal{S} \times (0,1)$, *we have:* $\mathrm{ROC}_s^{-1}(\alpha) = \mathrm{MV}_s(\alpha)$. *We also have* $\mathcal{S}^* = \mathcal{S}^*_{U,F}$, *and*

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

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

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

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 thus 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* 1 *and volume-axis coordinate* $\lambda(\mathrm{supp}F) \leq 1$, *the corresponding curve* ROC^* *exhibiting a plateau: it reaches* 1 *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* $X \in \mathcal{C}$, *where* $\mathcal{C}$ *denotes any compact set containing the observations* $X_1, \ldots, X_n$.

4. Extending Bipartite Methods

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.

4.1. Sampling

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

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

Except the choice of the algorithm $\mathcal{A}$ 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 (borelian) 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 section 14.2.4 in (Friedman et al., 2009), where it is used in particular to build *generalized association rules*. This idea is also exploited in (Steinwart et al., 2005) for anomaly detection, see also (Scott & Davenport, 2007). In this respect, it should be mentioned that a variety of techniques, including that proposed in (Schölkopf et al., 2001) 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 $\{x \in \mathcal{X} : \widehat{f}(x) > t_\alpha\}$ and one could consider using the decision function $\widehat{f}(x)$ as scoring function, one should keep in mind that there is no statistical guarantee that the ensembles $\{x \in \mathcal{X} : \widehat{f}(x) > t\}$ are good estimates of density level sets for $t \neq t_\alpha$. This explains the poor performance of such a "plug-in" approach in practice, as exhibited in section 5.

4.2. Unsupervised TreeRank

The TreeRank algorithm, a bipartite ranking technique optimizing the ROC curve in a recursive fashion, has been introduced in (Clémençon & Vayatis, 2009) and its properties have been investigated in detail in (Clémençon et al., 2011). Its output consists of an ordered partition of the feature space $\mathcal{X}$ (defining thus a ranking, for which elements of a same cell being viewed as ties). The ordered recursive partitioning process is described by a left-to-right oriented binary tree structure, referred to as *ranking tree*, with fixed maximum depth $J \geq 0$. At depth $j \leq J$, there are 2^j nodes, indexed by (j,k) with $0 \leq k < 2^j$. The root node depicts the entire space $\mathcal{C}_{0,0} = \mathcal{X}$ and each *internal node* (j,k) with $j < J$ and $k \in \{0, \ldots, 2^j - 1\}$ represents a subset $\mathcal{C}_{j,k} \subset \mathcal{X}$, whose left and right siblings respectively correspond to disjoint subsets $\mathcal{C}_{j+1,2k}$ and $\mathcal{C}_{j+1,2k+1}$ such that $\mathcal{C}_{j,k} = \mathcal{C}_{j+1,2k} \cup \mathcal{C}_{j+1,2k+1}$. At the root, one starts with a constant scoring function $s_1(x) = \mathbb{I}\{x \in \mathcal{C}_{0,0}\} \equiv 1$ and after $m = 2^j + k$ iterations, $0 \leq k < 2^j$, the current scoring function is $s_m(x) = \sum_{l=0}^{2k-1} (m-l) \cdot \mathbb{I}\{x \in \mathcal{C}_{j+1,l}\} + \sum_{l=k}^{2^j-1} (m-k-l) \cdot \mathbb{I}\{x \in \mathcal{C}_{j,l}\}$ and the cell $\mathcal{C}_{j,k}$ is split in order to form a refined version of the scoring function, $s_{m+1}(x) = \sum_{l=0}^{2k} (m-l) \cdot \mathbb{I}\{x \in \mathcal{C}_{j+1,l}\} + \sum_{l=k+1}^{2^j-1} (m-k-l) \cdot \mathbb{I}\{x \in \mathcal{C}_{j,l}\}$ namely, with maximum (empirical) AUC. Therefore, it happens that this problem boils down to solve a cost-sensitive binary classification problem on the set $\mathcal{C}_{j,k}$, see subsection 3.3 in (Clémençon et al., 2011) for further details. Indeed, one may write the AUC increment as $\mathrm{AUC}(s_{m+1}) - \mathrm{AUC}(s_m) = \frac{1}{2} H(\mathcal{C}_{j,k}) G(\mathcal{C}_{j,k}) \times (1 - \Lambda(\mathcal{C}_{j+1,2k} \mid \mathcal{C}_{j,k}))$, where $\Lambda(\mathcal{C}_{j+1,2k} \mid \mathcal{C}_{j,k}) \stackrel{\text{def}}{=} G(\mathcal{C}_{j,k} \setminus \mathcal{C}_{j+1,2k}) / G(\mathcal{C}_{j,k}) + H(\mathcal{C}_{j+1,2k}) / H(\mathcal{C}_{j,k})$. Setting $p = G(\mathcal{C}_{j,k}) / (H(\mathcal{C}_{j,k}) + G(\mathcal{C}_{j,k}))$, the crucial point of the TreeRank approach is that the quantity $2p(1-p)\Lambda(\mathcal{C}_{j+1,2k} \mid \mathcal{C}_{j,k})$ can be interpreted as the cost-sensitive error of a classifier on $\mathcal{C}_{j,k}$ predicting positive label on $\mathcal{C}_{j+1,2k}$ and negative label on $\mathcal{C}_{j,k} \setminus \mathcal{C}_{j+1,2k}$ 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 $\mathcal{C}_{j,k}$ into $\mathcal{C}_{j+1,2k}$ and $\mathcal{C}_{j+1,2k+1}$. As described at length in (Clémençon et al., 2011), 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 $dG/dH(x)$. In general, the growing stage is 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 (*cf* section 4 in (Clémençon et al., 2011)). Under adequate assumptions, consistency results and rate bounds for the TreeRank approach (in the sup norm sense and for the AUC deficit both at the same time) are established in (Clémençon & Vayatis, 2009) and (Clémençon et al., 2011), an extensive experimental study can be found in (Clémençon et al., 2012).

In the anomaly ranking context, the "negative distribution" is $U(dx)$. Therefore, in the situation where LeafRank is chosen as a cost-sensitive version of the CART algorithm with axis parallel splits (see (Breiman et al., 1984)), all the cells $\mathcal{C}_{j,k}$ can be expressed as union of hypercubes. The exact computation of the volume $U(\mathcal{C}_{j,k})$ 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. Hence, only empirical counterparts of the quantities $F(\mathcal{C})$ for subset $\mathcal{C} \subset [0,1]^d$ candidates,

$\widehat{F}_n(\mathcal{C}) = (1/n)\sum_{i=1}^{n} \mathbb{I}\{X \in \mathcal{C}\}$, 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 subsection 4.1. 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 (Clémençon & Vayatis, 2009).

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 section 5.1 in (Clémençon et al., 2011) for further details. This permits to identify the variables which have most relevance to detect anomalies.

5. Numerical Experiments

We now illustrate the points put forward in sections 3 and 4 by means of numerical experiments, based on unlabeled synthetic/real datasets. Precisely, we implemented the modification of the TREERANK procedure based on locally weighted versions of the CART method (with axis parallel splits) described at length in subsection 4.2, using the package for R statistical software (see http://www.r-project.org), available at http://treerank.sourceforge.net (with parameters: minsplit = 1, maxdepth = 4, in the LEAFRANK), see (Baskiotis et al., 2009). We have also used RankBoost (aggregating 30 stumps, see (Rudin et al., 2005)) and SVMRank (with linear and Gaussian kernels with cross-validated parameters, see (Herbrich et al., 2000)), using the SVM-light implementation available at http://svmlight.joachims.org/. The RankRLS method (http://www.tucs.fi/RLScore, see (Pahikkala et al., 2007)) that implements a regularized least square algorithm with linear kernel ("bias = 1") and with Gaussian kernel ($\gamma = 0.01$) has also been used, selection of the intercept on a grid being performed through a leave-one-out procedure. The anomaly ranking procedure based on the latter algorithms required to sample uniformly distributed "negative" data, as explained in subsection 4.1. As said at the end of section 4, the decision function output by the 1-class SVM procedure can be used as a scoring function, improperly however because the objective function it optimizes is related to a single point of the target MV curve (see the toy example below). We used the R-package Kernlab with gaussian kernel, with parameters chosen automatically by cross-validation. In the tables displayed below, the anomaly scoring function produced by RankBoost is referred to as "RankBoost", those computed by means of SVMrank (respectively, by means of RankRLS) based on a linear and a Gaussian kernels as "SVMlin" and "SVMgauss" (resp. "RLSlin" and "RLSgauss") and that produced by one-class SVM as "1cSVM".

In the following experiment, an estimate of the area under the MV-curve (AMV in short) is computed over 5 replications of a 5-fold cross validation as well as the overall standard deviation (denoted by σ).

5.1. Toy Examples and Synthetic Data

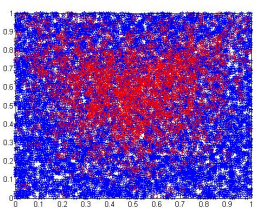
Let Z be a q-dimensional Gaussian r.v. with mean μ and covariance matrix Γ, and consider a borelian subset $C \subset \mathbb{R}^q$ with non zero Lebesgue measure. We denote by $\mathcal{N}_C(\mu, \Gamma)$ the conditional distribution of Z given $Z \in C$. Equipped with this notation, we can write the probability distribution used as toy example here as:

$$f(x) = \frac{1}{2}\mathcal{N}_{[0,1]^2}\left(\begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}, \begin{pmatrix} 1/30 & 0 \\ 0 & 1/30 \end{pmatrix}\right) + \frac{1}{4}\mathcal{N}_{[0,1]^2}\left(\begin{pmatrix} 3/10 \\ 7/10 \end{pmatrix}, \begin{pmatrix} 1/30 & -1/60 \\ -1/60 & 1/30 \end{pmatrix}\right) + \frac{1}{4}\mathcal{N}_{[0,1]^2}\left(\begin{pmatrix} 7/10 \\ 7/10 \end{pmatrix}, \begin{pmatrix} 1/30 & 1/60 \\ 1/60 & 1/30 \end{pmatrix}\right)$$

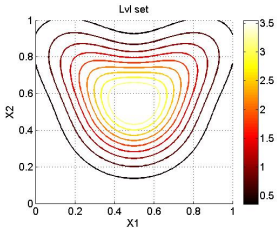
The simulated dataset is plotted in Fig. 2a, while some level sets of the density f are represented in Fig. 2b. We have independently sampled 5000 independent observations from distribution $f(x)dx$ and 5000 independent uniformly distributed points. The optimal AMV is denoted by AMV^* (knowing the density, it can be estimated by a basic Monte-Carlo scheme). As expected, given the distribution of the data to be ranked,

Table 1. Comparison of the AMV test - $AMV^* = 0.2393$

Method	TreeRank	RankBoost	SVMlin	SVMgauss	RLSlin	RLSgauss	ocSVM
AMV test	0.2448	0.2598	0.4128	0.2502	0.4129	0.2443	0.3373
σ	0.0124	0.0119	0.0125	0.0122	0.0123	0.0122	0.0042



a. Pooled sample: red circles represent instances with density f and blue stars those with uniform distribution.



b. Density level sets.

Figure 1. Mixture of Gaussian distributions

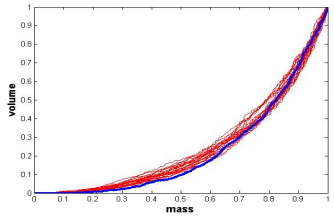


Figure 2. MV-curves: in blue, the MV-curve using the posterior distribution, in red the MV-curves using TreeRank.

the linear methods perform worst. Notice in addition that TreeRank and RLSgauss yield comparable results and outperform RankBoost, SVMgauss on this example. Among nonlinear rules, 1cSVM yields the poorest performance (*cf* section 4). The MV curves produced by the TreeRank algorithm in Fig. 3.

5.2. A Real-World Example

We also used a benchmark dataset in anomaly detection (computer network intrusion detection namely) proposed as a challenge for intrusion detection in the CMDC2013, see http://www.csmining.org/cdmc2013/ and (Song, 2013). In the present analysis, we kept the three features dst.host.rerror.rate, rerror.rate and serror.rate and removed degenerate features, yielding a training set of 6802 instances. Then, we simulated 10000 extra "negative" data, uniformly distributed over the cube $[0, 1]^3$. Estimates of the area under the MV curve (AMV) have been computed over five replications of a five-fold cross validation and the overall standard deviation is reported in Table 2. The modified TreeRank procedure outperforms all the other methods, illustrating the advantage of using a method avoiding any sampling stage. Given the clear superiority of the methods based on Gaussian kernels over linear techniques, one may also guess that the level sets of the underlying density are highly nonlinear.

6. Conclusion

Here we shed light on the connection between *anomaly ranking*, cast as *Mass Volume curve* minimization, and bipartite ranking when the distribution $F(dx)$ of the (unlabeled) training data is compactly supported. Assuming (rather than rescaling the observations) that the support is included in $[0, 1]^d$, the related bipartite problem corresponds to the situation where $F(dx)$ is the "positive" distribution and the "negative" one is uniform on $[0, 1]^d$. We thus proved that, following the generation of uniformly distributed data, bipartite ranking algorithms can be readily used to build scoring functions which nearly solve MV curve minimization. We illustrated this through numerical experiments.

References

Agarwal, S., Graepel, T., Herbrich, R., Har-Peled, S., and Roth, D. Generalization bounds for the area under the ROC curve. *JMLR*, 6:393–425, 2005.

Baskiotis, N., Clémençon, S., Depecker, M., and Vay-

Table 2. Comparison of the AMV test - Experiment CMD2013

Method	TreeRank	RankBoost	SVMlin	SVMgauss	RLSlin	RLSgauss
AMV test	0.0225	0.1500	0.4029	0.2501	0.4070	0.0481
σ	0.0027	0.0133	0.0136	0.0125	0.0135	0.0051

atis, N. R-implementation of the TreeRank algorithm. *Submitted for publication*, 2009.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. *Classification and Regression Trees*. Wadsworth and Brooks, 1984.

Clémençon, S. and Jakubowicz, J. Scoring anomalies: a M-estimation formulation. In *Proceedings of AISTATS*, 2013.

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

Clémençon, S., Lugosi, G., and Vayatis, N. Ranking and empirical risk minimization of U-statistics. *Ann. Stat.*, 36(2):844–874, 2008.

Clémençon, S., Depecker, M., and Vayatis, N. Adaptive partitioning schemes for bipartite ranking. *Machine Learning*, 43(1):31–69, 2011.

Clémençon, S., Depecker, M., and Vayatis, N. An empirical comparison of learning algorithms for nonparametric scoring: the treerank algorithm and other methods. *Patt. Analys. Appl.*, 2012.

Clémençon, S., Depecker, M., and Vayatis, N. Ranking Forests. *J. Mach. Learn. Res.*, 2013.

Duchi, J., Mackey, L., and Jordan, M. On the consistency of ranking algorithms. In *Proceedings of ICML*, 2010.

Einmahl, J.H.J. and Mason, D.M. Generalized quantile process. *Ann. Stat.*, 20:1062–1078, 1992.

Fawcett, T. An Introduction to ROC Analysis. *Letters in Pattern Recognition*, 27(8):861–874, 2006.

Freund, Y., Iyer, R., Schapire, R., and Singer, Y. An efficient boosting algorithm for combining preferences. *JMLR*, 4:933–969, 2003.

Friedman, J., Hastie, T., and Tibshirani, R. *The Elements of Statistical Learning*. Springer, 2009.

Herbrich, R., Graepel, T., and Obermayer, K. *Advances in Large Margin Classifiers*, chapter Large margin rank boundaries for ordinal regression, pp. 115–132. MIT Press, 2000.

Pahikkala, T., Tsivtsivadze, E., Airola, A., Boberg, J., and Salakoski, T. Learning to rank with pairwise regularized least-squares. In *Proceedings of SIGIR*, pp. 27–33, 2007.

Polonik, W. Minimum volume sets and generalized quantile processes. *Stochastic Processes and their Applications*, 69(1):1–24, 1997.

Rakotomamonjy, A. Optimizing Area Under Roc Curve with SVMs. In *Proceedings of the First Workshop on ROC Analysis in AI*, 2004.

Rudin, C., Cortes, C., Mohri, M., and Schapire, R. E. Margin-based ranking and boosting meet in the middle. In *Proceedings of COLT*, 2005.

Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A., and Williamson, R. Estimating the Support of a High-Dimensional Distribution. *Neural Computation*, 13(7):1443–1471, 2001.

Scott, C. and Davenport, M. Regression level set estimation via cost-sensitive classification. *IEEE Transactions on Signal Processing*, 55, 2007.

Scott, C. and Nowak, R. Learning minimum volume sets. *JMLR*, 7:665–704, 2006.

Song, J. Cdmc2013 intrusion detection dataset, 2013.

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

Vert, R. and Vert, J.-P. Consistency and convergence rates of one-class svms and related algorithms. *JMLR*, 7:817–854, 2006.

Viswanathan, K., Choudur, L., Talwar, V., Wang, C., Macdonald, G., and Satterfield, W. Ranking anomalies in data centers. In R.D.James (ed.), *Network Operations and System Management*, pp. 79–87. IEEE, 2012.