Journal of Machine Learning Research 6 (2005) 211–232 Submitted 11/04; Published 3/05

A Classification Framework for Anomaly Detection

Ingo Steinwart INGO@LANL.GOV
Don Hush DHUSH@LANL.GOV
Clint Scovel JCS@LANL.GOV
Modeling, Algorithms and Informatics Group, CCS-3
Los Alamos National Laboratory
Los Alamos, NM 87545, USA

Editor: Bernhard Schölkopf

Abstract

One way to describe anomalies is by saying that anomalies are not concentrated. This leads to the problem of finding level sets for the data generating density. We interpret this learning problem as a binary classification problem and compare the corresponding classification risk with the standard performance measure for the density level problem. In particular it turns out that the empirical classification risk can serve as an empirical performance measure for the anomaly detection problem. This allows us to compare different anomaly detection algorithms *empirically*, i.e. with the help of a test set. Furthermore, by the above interpretation we can give a strong justification for the well-known heuristic of artificially sampling "labeled" samples, provided that the sampling plan is well chosen. In particular this enables us to propose a support vector machine (SVM) for anomaly detection for which we can easily establish universal consistency. Finally, we report some experiments which compare our SVM to other commonly used methods including the standard one-class SVM.

Keywords: unsupervised learning, anomaly detection, density levels, classification, SVMs

1. Introduction

Anomaly (or novelty) detection aims to detect anomalous observations from a system. In the machine learning version of this problem we cannot directly model the normal behaviour of the system since it is either unknown or too complex. Instead, we have some sample observations from which the normal behaviour is to be learned. This anomaly detection learning problem has many important applications including the detection of e.g. anomalous jet engine vibrations (see Nairac et al., 1999; Hayton et al., 2001; King et al., 2002), abnormalities in medical data (see Tarassenko et al., 1995; Campbell and Bennett, 2001), unexpected conditions in engineering (see Desforges et al., 1998) and network intrusions (see Manikopoulos and Papavassiliou, 2002; Yeung and Chow, 2002; Fan et al., 2001). For more information on these and other areas of applications as well as many methods for solving the corresponding learning problems we refer to the recent survey of Markou and Singh (2003a,b).

It is important to note that a typical feature of these applications is that only unlabeled samples are available, and hence one has to make some a-priori assumptions on anomalies in order to be able to distinguish between normal and anomalous future oberservations. One of the most common ways to define anomalies is by saying that *anomalies are not concentrated* (see e.g. Ripley, 1996;

Schölkopf and Smola, 2002). To make this precise let Q be our *unknown data-generating distribution* on the input space X which has a density h with respect to a *known reference distribution* μ on X. Obviously, the density level sets $\{h > \rho\}$, $\rho > 0$, describe the concentration of Q. Therefore to define anomalies in terms of the concentration one only has to fix a threshold level $\rho > 0$ so that a sample $x \in X$ is considered to be anomalous whenever $h(x) \leq \rho$. Consequently, our aim is to find the set $\{h \leq \rho\}$ to detect anomalous observations, or equivalently, the ρ-*level set* $\{h > \rho\}$ to describe normal observations.

We emphasize that given the data-generating distribution Q the choice of μ determines the density h, and consequently *anomalies are actually modeled by both* μ *and* ρ. Unfortunately, many popular algorithms are based on density estimation methods that implicitly assume μ to be the uniform distribution (e.g. Gaussian mixtures, Parzen windows and k-nearest neighbors density estimates) and therefore for these algorithms defining anomalies is restricted to the choice of ρ. With the lack of any further knowledge one might feel that the uniform distribution is a reasonable choice for μ, however there are situations in which a different μ is more appropriate. In particular, this is true if we consider a modification of the anomaly detection problem where μ is not known but can be sampled from. We will see that unlike many others our proposed method can handle both problems.

Finding level sets of an unknown density is also a well known problem in statistics which has some important applications different from anomaly detection. For example, it can be used for the problem of cluster analysis as described in by Hartigan (1975) and Cuevas et al. (2001), and for testing of multimodality (see e.g. Müller and Sawitzki, 1991; Sawitzki, 1996). Some other applications including estimation of non-linear functionals of densities, density estimation, regression analysis and spectral analysis are briefly described by Polonik (1995). Unfortunately, the algorithms considered in these articles cannot be used for the anomaly detection problem since the imposed assumptions on h are often tailored to the above applications and are in general unrealistic for anomalies.

One of the main problems of anomaly detection—or more precisely density level detection—is the lack of an empirical performance measure which allows us to compare the generalization performance of different algorithms by test samples. By interpreting the density level detection problem as binary classification with respect to an appropriate measure, we show that the corresponding empirical classification risk can serve as such an empirical performance measure for anomaly detection. Furthermore, we compare the excess classification risk with the standard performance measure for the density level detection problem. In particular, we show that both quantities are asymptotically equivalent and that simple inequalities between them are possible under mild conditions on the density h.

A well-known heuristic (see e.g. Fan et al., 2001; González and Dagupta, 2003; Yu et al., 2004; Theiler and Cai., 2003) for anomaly detection is to generate a labeled data set by assigning one label to the original unlabeled data and another label to a set of artificially generated data, and then apply a binary classification algorithm. By interpreting the density level detection problem as a binary classification problem we can show that this heuristic can be strongly justified provided that the sampling plan for the artificial samples is chosen in a certain way and the used classification algorithm is well-adopted to this plan. We will work out this justification in detail by showing how to modify the standard support vector machine (SVM) for classification, and establishing a consistency result for this modification. Finally we report some experiments comparing the modified SVM with some other commonly used algorithms including Gaussian maximum-likelihood methods, and the standard one-class SVM proposed by Schölkopf et al. (2001).

2. Detecting Density Levels is a Classification Problem

We begin with rigorously defining the density level detection (DLD) problem. To this end let $(X, \mathcal{A})$ be a measurable space and μ a *known* distribution on $(X, \mathcal{A})$. Furthermore, let Q be an *unknown* distribution on $(X, \mathcal{A})$ which has an *unknown* density h with respect to μ, i.e. $dQ = hd\mu$. Given a $\rho > 0$ the set $\{h > \rho\}$ is called the ρ-*level set* of the density h. Throughout this work we assume that $\{h = \rho\}$ is a μ-zero set and hence it is also a Q-zero set. For the density level detection problem and related tasks this is a common assumption (see e.g. Polonik, 1995; Tsybakov, 1997).

Now, the goal of the DLD problem is to find an estimate of the ρ-level set of h. To this end we need some information which in our case is given to us by a training set $T = (x_1, \ldots, x_n) \in X^n$. We will assume in the following that T is i.i.d. drawn from Q. With the help of T a DLD algorithm constructs a function $f_T : X \to \mathbb{R}$ for which the set $\{f_T > 0\}$ is an estimate of the ρ-level set $\{h > \rho\}$ of interest. Since in general $\{f_T > 0\}$ does not excactly coincide with $\{h > \rho\}$ we need a *performance measure* which describes how well $\{f_T > 0\}$ approximates the set $\{h > \rho\}$. Probably the best known performance measure (see e.g. Tsybakov, 1997; Ben-David and Lindenbaum, 1997, and the references therein) for measurable functions $f : X \to \mathbb{R}$ is

$$\mathcal{S}_{\mu,h,\rho}(f) \;:=\; \mu\Big(\{f > 0\} \triangle \{h > \rho\}\Big),$$

where $\triangle$ denotes the symmetric difference. Obviously, the smaller $\mathcal{S}_{\mu,h,\rho}(f)$ is, the more $\{f > 0\}$ coincides with the ρ-level set of h and a function f minimizes $\mathcal{S}_{\mu,h,\rho}$ if and only if $\{f > 0\}$ is μ-almost surely identical to $\{h > \rho\}$. Furthermore, for a sequence of functions $f_n : X \to \mathbb{R}$ with $\mathcal{S}_{\mu,h,\rho}(f_n) \to 0$ we easily see that $\operatorname{sign} f_n(x) \to 1_{\{h>\rho\}}(x)$ for μ-almost all $x \in X$, and since Q is absolutely continuous with respect to μ the same convergence holds Q-almost surely. Finally, it is important to note, that the performance measure $\mathcal{S}_{\mu,h,\rho}$ is insensitive to μ-zero sets. Since we cannot detect μ-zero sets using a training set T drawn from Q^n this feature is somehow natural for our model.

Although $\mathcal{S}_{\mu,h,\rho}$ seems to be well-adapted to our model, it has a crucial disadvantage in that we cannot compute $\mathcal{S}_{\mu,h,\rho}(f)$ since $\{h > \rho\}$ is unknown to us. Therefore, we have to *estimate* it. In our model the only information we can use for such an estimation is a *test set* $W = (\hat{x}_1, \ldots, \hat{x}_m)$ which is i.i.d. drawn from Q. Unfortunately, there is no method known to estimate $\mathcal{S}_{\mu,h,\rho}(f)$ from W with *guaranteed* accuracy in terms of m, f, μ and ρ, and we strongly believe that such a method cannot exist. Because of this lack, we cannot *empirically* compare different algorithms in terms of the performance measure $\mathcal{S}_{\mu,h,\rho}$.

Let us now describe another performance measure which has merits similar to $\mathcal{S}_{\mu,h,\rho}$ but additionally has an empirical counterpart, i.e. a method to estimate its value with guaranteed accuracy by only using a test set. This performance measure is based on interpreting the DLD problem as a binary classification problem in which T is assumed to be positively labeled and infinitely many negatively labeled samples are available by the knowledge of μ. To make this precise we write $Y := \{-1, 1\}$ and define

Definition 1 *Let μ and Q be probability measures on X and $s \in (0,1)$. Then the probability measure $Q \ominus_s \mu$ on $X \times Y$ is defined by*

$$Q \ominus_s \mu(A) \;:=\; s\mathbb{E}_{x\sim Q} 1_A(x, 1) + (1-s)\mathbb{E}_{x\sim \mu} 1_A(x, -1)$$

for all measurable subsets $A \subset X \times Y$. Here we used the shorthand $1_A(x,y) := 1_A((x,y))$.

Roughly speaking, the distribution $Q \ominus_s \mu$ measures the "1-slice" of $A \subset X \times Y$ by sQ and the "-1-slice" by $(1-s)\mu$. Moreover, the measure $P := Q \ominus_s \mu$ can obviously be associated with a binary classification problem in which positive samples are drawn from sQ and negative samples are drawn from $(1-s)\mu$. Inspired by this interpretation let us recall that the binary classification risk for a measurable function $f : X \to \mathbb{R}$ and a distribution P on $X \times Y$ is defined by

$$\mathcal{R}_P(f) \;=\; P\big(\{(x,y) : \operatorname{sign} f(x) \neq y\}\big)\,,$$

where we define $\operatorname{sign} t := 1$ if $t > 0$ and $\operatorname{sign} t = -1$ otherwise. Furthermore, the *Bayes risk* $\mathcal{R}_P$ of P is the smallest possible classification risk with respect to P, i.e.

$$\mathcal{R}_P \;:=\; \inf\Big\{ \mathcal{R}_P(f) \,\Big|\, f : X \to \mathbb{R} \text{ measurable} \Big\}\,.$$

We will show in the following that learning with respect to $\mathcal{S}_{\mu,h,\rho}$ is equivalent to learning with respect to $\mathcal{R}_P(.)$. To this end we begin by computing the marginal distribution P_X and the *supervisor* $\eta(x) := P(y = 1|x)$, $x \in X$, of $P := Q \ominus_s \mu$:

Proposition 2 *Let μ and Q be probability measures on X such that Q has a density h with respect to μ, and let $s \in (0,1)$. Then the marginal distribution of $P := Q \ominus_s \mu$ on X is $P_X = sQ + (1-s)\mu$. Furthermore, we P_X-a.s. have*

$$P(y = 1|x) = \frac{sh(x)}{sh(x) + 1 - s}\,.$$

Proof As recalled in the appendix, $P(y = 1|x)$, $x \in X$, is a regular conditional probability and hence we only have to check the condition of Corollary 19. To this end we first observe by the definition of $P := Q \ominus_s \mu$ that for all non-negative, measurable functions $f : X \times Y \to \mathbb{R}$ we have

$$\int_{X\times Y} f\,dP \;=\; s\int_X f(x,1)Q(dx) + (1-s)\int_X f(x,-1)\mu(dx)\,.$$

Therefore, for $A \in \mathcal{A}$ we obtain

$$\begin{aligned}
&\int_{A\times Y} \frac{sh(x)}{sh(x)+1-s} P(dx,dy) \\
&= s\int_A \frac{sh(x)}{sh(x)+1-s} h(x)\mu(dx) + (1-s)\int_A \frac{sh(x)}{sh(x)+1-s}\mu(dx) \\
&= \int_A sh(x)\mu(dx) \\
&= s\int_A 1_{X\times\{1\}}(x,1)Q(dx) + (1-s)\int_A 1_{X\times\{1\}}(x,-1)\mu(dx) \\
&= \int_{A\times Y} 1_{X\times\{1\}}\,dP\,.
\end{aligned}$$

■

Note that the formula for the marginal distribution P_X in particular shows that the μ-zero sets of X are exactly the P_X-zero sets of X. As an immediate consequence of the above proposition we additionally obtain the following corollary which describes the ρ-level set of h with the help of the supervisor η:

Corollary 3 *Let μ and Q be probability measures on X such that Q has a density h with respect to μ. For $\rho > 0$ we write $s := \frac{1}{1+\rho}$ and define $P := Q \ominus_s \mu$. Then for $\eta(x) := P(y = 1|x)$, $x \in X$, we have*

$$\mu\Big(\{\eta > 1/2\} \triangle \{h > \rho\}\Big) \;=\; 0\,,$$

i.e. $\{\eta > 1/2\}$ μ-almost surely coincides with $\{h > \rho\}$.

Proof By Proposition 2 we see that $\eta(x) > \frac{1}{2}$ is μ-almost surely equivalent to $\frac{sh(x)}{sh(x)+1-s} > \frac{1}{2}$ which is equivalent to $h(x) > \frac{1-s}{s} = \rho$. ∎

The above results in particular show that every distribution $P := Q \ominus_s \mu$ with $dQ := hd\mu$ and $s \in (0,1)$ determines a triple (μ, h, ρ) with $\rho := (1-s)/s$ and vice-versa. In the following we therefore use the shorthand $\mathcal{S}_P(f) := \mathcal{S}_{\mu,h,\rho}(f)$.

Let us now compare $\mathcal{R}_P(.)$ with $\mathcal{S}_P(.)$. To this end recall, that binary classification aims to discriminate $\{\eta > 1/2\}$ from $\{\eta < 1/2\}$. In view of the above corollary it is hence no surprise that $\mathcal{R}_P(.)$ can serve as a surrogate for $\mathcal{S}_P(.)$ as the following theorem shows:

Theorem 4 *Let μ and Q be probability measures on X such that Q has a density h with respect to μ. Let $\rho > 0$ be a real number which satisfies $\mu(\{h = \rho\}) = 0$. We write $s := \frac{1}{1+\rho}$ and define $P := Q \ominus_s \mu$. Then for all sequences (f_n) of measurable functions $f_n : X \to \mathbb{R}$ the following are equivalent:*

i) $\mathcal{S}_P(f_n) \to 0$.

ii) $\mathcal{R}_P(f_n) \to \mathcal{R}_P$.

In particular, for a measurable function $f : X \to \mathbb{R}$ we have $\mathcal{S}_P(f) = 0$ if and only if $\mathcal{R}_P(f) = \mathcal{R}_P$.

Proof For $n \in \mathbb{N}$ we define $E_n := \{f_n > 0\} \triangle \{h > \rho\}$. Since by Corollary 3 we know $\mu(\{h > \rho\} \triangle \{\eta > \frac{1}{2}\}) = 0$ it is easy to see that the classification risk of f_n can be computed by

$$\mathcal{R}_P(f_n) \;=\; \mathcal{R}_P + \int_{E_n} |2\eta - 1| dP_X\,. \tag{1}$$

Now, $\{|2\eta - 1| = 0\}$ is a μ-zero set and hence a P_X-zero set. The latter implies that the measures $|2\eta - 1|dP_X$ and P_X are absolutely continuous with respect to each other, and hence we have

$$|2\eta - 1|dP_X(E_n) \to 0 \qquad \text{if and only if} \qquad P_X(E_n) \to 0\,.$$

Furthermore, we have already observed after Proposition 2 that P_X and μ are absolutely continuous with respect to each other, i.e. we also have

$$P_X(E_n) \to 0 \qquad \text{if and only if} \qquad \mu(E_n) \to 0\,.$$

Therefore, the assertion follows from $\mathcal{S}_P(f_n) = \mu(E_n)$. ∎

Theorem 4 shows that instead of using $\mathcal{S}_P$ as a performance measure for the density level detection problem one can alternatively use the classification risk $\mathcal{R}_P(.)$. Therefore, we will establish some basic properties of this performance measure in the following. To this end we write $I(y,t) := 1_{(-\infty,0]}(yt)$, $y \in Y$ and $t \in \mathbb{R}$, for the standard classification loss function. With this notation we can compute $\mathcal{R}_P(f)$:

Proposition 5 *Let μ and Q be probability measures on X. For $\rho > 0$ we write $s := \frac{1}{1+\rho}$ and define $P := Q \ominus_s \mu$. Then for all measurable $f : X \to \mathbb{R}$ we have*

$$\mathcal{R}_P(f) \;=\; \frac{1}{1+\rho}\mathbb{E}_Q I(1, \operatorname{sign} f) + \frac{\rho}{1+\rho}\mathbb{E}_\mu I(-1, \operatorname{sign} f)\,.$$

Furthermore, for the Bayes risk we have

$$\mathcal{R}_P \;\leq\; \min\Big\{\frac{1}{1+\rho}, \frac{\rho}{1+\rho}\Big\}$$

and

$$\mathcal{R}_P \;=\; \frac{1}{1+\rho}\mathbb{E}_Q 1_{\{h \leq \rho\}} + \frac{\rho}{1+\rho}\mathbb{E}_\mu 1_{\{h > \rho\}}\,.$$

Proof The first assertion directly follows from

$$\begin{aligned}
\mathcal{R}_P(f) &= P\big(\{(x,y) : \operatorname{sign} f(x) \neq y\}\big) \\
&= P\big(\{(x,1) : \operatorname{sign} f(x) = -1\}\big) + P\big(\{(x,-1) : \operatorname{sign} f(x) = 1\}\big) \\
&= sQ\big(\{\operatorname{sign} f = -1\}\big) + (1-s)\mu\big(\{\operatorname{sign} f = 1\}\big) \\
&= s\mathbb{E}_Q I(1, \operatorname{sign} f) + (1-s)\mathbb{E}_\mu I(-1, \operatorname{sign} f)\,.
\end{aligned}$$

The second assertion directly follows from $\mathcal{R}_P \leq \mathcal{R}_P(1_X) \leq s$ and $\mathcal{R}_P \leq \mathcal{R}_P(-1_X) \leq 1-s$. Finally, for the third assertion recall that $f = 1_{\{h>\rho\}} - 1_{\{h\leq\rho\}}$ is a function which realizes the Bayes risk. ■

As described at the beginning of this section our main goal is to find a performance measure for the density level detection problem which has an empirical counterpart. In view of Proposition 5 the choice of an empirical counterpart for $\mathcal{R}_P(.)$ is rather obvious:

Definition 6 *Let μ be a probability measure on X and $\rho > 0$. Then for $T = (x_1, \ldots, x_n) \in X^n$ and a measurable function $f : X \to \mathbb{R}$ we define*

$$\mathcal{R}_T(f) \;:=\; \frac{1}{(1+\rho)n}\sum_{i=1}^{n} I(1, \operatorname{sign} f(x_i)) + \frac{\rho}{1+\rho}\mathbb{E}_\mu I(-1, \operatorname{sign} f)\,.$$

If we identify T with the corresponding empirical measure it is easy to see that $\mathcal{R}_T(f)$ is the classification risk with respect to the measure $T \ominus_s \mu$ for $s := \frac{1}{1+\rho}$. Then for measurable functions $f : X \to \mathbb{R}$, e.g. Hoeffding's inequality shows that $\mathcal{R}_T(f)$ approximates the true classification risk $\mathcal{R}_P(f)$ in a fast and controllable way.

It is highly interesting that the classification risk $\mathcal{R}_P(.)$ is strongly connected with another approach for the density level detection problem which is based on the so-called *excess mass* (see e.g. Hartigan, 1987; Müller and Sawitzki, 1991; Polonik, 1995; Tsybakov, 1997, and the references therein). To be more precise let us first recall that the excess mass of a measurable function $f : X \to \mathbb{R}$ is defined by

$$\mathcal{E}_P(f) \;:=\; Q(\{f > 0\}) - \rho\mu(\{f > 0\})\,,$$

where Q, ρ and μ have the usual meaning. The following proposition shows that $\mathcal{R}_P(.)$ and $\mathcal{E}_P(.)$ are essentially the same:

Proposition 7 *Let μ and Q be probability measures on X. For $\rho > 0$ we write $s := \frac{1}{1+\rho}$ and define $P := Q \ominus_s \mu$. Then for all measurable $f : X \to \mathbb{R}$ we have*

$$\mathcal{E}_P(f) \;=\; 1 - (1+\rho)\mathcal{R}_P(f)\,.$$

Proof We obtain the assertion by the following simple calculation:

$$\begin{aligned}
\mathcal{E}_P(f) &= Q(\{f > 0\}) - \rho\mu(\{f \geq 0\}) \\
&= 1 - Q(\{f \leq 0\}) - \rho\mu(\{f > 0\}) \\
&= 1 - Q\big(\{\operatorname{sign} f = -1\}\big) - \rho\mu\big(\{\operatorname{sign} f = 1\}\big) \\
&= 1 - (1+\rho)\mathcal{R}_P(f)\,.
\end{aligned}$$

■

If Q is an empirical measure based on a training set T in the definition of $\mathcal{E}_P(.)$ then we obtain am empirical performance measure which we denote by $\mathcal{E}_T(.)$. By the above proposition we have

$$\mathcal{E}_T(f) = 1 - \frac{1}{n}\sum_{i=1}^{n} I(1, \operatorname{sign} f(x_i)) - \rho\mathbb{E}_\mu I(-1, \operatorname{sign} f) = 1 - (1+\rho)\mathcal{R}_T(f) \tag{2}$$

for all measurable $f : X \to \mathbb{R}$. Now, given a class $\mathcal{F}$ of measurable functions from X to $\mathbb{R}$ the (empirical) excess mass approach considered e.g. by Hartigan (1987); Müller and Sawitzki (1991); Polonik (1995); Tsybakov (1997), chooses a function $f_T \in \mathcal{F}$ which maximizes $\mathcal{E}_T(.)$ within $\mathcal{F}$. By equation (2) we see that this approach is actually a type of empirical risk minimization (ERM). Surprisingly, this connection has not been observed, yet. In particular, the excess mass has only been considered as an algorithmic tool, but not as a performance measure. Instead, the papers dealing with the excess mass approach measures the performance by $\mathcal{S}_P(.)$. In their analysis an additional assumption on the behaviour of h around the level ρ is required. Since this condition can also be used to establish a quantified version of Theorem 4 we will recall it now:

Definition 8 *Let μ be a distribution on X and $h : X \to [0,\infty)$ be a measurable function with $\int h d\mu = 1$, i.e. h is a density with respect to μ. For $\rho > 0$ and $0 \leq q \leq \infty$ we say that h has* ρ-exponent q *if there exists a constant $C > 0$ such that for all sufficiently small $t > 0$ we have*

$$\mu\big(\{|h - \rho| \leq t\}\big) \;\leq\; Ct^q\,. \tag{3}$$

Condition (3) was first considered by Polonik (1995, Thm. 3.6). This paper also provides an example of a class of densities on $\mathbb{R}^d$, $d \geq 2$, which has exponent $q = 1$. Later, Tsybakov (1997, p. 956) used (3) for the analysis of a density level detection method which is based on a localized version of the empirical excess mass approach.

Interestingly, condition (3) is closely related to a concept for binary classification called the Tsybakov noise exponent (see e.g. Mammen and Tsybakov, 1999; Tsybakov, 2004; Steinwart and Scovel, 2004) as the following proposition proved in the appendix shows:

Proposition 9 *Let μ and Q be distributions on X such that Q has a density h with respect to μ. For $\rho > 0$ we write $s := \frac{1}{1+\rho}$ and define $P := Q \ominus_s \mu$. Then for $0 < q \leq \infty$ the following are equivalent:*

i) h has ρ-exponent *q.*

ii) P has Tsybakov noise exponent q, i.e. there exists a constant $C > 0$ *such that for all sufficiently small* $t > 0$ *we have*

$$P_X\big(|2\eta - 1| \le t\big) \;\le\; C \cdot t^q \tag{4}$$

In recent years Tsybakov's noise exponent has played a crucial role for establishing learning rates faster than $n^{-\frac{1}{2}}$ for certain algorithms (see e.g. Mammen and Tsybakov, 1999; Tsybakov, 2004; Steinwart and Scovel, 2004). Remarkably, it was already observed by Mammen and Tsybakov (1999), that the classification problem can be analyzed by methods originally developed for the DLD problem. However, to our best knowledge the exact relation between the DLD problem and binary classification has not been presented, yet. In particular, it has not been observed yet, that this relation opens a *non-heuristic* way to use classification methods for the DLD problem as we will discuss in the next section.

As already announced we can also establish inequalities between $\mathcal{S}_P$ and $\mathcal{R}_P(.)$ with the help of the ρ-exponent. This is done in the following theorem:

Theorem 10 *Let* $\rho > 0$ *and* μ *and* Q *be probability measures on* X *such that* Q *has a density* h *with respect to* μ. *For* $s := \frac{1}{1+\rho}$ *we write* $P := Q \ominus_s \mu$. *Then the following statements hold:*

i) If h *is bounded then there exists a constant* $c > 0$ *such that for all measurable* $f : X \to \mathbb{R}$ *we have*

$$\mathcal{R}_P(f) - \mathcal{R}_P \;\le\; c\,\mathcal{S}_P(f)\,.$$

ii) If h *has* ρ*-exponent* $q \in (0,\infty]$ *then there exists a constant* $c > 0$ *such that for all measurable* $f : X \to \mathbb{R}$ *we have*

$$\mathcal{S}_P(f) \;\le\; c\big(\mathcal{R}_P(f) - \mathcal{R}_P\big)^{\frac{q}{1+q}}\,.$$

Proof The first assertion directly follows from (1) and Proposition 2. The second assertion follows from Proposition 9 and a result of Tsybakov (2004, Prop. 1). ■

Remark 11 *We note that many of the results of this section can be generalized to the case where* Q *is not absolutely continuous with respect to* μ. *Indeed, select an auxilliary measure* ν *such that both* Q *and* μ *are absolutely continuous with respect to* ν. *For example one could choose* $\nu = \frac{Q+\mu}{2}$. *Consequently we have* $Q = h_1\nu$ *and* $\mu = h_2\nu$ *for some real valued functions* h_1 *and* h_2. *Then Proposition 2 holds with* $h(x) := \frac{h_1(x)}{h_2(x)}$, *where one defines the righthand side to be* 0 *when* $h_1(x) = h_2(x) = 0$. *One can also show that* h *is* P_X*-a.s. independent of the choice of* ν. *Corollary 3 holds where the measure of the symmetric difference is evaluated with either* Q *or* μ. *However it appears that only the "*$\mathcal{R}_P(f_n) \to \mathcal{R}_P \Rightarrow \mathcal{S}_P(f_n) \to 0$*" assertion of Theorem 4 holds instead of equivalence. Finally, Propositions 5 and 7 hold, Proposition 9 holds with a suitable generalization of Definition 8 of* ρ*-exponent, and the second assertion of Theorem 10 holds.*

3. A Support Vector Machine for Density Level Detection

In the previous section we have shown that the DLD problem can be interpreted as a binary classification problem in which one conditional class probability is known. We now show that this interpretation has far reaching algorithmic consequences. To this end let us assume that we give each sample of our training set $T = (x_1, \ldots, x_n)$ drawn from Q the label 1. Additionally we generate a second training set $T' = (x'_1, \ldots, x'_{n'})$ from μ and label each sample of it with -1. Merging these labeled sample sets gives a new training set which then can be used by a binary classification algorithm. By our considerations in the previous section it seems reasonable to expect that the used classification algorithm actually learns the DLD problem provided that the algorithm is well-adjusted to the sample set sizes and the parameter ρ.

In the following we work this high-level approach out by constructing an SVM for the DLD problem. To this end let $k : X \times X \to \mathbb{R}$ be a positive definite kernel with reproducing kernel Hilbert space (RKHS) H. Furthermore, let μ be a known probability measure on X and $l : Y \times \mathbb{R} \to [0,\infty)$ be the *hinge* loss function, i.e. $l(y,t) := \max\{0, 1-yt\}$, $y \in Y$, $t \in \mathbb{R}$. Then for a training set $T = (x_1, \ldots, x_n) \in X^n$, a regularization parameter $\lambda > 0$, and $\rho > 0$ we initially define

$$f_{T,\mu,\lambda} := \arg\min_{f \in H} \lambda \|f\|_H^2 + \frac{1}{(1+\rho)n} \sum_{i=1}^{n} l(1, f(x_i)) + \frac{\rho}{1+\rho} \mathbb{E}_{x\sim\mu} l(-1, f(x)), \tag{5}$$

and

$$(\tilde{f}_{T,\mu,\lambda}, \tilde{b}_{T,\mu,\lambda}) := \arg\min_{\substack{f \in H \\ b \in R}} \lambda \|f\|_H^2 + \frac{1}{(1+\rho)n} \sum_{i=1}^{n} l(1, f(x_i)+b) + \frac{\rho}{1+\rho} \mathbb{E}_{x\sim\mu} l(-1, f(x)+b). \tag{6}$$

The decision function of the *SVM without offset* is $f_{T,\mu,\lambda} : X \to \mathbb{R}$ and analogously, the *SVM with offset* has the decision function $\tilde{f}_{T,\mu,\lambda} + \tilde{b}_{T,\mu,\lambda} : X \to \mathbb{R}$.

Although the measure μ is known, almost always the expectation $\mathbb{E}_{x\sim\mu} l(-1, f(x))$ can only be numerically computed which requires finitely many function evaluations of f. If the integrand of this expectation was smooth we could use some known deterministic methods to choose these function evaluations efficiently. However, since the hinge loss is not differentiable there is no such method known to us. According to our above plan we will therefore use points $T' := (x'_1, \ldots, x'_{n'})$ which are randomly sampled from μ to approximate $\mathbb{E}_{x\sim\mu} l(-1, f(x))$ by $\frac{1}{n'} \sum_{i=1}^{n'} l(-1, f(x'_i))$. We denote the corresponding approximate solutions of (5) and (6) by $f_{T,T',\lambda}$ and $(\tilde{f}_{T,T',\lambda}, \tilde{b}_{T,T',\lambda})$, respectively. Furthermore, in these cases the formulations (5) and (6) are identical to the standard L1-SVM formulations besides the weighting factors in front of the empirical error terms. Therefore, the derivation of the corresponding dual problems is straightforward. For example, the dual problem for (6) can be written as follows:

$$\begin{aligned}
\max \quad & \sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n'} \alpha'_i - \tfrac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) - \tfrac{1}{2} \sum_{i,j=1}^{n'} \alpha'_i \alpha'_j k(x'_i, x'_j) + \sum_{i,j=1}^{n,n'} \alpha_i \alpha'_j k(x_i, x'_j) \\
\text{s.t.} \quad & \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n'} \alpha'_i = 0, \\
& 0 \le \alpha_i \le \tfrac{2}{\lambda(1+\rho)n}, \qquad i = 1, \ldots, n, \\
& 0 \le \alpha'_i \le \tfrac{2\rho}{\lambda(1+\rho)n'}, \qquad i = 1, \ldots, n'.
\end{aligned} \tag{7}$$

The fact that the SVM for DLD essentially coincides with the standard L1-SVM also allows us to modify many known results for these algorithms. For simplicity we will only state a consistency result which describes the case where we use $n' = n$ random samples from μ in order to approximate the expectation with respect to μ. However, it is straight forward to extend the result to the more general case of $n' = rn$ samples for some positive $r \in \mathbb{Q}$. In order to formulate the result we have to recall the notion of universal kernels (see Steinwart, 2001). To this end let X be a compact metric space, say a closed and bounded subset of $\mathbb{R}^d$. We denote the space of all continuous functions on X by $C(X)$. As usual, this space is equipped with the supremum norm $\|.\|_\infty$. Then the RKHS H of a continuous kernel k on X is embedded into $C(X)$, i.e. $H \subset C(X)$, where the inclusion is continuous. We say that the kernel k is *universal*, if in addition H is dense in $C(X)$, i.e. for every $f \in C(X)$ and every $\varepsilon > 0$ there exists a $g \in H$ with $\|f - g\|_\infty < \varepsilon$. Some examples of universal kernels including the Gaussian RBF kernels were presented by Steinwart (2001).

Now we can formulate the announced result:

Theorem 12 (Universal consistency) *Let X be a compact metric space and k be a universal kernel on X. Furthermore, let $\rho > 0$, and μ and Q be probability measures on X such that Q has a density h with respect to μ. For $s := \frac{1}{1+\rho}$ we write $P := Q \ominus_s \mu$. Then for all sequences (λ_n) of positive numbers with $\lambda_n \to 0$ and $n\lambda_n^2 \to \infty$ and for all $\varepsilon > 0$ we have*

$$(Q \otimes \mu)^n \Big((T, T') \in X^n \times X^n : \mathcal{R}_P(f_{T,T',\lambda_n}) \leq \mathcal{R}_P + \varepsilon \Big) \to 0 \,,$$

for $n \to \infty$. The same result holds for the SVM with offset if one replaces the condition $n\lambda_n^2 \to \infty$ by the slightly stronger assumption $n\lambda_n^2 / \log n \to \infty$. Finally, for both SVMs it suffices to assume $n\lambda_n^{1+\delta} \to \infty$ for some $\delta > 0$ if one uses a Gaussian RBF kernel.

Sketch of the Proof Let us introduce the shorthand $\nu = Q \otimes \mu$ for the product measure of Q and μ. Moreover, for a measurable function $f : X \to \mathbb{R}$ we define the function $l \odot f : X \times X \to \mathbb{R}$ by

$$l \odot f(x, x') := \frac{1}{1+\rho} l(1, f(x)) + \frac{\rho}{1+\rho} l(-1, f(x')) \,, \qquad x, x' \in X.$$

Furthermore, we write $l \circ f(x, y) := l(y, f(x))$, $x \in X$, $y \in Y$. Then it is easy to check that we always have $\mathbb{E}_\nu l \odot f = \mathbb{E}_P l \circ f$. Analogously, we see $\mathbb{E}_{T \otimes T'} l \odot f = \mathbb{E}_{T \ominus_s T'} l \circ f$, if $T \otimes T'$ denotes the product measure of the empirical measures based on T and T'. Now, using Hoeffding's inequality for ν it is easy to establish a concentration inequality in the sense of Steinwart (2005, Lem. III.5). The rest of the proof is analogous to the steps of Steinwart (2005). ∎

Recall that by Theorem 4 consistency with respect to $\mathcal{R}_P(.)$ is equivalent to consistency with respect to $\mathcal{S}_P(.)$. Therefore we immediately obtain the following corollary

Corollary 13 *Under the assumptions of Theorem 12 both the DLD-SVM with offset and without offset are universally consistent with respect to $\mathcal{S}_P(.)$, i.e. $\mathcal{S}_P(\tilde{f}_{T,\mu,\lambda} + \tilde{b}_{T,\mu,\lambda}) \to 0$ and $\mathcal{S}_P(f_{T,T',\lambda_n}) \to 0$ in probability.*

Remark 14 *We have just seen that our DLD-SVM whose design was based on the plan described in the beginning of this section can learn arbitrary DLD problems. It should be almost clear that*

a similar approach and analysis is possible for many other classification algorithms. This gives a strong justification for the well-known heuristic of adding artificial samples to anomaly detection problems with unlabeled data. However, it is important to note that this justification only holds for the above sampling plan and suitably adjusted classification algorithms, and that other, heuristic sample plans may actually lead to bad learning performance (cf. the second part of Section 5)

4. Experiments

We present experimental results for anomaly detection problems where the set X is a subset of $\mathbb{R}^d$. A total of four different learning algorithms are used to produce functions f which declare the set $\{x : f(x) \leq 0\}$ anomalous. A distinct advantage of the formulation in Section 2 is that it allows us to make *quantitative* comparisons of different functions by comparing estimates of their risk $\mathcal{R}_P(f)$ which can be computed from sample data. In particular consider a data set pair (S, S') where S contains samples drawn from Q and S' contains samples drawn from μ (in what follows (S, S') is either training data, validation data, or test data). Based on Definition 6 we define the empirical risk of f with respect to (S, S') to be

$$\mathcal{R}_{(S,S')}(f) = \frac{1}{(1+\rho)|S|} \sum_{x \in S} I(1, \operatorname{sign} f(x)) + \frac{\rho}{(1+\rho)|S'|} \sum_{x \in S'} I(-1, \operatorname{sign} f(x)). \tag{8}$$

A smaller risk indicates a better solution to the DLD problem. Since the risk $\mathcal{R}_{\mathcal{P}}(\cdot)$ depends explicitly on ρ additional insight into the performance of f can be obtained from the two error terms. Specifically the quantity $\frac{1}{|S|}\sum_{x \in S} I(1, \operatorname{sign} f(x))$ is an estimate of $Q(\{f \leq 0\})$ which we call the *alarm rate* (i.e. the rate at which samples will be labeled anomalous by f), and the quantity $\frac{1}{|S'|}\sum_{x \in S'} I(-1, \operatorname{sign} f(x))$ is an estimate of $\mu(\{f > 0\})$ which we call the *volume* of the predicted normal set. There is an obvious trade-off between these two quantities, i.e. for the optimal solutions for fixed ρ smaller alarm rates correspond to larger volumes and vice versa. Also, from the expression for the risk in Proposition 5 it is clear that for any two functions with the same alarm rate we prefer the function with the smaller volume and vice versa. More generally, when comparing different solution methods it is useful to consider the values of these quantities that are achieved by varying the value of ρ in the design process. Such *performance curves* are presented in the comparisons below.

We consider three different anomaly detection problems, two are synthetic and one is an application in cybersecurity. In each case we define a problem instance to be a triplet consisting of samples from Q, samples from μ, and a value for the density level ρ. We compare four learning algorithms that accept a problem instance and automatically produce a function f: the density level detection support vector machine (DLD–SVM), the one-class support vector machine (1CLASS–SVM), the Gaussian maximum-likelihood (GML) method, the mixture of Gaussians maximum-likelihood (MGML) method.[1] The first is the algorithm introduced in this paper, the second is an algorithm based on the the one-class support vector machine introduced by Schölkopf et al. (2001) and the others (including the Parzen windows method) are based on some of the most common parametric and non-parametric statistical methods for density-based anomaly detection in $\mathbb{R}^d$. Each of the four learning algorithms is built on a core procedure that contains one or more free parameters. The availability of a computable risk estimate makes it possible to determine values for these parameters

1. We also experimented with a Parzen windows method, but do not include the results because they were substantially worse than the other methods in every case.

using a principled approach that is applied uniformly to all four core procedures. In particular this is accomplished as follows in our experiments. The data in each problem instance is partitioned into three pairs of sets; the training sets (T,T'), the validation sets (V,V') and the test sets (W,W'). The core procedures are run on the training sets and the values of the free parameters are chosen to minimize the empirical risk (8) on the validation sets. The test sets are used to estimate performance. We now describe the four learning algorithms in detail.

In the DLD–SVM algorithm we employ the SVM *with offset* described in Section 3 with a Gaussian RBF kernel

$$k(x,x') = e^{-\sigma^2\|x-x'\|^2}.$$

With λ and σ^2 fixed and the expected value $\mathbb{E}_{x\sim\mu} l(-1, f(x)+b)$ in (6) replaced with an empirical estimate based on T' this formulation can be solved using, for example, the `C-SVC` option in the `LIBSVM` software (see Chang and Lin, 2004) by setting $C=1$ and setting the class weights to $w_1 = 1/\big(\lambda|T|(1+\rho)\big)$ and $w_{-1} = \rho/\big(\lambda|T'|(1+\rho)\big)$. The regularization parameters λ and σ^2 are chosen to (approximately) minimize the empirical risk $\mathcal{R}_{(V,V')}(f)$ on the validation sets. This is accomplished by employing a grid search over λ and a combined grid/iterative search over σ^2. In particular, for each value of λ from a fixed grid we seek a minimizer over σ^2 by evaluating the validation risk at a coarse grid of σ^2 values and then performing a Golden search over the interval defined by the two σ^2 values on either side of the coarse grid minimum.[2] As the overall search proceeds the (λ,σ^2) pair with the smallest validation risk is retained.

The 1CLASS–SVM algorithm is based on the one-class support vector machine introduced by Schölkopf et al. (2001). Recall that this method neither makes the assumption that there is a reference distribution μ nor uses T' in the production of its decision function f. Consequently it may be harder to compare the empirical results of the 1CLASS–SVM with those of the other methods in a fair way. Again we employ the Gaussian RBF kernel with width parameter σ^2. The one-class formulation of Schölkopf et al. (2001) contains a parameter ν which controls the size of the set $\{x\in T : f(x)\le 0\}$ (and therefore controls the measure $Q(\{f\le 0\})$ through generalization). With ν and σ^2 fixed a solution can be obtained using the `one-class-SVM` option in the `LIBSVM` software. To use this 1-class algorithm to solve an instance of the DLD problem we determine ν automatically as a function of ρ. In particular both ν and σ^2 are chosen to (approximately) minimize the validation risk using the search procedure described above for the DLD–SVM where the grid search for λ is replaced by a Golden search (over $[0,1]$) for ν.

The GML algorithm produces a function $f = g - t$ where t is an offset and g is a Gaussian probability density function whose mean and inverse covariance are determined from maximum likelihood estimates formed from the training data T (see e.g. Duda et al., 2000). In particular the inverse covariance takes the form $(\Sigma+\lambda I)^{-1}$ where Σ is the maximum likelihood covariance estimate and the regularization term λI is a scaled identity matrix which guarantees that the inverse is well-defined and numerically stable. Once the parameters of g are determined the offset t is chosen to minimize the training risk $\mathcal{R}_{(T,T')}$. The regularization parameter λ is chosen to (approximately) minimize the validation risk by searching a fixed grid of λ values.

The MGML algorithm is essentially the same as the GML method except that g is a mixture of K Gaussians whose maximum likelihood parameter estimates are determined using the Expectation-Maximization (EM) algorithm of Dempster et al. (1977). The same regularization parameter is used

2. If the minimum occurs at more than one grid point or at an end point the Golden search interval is defined by the nearest grid points that include all minimal values.

	Train	Validate	Test
Number of Q samples	1000	500	100,000
Number of μ samples	2000	2000	100,000

λ grid (DLD–SVM/GML/MGML)	1.0, 0.5, 0.1, 0.05, 0.01, ..., 0.0000005, 0.0000001
σ^2 grid (DLD–SVM/1CLASS–SVM)	0.001, 0.01, 0.05, 0.1, 0.5, 1.0, 5.0, 10.0, 100.0

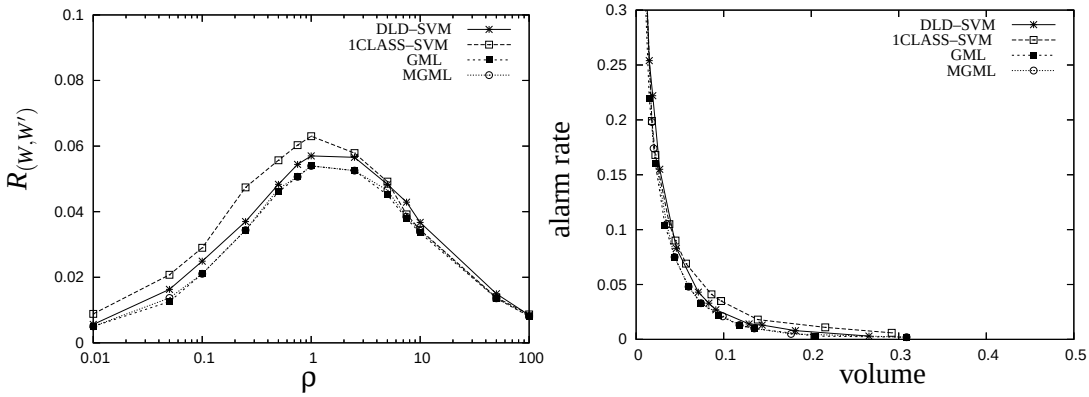
Table 1: Parameters for experiments 1 and 2.

for all inverse covariance estimates and both λ and K are chosen to (approximately) minimize the validation risk by searching a fixed grid of (λ, K) values.

Data for the first experiment are generated using an approach designed to mimic a type of real problem where x is a feature vector whose individual components are formed as linear combinations of raw measurements and therefore the central limit theorem is used to invoke a Gaussian assumption for Q. Specifically, samples of the random variable $x \sim Q$ are generated by transforming samples of a random variable u that is uniformly distributed over $[0,1]^{27}$. The transform is $x = Au$ where A is a 10–by–27 matrix whose rows contain between $m = 2$ and $m = 5$ non-zero entries with value $1/m$ (i.e. each component of x is the average of m uniform random variables). Thus Q is approximately Gaussian with mean $(0.5, 0.5)$ and support $[0,1]^{10}$. Partial overlap in the nonzero entries across the rows of A guarantee that the components of x are partially correlated. We chose μ to be the uniform distribution over $[0,1]^{10}$. Data for the second experiment are identical to the first except that the vector $(0,0,0,0,0,0,0,0,0,1)$ is added to the samples of x with probability 0.5. This gives a bi-modal distribution Q that approximates a mixture of Gaussians. Also, since the support of the last component is extended to $[0,2]$ the corresponding component of μ is also extended to this range. A summary of the data and algorithm parameters for experiments 1 and 2 is shown in Table 1. Note that the test set sizes are large enough to provide very accurate estimates of the risk.

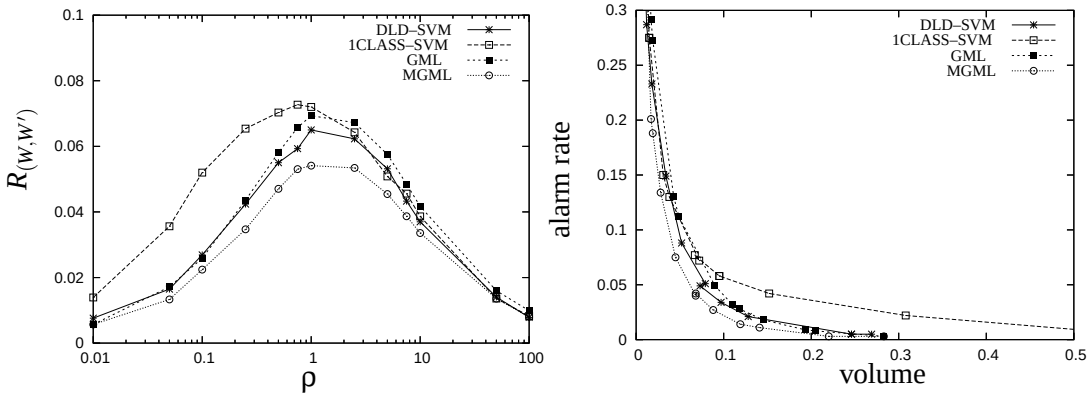
The four learning algorithms were applied for values of ρ ranging from .01 to 100 and the results are shown in Figure 1. Figures 1(a) and 1(c) plot the empirical risk $R_{(W,W')}$ versus ρ while Figures 1(b) and 1(d) plot the corresponding performance curves. Since the data is approximately Gaussian it is not surprising that the best results are obtained by GML (first experiment) and MGML (both experiments). However, for most values of ρ the next best performance is obtained by DLD–SVM (both experiments). The performance of 1CLASS–SVM is clearly worse than the other three at smaller values of ρ (i.e. larger values of the volume), and this difference is more substantial in the second experiment. In addition, although we do not show it, this difference is even more pronounced (in both experiments) at smaller training and validation set sizes. These results are significant because values of ρ substantially larger than one appear to have little utility here since they yield alarm rates that do not conform to our notion that anomalies are rare events. In addition $\rho \gg 1$ appears to have little utility in the general anomaly detection problem since it defines anomalies in regions where the concentration of Q is much larger than the concentration of μ, which is contrary to our premise that anomalies are not concentrated.

The third experiment considers an application in cybersecurity. The goal is to monitor the network traffic of a computer and determine when it exhibits anomalous behavior. The data for this experiment was collected from an active computer in a normal working environment over the course of 16 months. The features in Table 2 were computed from the outgoing network traffic.



(a) Risk curves for $Q \approx$ Gaussian.

(b) Performance curves for $Q \approx$ Gaussian.



(c) Risk curves for $Q \approx$ Gaussian mixture.

(d) Performance curves for $Q \approx$ Gaussian mixture.

Figure 1: Synthetic data experiments.

The averages were computed over one hour time windows giving a total of 11664 feature vectors. The feature values were normalized to the range $[0,1]$ and treated as samples from Q. Thus Q has support in $[0,1]^{12}$. Although we would like to choose μ to capture a notion of anomalous behavior for this application, only the DLD–SVM method allows such a choice. Thus, since both GML and MGML define densities with respect to a uniform measure and we wish to compare with these methods, we chose μ to be the uniform distribution over $[0,1]^{12}$. A summary of the data and algorithm parameters for this experiment is shown in Table 3. Again, we would like to point out that this choice may actually penalize the 1CLASS-SVM since this method is not based on the notion of a reference measure. However, we currently do not know any other approach which treats the 1CLASS-SVM with its special strucure in a fairer way.

The four learning algorithms were applied for values of ρ ranging from .005 to 50 and the results are summarized by the empirical risk curve in Figure 2(a) and the corresponding performance curve in Figure 2(b). The empirical risk values for DLD–SVM and MGML are nearly identical except for $\rho = 0.05$ where the MGML algorithm happened to choose $K = 1$ to minimize the validation risk

Feature Number	Description
1	Number of sessions
2	Average number of source bytes per session
3	Average number of source packets per session
4	Average number of source bytes per packet
5	Average number of destination bytes per session
6	Average number of destination packets per session
7	Average number of destination bytes per packet
8	Average time per session
9	Number of unique destination IP addresses
10	Number of unique destination ports
11	Number of unique destination IP addresses divided by total number of sessions
12	Number of unique destination ports divided by total number of sessions

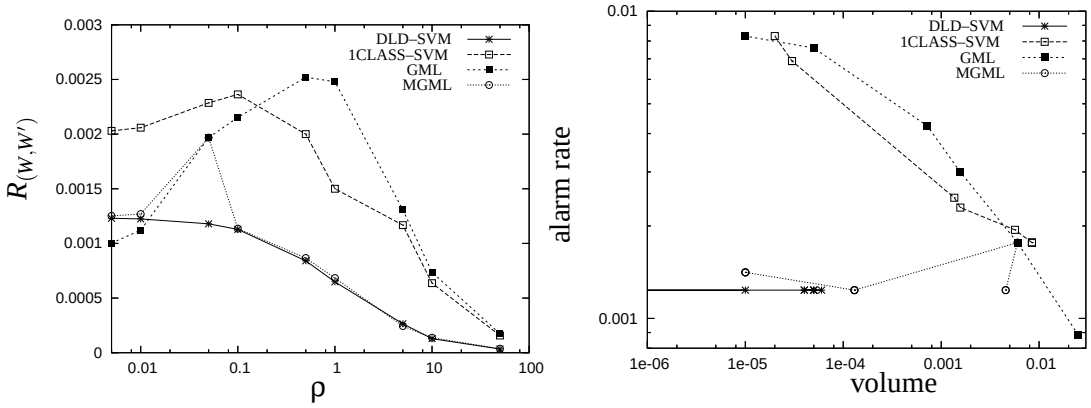
Table 2: Outgoing network traffic features.

	Train	Validate	Test
Number of Q samples	4000	2000	5664
Number of μ samples	10,000	100,000	100,000

λ grid (DLD–SVM/GML/MGML)	0.1, 0.01, 0.001, . . . , 0.0000001
σ^2 grid (DLD–SVM/1CLASS–SVM)	0.001, 0.01, 0.05, 0.1, 0.5, 1.0, 5.0, 10.0, 100.0

Table 3: Parameters for cybersecurity experiment.

(i.e. the MGML and GML solutions are identical at $\rho = 0.05$). Except for this case the empirical risk values for DLD–SVM and MGML are much better than 1CLASS–SVM and GML at nearly all values of ρ. The performance curves confirm the superiority of DLD–SVM and MGML, but also reveal differences not easily seen in the empirical risk curves. For example, all four methods produced some solutions with identical performance estimates for different values of ρ which is reflected by the fact that the performance curves show fewer points than the corresponding empirical risk curves.



(a) Risk curves.

(b) Performance curves.

Figure 2: Cybersecurity experiment.

5. Discussion

A review of the literature on anomaly detection suggests that there are many ways to characterize anomalies (see e.g. Markou and Singh, 2003a,b). In this work we assumed that anomalies are not concentrated. This assumption can be specified by choosing a reference measure μ which determines a density and a level value ρ. The density then quantifies the degree of concentration and the density level ρ establishes a threshold on the degree that determines anomalies. Thus, μ and ρ play key roles in the *definition* of anomalies. In practice the user chooses μ and ρ to capture some notion of anomaly that he deems relevant to the application.

This paper advances the existing state of "density based" anomaly detection in the following ways.

- Most existing algorithms make an implicit choice of μ (usually the Lebesgue measure) whereas our approach allows μ to be any measure that defines a density. Therefore we accommodate a larger class of anomaly detection problems. This flexibility is in particular important when dealing with e.g. categorical data. In addition, it is the key ingredient when dealing with *hidden classification problems*, which we will discuss below.

- Prior to this work there have been no methods known to rigorously estimate the performance based on *unlabeled* data. Consequently, it has been difficult to compare different methods for anomaly detection in practice. We have introduced an empirical performance measure,

namely the empirical classification risk, that enables such a comparision. In particular, it can be used to perform a model selection based on cross validation. Furthermore, the infinite sample version of this empirical performance measure is asymptotically equivalent to the standard performance measure for the DLD problem and under mild assumptions inequalities between them have been obtained.

- By interpreting the DLD problem as a binary classification problem we can use well-known classification algorithms for DLD if we generate artificial samples from μ. We have demonstrated this approach which is a rigorous variant of a well-known heuristic for anomaly detection in the formulation of the DLD-SVM.

These advances have created a situation in which much of the knowledge on classification can now be used for anomaly detection. Consequently, we expect substantial advances in anomaly detection in the future.

Finally let us consider a different learning scenario in which anomaly detection methods are also commonly employed. In this scenario we are interested in solving a binary classification problem given only unlabeled data. More precisely, suppose that there is a distribution ν on $X \times \{-1,1\}$ and the samples are obtained from the *marginal* distribution ν_X on X. Since labels exist but are hidden from the user we call this a *hidden classification problem (HCP)*. Hidden classification problems for example occur in network intrusion detection problems where it is impractical to obtain labels. Obviously, solving a HCP is intractable if no assumptions are made on the labeling process. One such assumption is that one class consists of anomalous, lowly concentrated samples (e.g. intrusions) while the other class reflects normal behaviour. Making this assumption rigorous requires the specification of a reference measure μ and a threshold ρ. Interestingly, when ν_X is absolutely continuous[3] with respect to $\nu(.|y=1)$ solving the DLD problem with

$$\begin{aligned} Q &:= \nu_X \\ \mu &:= \nu(.|y=1) \\ \rho &:= 2\nu(X \times \{1\}) \end{aligned}$$

gives the Bayes classifier for the binary classification problem associated with ν. Therefore, in principle the DLD formalism can be used to solve the binary classification problem. In the HCP however, although information about $Q = \nu_X$ is given to us by the samples, we must rely entirely on first principle knowledge to *guess* μ and ρ. Our inability to choose μ and ρ correctly determines the *model error* that establishes the limit on how well the classification problem associated with ν can be solved with unlabeled samples. This means for example that when an anomaly detection method is used to produce a classifier f for a HCP its anomaly detection performance $\mathcal{R}_P(f)$ with $P := Q \ominus_s \mu$ and $s := \frac{1}{1+\rho}$ may be very different from its hidden classification performance $\mathcal{R}_\nu(f)$. In particular $\mathcal{R}_P(f)$ may be very good, i.e. very close to $\mathcal{R}_P$, while $\mathcal{R}_\nu(f)$ may be very poor, i.e. far above $\mathcal{R}_\nu$. Another consequence of the above considerations is that the common practice of measuring the performance of anomaly detection algorithms on (hidden) binary classification problems is problematic. Indeed, the obtained classification errors depend on the model error and thus they provide an inadequate description how well the algorithms solve the anomaly detection problem.

3. This assumption is actually superfluous by Remark 11.

Furthermore, since the model error is strongly influenced by the particular HCP it is almost impossible to generalize from the reported results to more general statements on the hidden classification performance of the considered algorithms.

In conclusion although there are clear similiarities between the use of the DLD formalism for anomaly detection and its use for the HCP there is also an important difference. In the first case the specification of μ and ρ determines the *definition* of anomalies and therefore there is no model error, whereas in the second case the model error is determined by the choice of μ and ρ.

Acknowledgments

We would like to thank J. Theiler who inspired this work when giving a talk on a recent paper (Theiler and Cai., 2003).

Appendix A. Regular Conditional Probabilities

In this apendix we recall some basic facts on conditional probabilities and regular conditional probabilities. We begin with

Definition 15 *Let $(X, \mathcal{A}, P)$ be a probability space and $\mathcal{C} \subset \mathcal{A}$ a sub-σ-algebra. Furthermore, let $A \in \mathcal{A}$ and $g : (X, \mathcal{C}) \to \mathbb{R}$ be $P_{|\mathcal{C}}$-integrable. Then g is called a conditional probability of A with respect to $\mathcal{C}$ if*

$$\int_C 1_A dP \;=\; \int_C g dP$$

for all $C \in \mathcal{C}$. In this case we write $P(A|\mathcal{C}) := g$.

Furthermore we need the notion of regular conditional probabilities. To this end let $(X, \mathcal{A})$ and $(Y, \mathcal{B})$ be measurable spaces and P be a probability measure on $(X \times Y, \mathcal{A} \otimes \mathcal{B})$. Denoting the projection of $X \times Y$ onto X by π_X we write $\pi_X^{-1}(\mathcal{A})$ for the sub-σ-Algebra of $\mathcal{A} \otimes \mathcal{B}$ which is induced by π_X. Recall, that this sub-σ-Algebra is generated by the collection of the sets $A \times Y$, $A \in \mathcal{A}$. For later purpose, we also notice that this collection is obviously stable against finite intersections. Finally, P_X denotes the marginal distribution of P on X, i.e. $P_X(A) = P(\pi_X^{-1}(A))$ for all $A \in \mathcal{A}$.

Now let us recall the definition of regular conditional probabilities:

Definition 16 *A map $P(\,.\,|\,.\,) : \mathcal{B} \times X \to [0,1]$ is called a* regular conditional probability *of P if the following conditions are satisfied:*

i) *$P(\,.\,|x)$ is a probability measure on $(Y, \mathcal{B})$ for all $x \in X$.*

ii) *$x \mapsto P(B|x)$ is $\mathcal{A}$-measurable for all $B \in \mathcal{B}$.*

iii) *For all $A \in \mathcal{A}$, $B \in \mathcal{B}$ we have*

$$P(A \times B) \;=\; \int_A P(B|x) P_X(dx)\,.$$

Under certain conditions such regular conditional probabilities exist. To be more precise, recall that a topological space is called *Polish* if its topology is metrizable by a complete, separable metric. The following theorem in the book of Dudley (2002, Thm. 10.2.2) gives a sufficient condition for the existence of a regular conditional probability:

Theorem 17 *If Y is a Polish space then a regular conditional probability $P(.\,|\,.) : \mathcal{B} \times X \to [0,1]$ of P exists.*

Regular conditional probabilities play an important role in binary classification problems. Indeed, given a probability measure P on $X \times \{-1,1\}$ the aim in classification is to approximately find the set $\{P(y=1|x) > \frac{1}{2}\}$, where "approximately" is measured by the classification risk.

Let us now recall the connection between conditional probabilities and regular conditional probabilities (see Dudley, 2002, p. 342 and Thm. 10.2.1):

Theorem 18 *If a conditional probability $P(.\,|\,.) : \mathcal{B} \times X \to [0,1]$ of P exists then we P-a.s. have*

$$P(B|x) \;=\; P\big(X \times B|\pi_X^{-1}(\mathcal{A})\big)(x,y)\,.$$

As an immediate consequence of this theorem we can formulate the following "test" for regular conditional probabilities.

Corollary 19 *Let $B \in \mathcal{B}$ and $f : X \to [0,1]$ be $\mathcal{A}$-measurable. Then $f(x) = P(B|x)$ P_X-a.s. if*

$$\int_{A\times Y} f \circ \pi_X dP = \int_{A\times Y} 1_{X\times B} dP$$

for all $A \in \mathcal{A}$.

Proof The assertion follows from Theorem 18, the definition of conditional probabilities and the fact that the collection of the sets $A \times Y$, $A \in \mathcal{A}$ is stable against finite intersections. ■

Appendix B. Proof of Proposition 9

Proof of Proposition 9 By Proposition 2 we have $|2\eta - 1| = \left|\frac{h-\rho}{h+\rho}\right|$ and hence we observe

$$\begin{aligned}\big\{|2\eta-1| \le t\big\} &= \big\{|h-\rho| \le (h+\rho)t\big\}\\ &= \big\{-(h+\rho)t \le h-\rho \le (h+\rho)t\big\}\\ &= \Big\{\frac{1-t}{1+t}\rho \le h \le \frac{1+t}{1-t}\rho\Big\},\end{aligned}$$

whenever $0 < t < 1$.

Now let us first assume that P has Tsybakov exponent $q > 0$ with some constant $C > 0$. Then using

$$\big\{|h-\rho| \le t\rho\big\} \;=\; \big\{(1-t)\rho \le h \le (1+t)\rho\big\} \;\subset\; \Big\{\frac{1-t}{1+t}\rho \le h \le \frac{1+t}{1-t}\rho\Big\}$$

we find

$$P_X\big(\big\{|h-\rho| \le t\rho\big\}\big) \;\le\; P_X\big(\big\{|2\eta-1| \le t\big\}\big) \;\le\; Ct^q\,,$$

which by $P_X = \frac{1}{\rho+1}Q + \frac{\rho}{\rho+1}\mu$ shows that h has ρ-exponent q.

Now let us conversely assume that h has ρ-exponent q with some constant $C > 0$. Then for $0 < t < 1$ we have

$$\begin{aligned} Q\big(\{|h-\rho| \le t\}\big) &= \int_X 1_{\{|h-\rho|\le t\}} h\,d\mu \\ &= \int_{\{h\le 1+\rho\}} 1_{\{|h-\rho|\le t\}} h\,d\mu \\ &\le (1+\rho)\int_{\{h\le 1+\rho\}} 1_{\{|h-\rho|\le t\}}\,d\mu \\ &= (1+\rho)\mu\big(\{|h-\rho| \le t\}\big)\,. \end{aligned}$$

Using $P_X = \frac{1}{\rho+1}Q + \frac{\rho}{\rho+1}\mu$ we hence find

$$P_X\big(\{|h-\rho| \le t\}\big) \;\le\; 2\mu\big(\{|h-\rho| \le t\}\big) \;\le\; 2Ct^q$$

for all sufficiently small $t \in (0,1)$. Let us now define $t_l := \frac{2t}{1+t}$ and $t_r := \frac{2t}{1-t}$. This immediately gives $1-t_l = \frac{1-t}{1+t}$ and $1+t_r = \frac{1+t}{1-t}$. Furthermore, we obviously also have $t_l \le t_r$. Therefore we find

$$\begin{aligned} \left\{\frac{1-t}{1+t}\rho \le h \le \frac{1+t}{1-t}\rho\right\} &= \big\{(1-t_l)\rho \le h \le (1+t_r)\rho\big\} \\ &\subset \big\{(1-t_r)\rho \le h \le (1+t_r)\rho\big\} \\ &= \big\{|h-\rho| \le t_r\rho\big\}\,. \end{aligned}$$

Hence for all sufficiently small $t > 0$ with $t < \frac{1}{1+2\rho}$, i.e. $t_r\rho < 1$, we obtain

$$P_X\big(\{|2\eta-1| \le t\}\big) \;\le\; P_X\big(\{|h-\rho| \le t_r\rho\}\big) \;\le\; 2C(t_r\rho)^q \;\le\; 2C(1+2\rho)^q t^q\,.$$

From this we easily get the assertion. ∎

References

S. Ben-David and M. Lindenbaum. Learning distributions by their density levels: a paradigm for learning without a teacher. *J. Comput. System Sci.*, 55:171–182, 1997.

C. Campbell and K. P. Bennett. A linear programming approach to novelty detection. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 395–401. MIT Press, 2001.

C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, 2004.

A. Cuevas, M. Febrero, and R. Fraiman. Cluster analysis: a further approach based on density estimation. *Computat. Statist. Data Anal.*, 36:441–459, 2001.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. Ser. B (methodology)*, 39:1–38, 1977.

M. J. Desforges, P. J. Jacob, and J. E. Cooper. Applications of probability density estimation to the detection of abnormal conditions in engineering. *Proceedings of the Institution of Mechanical Engineers, Part C—Mechanical engineering science*, 212:687–703, 1998.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley, New York, 2000.

R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2002.

W. Fan, M. Miller, S. J. Stolfo, W. Lee, and P. K. Chan. Using artificial anomalies to detect unknown and known network intrusions. In *IEEE International Conference on Data Mining (ICDM'01)*, pages 123–130. IEEE Computer Society, 2001.

F. González and D. Dagupta. Anomaly detection using real-valued negative selection. *Genetic Programming and Evolvable Machines*, 4:383–403, 2003.

J. A. Hartigan. *Clustering Algorithms*. Wiley, New York, 1975.

J. A. Hartigan. Estimation of a convex density contour in 2 dimensions. *J. Amer. Statist. Assoc.*, 82: 267–270, 1987.

P. Hayton, B. Schölkopf, L. Tarassenko, and P. Anuzis. Support vector novelty detection applied to jet engine vibration spectra. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 946–952. MIT Press, 2001.

S. P. King, D. M. King, P. Anuzis, K. Astley, L. Tarassenko, P. Hayton, and S. Utete. The use of novelty detection techniques for monitoring high-integrity plant. In *IEEE International Conference on Control Applications*, pages 221–226. IEEE Computer Society, 2002.

E. Mammen and A. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27:1808–1829, 1999.

C. Manikopoulos and S. Papavassiliou. Network intrusion and fault detection: a statistical anomaly approach. *IEEE Communications Magazine*, 40:76–82, 2002.

M. Markou and S. Singh. Novelty detection: a review—Part 1: statistical approaches. *Signal Processing*, 83:2481–2497, 2003a.

M. Markou and S. Singh. Novelty detection: a review—Part 2: neural network based approaches. *Signal Processing*, 83:2499–2521, 2003b.

D. W. Müller and G. Sawitzki. Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.*, 86:738–746, 1991.

A. Nairac, N. Townsend, R. Carr, S. King, P. Cowley, and L. Tarassenko. A system for the analysis of jet engine vibration data. *Integrated Computer-Aided Engineering*, 6:53–56, 1999.

W. Polonik. Measuring mass concentrations and estimating density contour clusters—an excess mass aproach. *Ann. Statist.*, 23:855–881, 1995.

B. D. Ripley. *Pattern recognition and neural networks*. Cambridge University Press, 1996.

G. Sawitzki. The excess mass approach and the analysis of multi-modality. In W. Gaul and D. Pfeifer, editors, *From data to knowledge: Theoretical and practical aspects of classification, data analysis and knowledge organization*, Proc. 18th Annual Conference of the GfKl, pages 203–211. Springer, 1996.

B. Schölkopf, J. C. Platt, J. Shawe-Taylor, and A. J. Smola. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471, 2001.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.*, 2:67–93, 2001.

I. Steinwart. Consistency of support vector machines and other regularized kernel machines. *IEEE Trans. Inform. Theory*, 51:128–142, 2005.

I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *Ann. Statist.*, submitted, 2004. `http://www.c3.lanl.gov/~ingo/publications/ann-04a.pdf`.

L. Tarassenko, P. Hayton, N. Cerneaz, and M. Brady. Novelty detection for the identification of masses in mammograms. In *4th International Conference on Artificial Neural Networks*, pages 442–447, 1995.

J. Theiler and D. M. Cai. Resampling approach for anomaly detection in multispectral images. In *Proceedings of the SPIE 5093*, pages 230–240, 2003.

A. B. Tsybakov. On nonparametric estimation of density level sets. *Ann. Statist.*, 25:948–969, 1997.

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32:135–166, 2004.

D. Y. Yeung and C. Chow. Parzen-window network intrusion detectors. In *Proceedings of the 16th International Conference on Pattern Recognition (ICPR'02) Vol. 4*, pages 385–388. IEEE Computer Society, 2002.

H. Yu, J. Hen, and K. C. Chang. PEBL: Web page classification without negative examples. *IEEE. Trans. on Knowledge and Data Engineering*, 16:70–81, 2004.