Relative Novelty Detection

Alex J. Smola
Yahoo! Research
Santa Clara, CA

Le Song
Carnegie Mellon University
Pittsburgh, PA

Choon Hui Teo
Australian National University
Canberra, Australia

Abstract

Novelty detection is an important tool for unsupervised data analysis. It relies on finding regions of low density within which events are then flagged as novel. By design this is dependent on the underlying measure of the space. In this paper we derive a formulation which is able to address this problem by allowing for a reference measure to be given in the form of a sample from an alternative distribution. We show that this optimization problem can be solved efficiently and that it works well in practice.

1 Introduction

Novelty detection is useful in finding events which occur only rarely. The basic premise is that given a set of observations $X = \{x_1, \ldots, x_m\} \subseteq \mathcal{X}$, drawn from some distribution $p(x)$ one wants to find a function h whose value is below some threshold, say 0, only for those observations which can be considered novel. h can then be used to detect unusual activity in computer networks, e.g. for fault or intrusion detection, to supervise industrial processes and machines, or to clean a database. A family of algorithms that has been used with great success are one-class Support Vector Machine style estimators [Schölkopf et al., 2001, Tax and Duin, 1999]. They rely on the idea that regions of high density can be efficiently enclosed in a small set or alternatively efficiently separated by a hyperplane.

Experiments show that this approach outperforms the traditional strategy commonly used in statistics of estimating the density first and subsequently thresholding the density to detect unusual events. This is due to two reasons: density estimators attempt to perform particularly well in regions where the density is *high*. Moreover, considerable computation (the cost may be exponential depending on the density model) needs to be spent on normalizing the density. Both concerns are irrelevant for novelty detection hence one should be able to design algorithms which are immune to them.

While single class SVMs indeed resolve those problems, they suffer from a related issue: while not explicit, single class SVMs depend on the measure of the underlying space. In other words, by transforming the measure of the space we use for estimation, we arrive at significantly different estimates. For many applications, this problem cannot be addressed since we can only make educated guesses in terms of what the measure of the underlying space should be. For some cases, however, we will have data drawn from a reference measure at our disposition (e.g. from related machines in a server center).

In this paper we extend the one-class Support Vector Machine approach to thresholded estimates of likelihood ratios. In fact, we show that we are able to use similar objective function that is applied to single-class SVMs while retaining a convex optimization problem and without the need for intermediate density estimation. The optimization problem remains simple and it only requires a slight modification of the problem posed by Nguyen et al. [2008] to fit our needs. We show that estimation can be carried out by repeated invocation of convex optimization.

Note that our problem is closely related to binary classification between two sets of observations, in particular, the retrieval of particularly characteristic observations, such as Precision@k and the multivariate ranking proposed by Joachims [2005]. There exist subtle differences, though: binary classification is a *symmetric* setting for discrimination between two sets whereas we are interested in addressing the *asymmetric* problem of finding novel instances in one set relative to the

Appearing in Proceedings of the 12^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5. Copyright 2009 by the authors.

other. Secondly, formulations such as Precision@k are not specifically designed for the retrieval of density thresholded observations. We show that in practice our algorithm outperforms [Joachims, 2005] even for retrieval.

2 Novelty Detection

We begin with a nonstandard description of novelty detection by thresholded likelihood maximization along the lines of [Smola et al., 2005]. Denote by $\mathcal{H}$ the space of functions on the domain $\mathcal{X}$. In many cases we will assume that $\mathcal{H}$ is a Reproducing Kernel Hilbert Space with kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and with associated RKHS norm $\|h\|$ for all $h \in \mathcal{H}$. This choice, however, is not essential and one should note that our setting is entirely general in terms of how the complexity of the function $h \in \mathcal{H}$ is measured.

Define a density in the nonparametric exponential family via

$$p(x|h) := \exp\left(h(x) - g[h]\right) \tag{1}$$
$$\text{where } g[h] = \log \int_{x \in \mathcal{X}} \exp h(x) dx.$$

Typically computation of $g[h]$ or its derivative is intractable and only approximations of it may be considered. Nonetheless we can state a general purpose penalized log-likelihood maximization problem via

$$\underset{h}{\text{minimize}} \sum_{i=1}^{m} g[h] - h(x_i) + \lambda\Omega[h] \tag{2}$$

to obtain a good density estimate for the set of observations X. Here $\Omega[h]$ is a (typically convex) regularization functional of h, such as $\frac{1}{2}\|h\|^2$.

2.1 Single Class SVM

Conventionally in novelty detection one would try to obtain an estimate $\hat{h}$ by solving (2). Subsequently one may want to threshold the (unnormalized) log-likelihood $\hat{h}$ at some level ρ to declare all observations novel which satisfy $\hat{h}(x) \leq \rho$. This approach has several disadvantages:

1. Solving (2) is often intractable.
2. We do not need to know $g[h]$ for the purpose of novelty detection, yet it is $g[h]$ which renders the optimization problem intractable.
3. We only need the value of h relative to a threshold ρ rather than its actual value. In particular, we do *not* care about values of h for regions of *high* density but only for all x with $h(x) \leq \rho$.

These three problems can be addressed by replacing $-\log p(x|h)$ by a thresholded likelihood ratio, that is

$$\max\left(0, \log \frac{\exp(\rho - g[h])}{p(x|h)}\right) = \max(0, \rho - h(x)). \tag{3}$$

The latter is exactly the loss function proposed by Schölkopf et al. [2001] in the context of single-class SVMs. The objective function can be understood as follows: we are only interested in the likelihood ratio between $p(x|h)$ and some reference density $\exp(\rho - g[h])$. The normalization of $p(x|h)$, that is $\exp(-g[h])$ is not needed. Finally, we only care about regions where the density $p(x)$ is below a certain threshold, hence the $\max(0, \xi)$ term. Amongst other things, (3) explains the conundrum why the functions estimated with single-class SVM resemble density estimates: they *are* density estimates, albeit only for the low density regions.

2.2 Domain Reparametrization

The optimization problem arising from (3) is convex in h, provided that $\Omega[h]$ is. However, it suffers from a key problem: Assume that we have some diffeomorphism $\eta : \mathcal{Z} \to \mathcal{X}$ which reparametrizes the domain $\mathcal{X}$. In this case the density $p(x|h)$ needs to be rewritten as

$$p(z|h) = p(\eta(z)|h) \left|\partial_z \eta(z)\right|. \tag{4}$$

Here the determinant of the Jacobian $|\partial_z \eta(z)|$ is used to recalibrate the measure with respect to its new parametrization. Unfortunately, when plugging $p(z|h)$ into (3) the expression becomes

$$\max\left(0, \rho - h(\eta(z)) - \log|\partial_z \eta(z)|\right),$$

that is, we are now looking for relative novelty with respect to the new measure $dz = |\partial_z \eta(z)|^{-1} dx$ rather than the original measure dx. This means that regions which before might have been considered novel may cease being considered novel and vice versa due to the offset of $\log|\partial_z \eta(z)|$.

In other words: a simple reparametrization of an *arbitrary* measure on the data domain can skew what is considered an outlier. This is highly undesirable. Imagine a ring-shaped density: a change of parameters from x to $\log x$ would change the novelty score by $\log x$, all other terms being equal which has significant implications on whether the center of the ring-shaped density will be regarded as novel given a finite amount of data. While on simple vectorial domains the Lebesgue measure may be acceptable or it may be possible to make educated guesses, this is next to impossible on more structured domains such as strings, graphs, webpages and other data structures.

2.3 Relative Novelty Detection

One way of addressing this problem is by positing a reference measure q with respect to which it is possible to perform novelty detection. Since this measure undergoes the same transformations as the distribution under consideration the two effects cancel out and our analysis becomes measure invariant. This means that rather than dealing with (3) we estimate

$$f(p(x)/q(x)) := \max(0, \rho - \log p(x)/q(x)) \qquad (5)$$

or rather a witness of the log-likelihood ratio in the relevant range of $\rho \geq \log p(x)/q(x)$. Provided that the reference measure q is known, this leads to a simple modification of the original optimization problem. However, in most cases we will not know q but we may only have a sample $Y := \{y_1, \ldots, y_n\} \subseteq \mathcal{X}$ drawn from the reference q. While we could estimate q based on Y, it is more desirable to obtain a formulation which does not depend on a density estimate but rather on an expansion in terms of X and Y directly.

3 Estimation of f-divergences

The loss function $f(\xi)$ as given by (5) has an important property: its gradient vanishes for typical observations where the log likelihood ratio ξ exceeds ρ. Instead of estimating the ratio directly we resort to a technique proposed by Nguyen et al. [2008].

3.1 Variational Decomposition

Divergences between distributions, say p and q which can be expressed as the expectation over a function of a likelihood ratio can be estimated directly by solving a convex minimization problem which attempts to approximate a surrogate for $\log p(x)/q(x)$. This surrogate can then be used to in our quest to find regions of novel observations.

Definition 1 *Denote by $f : \mathbb{R}^+ \to \mathbb{R}$ a convex function with $f(1) = 0$. Moreover, let p and q be distributions on $\mathcal{X}$ and assume that the Radon Nikodym derivative $\frac{dp}{dq}$ exists. Then define the f-divergence via*

$$D_f(p, q) := \mathbf{E}_{x \sim p(x)} \left[f\left(\tfrac{q(x)}{p(x)}\right)\right]. \qquad (6)$$

This concept dates back to Ali and Silvey [1966] and is commonly referred to as the Csiszár f-divergence. Typically one requires that $f(1) = 0$ to ensure that whenever p and q are identical it follows that $D_f(p, q) = 0$. For instance, by choosing $f(\xi) = -\log \xi$ we obtain the Kullback-Leibler divergence. On the other hand, $f(\xi) = \xi \log \xi$ yields the reverse Kullback-Leibler divergence. In the context of this paper we are primarily interested in $f(\xi) = \max(0, \rho - \log \xi)$ as defined in (5) where $\rho \leq 0$.

Nguyen et al. [2008] propose a variational approach to estimating f-divergences between distributions. Their approach can be summarized by the decomposition:

$$D_f(p, q) = \sup_h \mathbf{E}_q [h(x)] - \mathbf{E}_p [f^*(h(x))]. \qquad (7)$$

Here h is a real valued function on $\mathcal{X}$ and f^* is the Fenchel-Legendre dual of f, defined as

$$f^*(\chi) := \sup_\xi \chi\xi - f(\xi). \qquad (8)$$

This means that we may recast the problem of estimating $D_f(p, q)$ as an optimization problem. Moreover, the solution $\bar{h}$ of (7) provides us with valuable information about the log likelihood ratio itself.

3.2 Properties of the Decomposition

Denote by $\bar{h}$ the solution of (7). By duality f satisfies

$$f(\xi) = \sup_\chi \chi\xi - f^*(\chi). \qquad (9)$$

and consequently $\xi \in (f^*)'(\chi)$. The set-inclusion arises from the fact that f or f^* may not be continuously differentiable, hence we need to deal with subdifferentials. Since $\bar{h}(x)$ needs to satisfy the optimality conditions imposed by (8) pointwise we have [Nguyen et al., 2008]

$$\tfrac{q(x)}{p(x)} \in (f^*)'\left(\bar{h}(x)\right). \qquad (10)$$

To make the relationship to the conventional KL-divergence more explicit let us compute the Fenchel-Legendre dual for both $f_{\mathrm{KL}}(\xi) = -\log \xi$ and $f_{\mathrm{nv}}(\xi) = \max(0, \rho - \log \xi)$. We have

$$f^*_{\mathrm{KL}}(\xi) = \begin{cases} \infty & \text{if } \xi \geq 0 \\ -1 - \log(-\xi) & \text{if } \xi < 0 \end{cases} \qquad (11)$$

$$(f^*_{\mathrm{KL}})'(\xi) = \begin{cases} -\xi^{-1} & \text{if } \xi < 0 \end{cases} \qquad (12)$$

for the KL divergence. Moreover, for novelty detection, we effectively obtain thresholded variants via

$$f^*_{\mathrm{nv}}(\xi) = \begin{cases} \infty & \text{if } \xi > 0 \\ \xi e^\rho & \text{if } \xi \in [-e^{-\rho}, 0] \\ -1 - \rho - \log(-\xi) & \text{if } \xi < -e^{-\rho} \end{cases} \qquad (13)$$

$$(f^*_{\mathrm{nv}})'(\xi) = \begin{cases} [e^\rho, \infty) & \text{if } \xi = 0 \\ e^\rho & \text{if } \xi \in [-e^{-\rho}, 0] \\ -\xi^{-1} & \text{if } \xi < -e^{-\rho} \end{cases} \qquad (14)$$

This means that whenever $\bar{h}(x) \leq -e^{-\rho}$ we are able to infer $q(x) = \frac{p(x)}{-\bar{h}(x)}$. For larger values $\bar{h}(x) > -e^{-\rho}$ we can only infer that $q(x) > e^\rho p(x)$, that is, we can infer that x is not novel in p with respect to q.

3.3 Reparametrization

Clearly, solving (7) outright with respect to h is impossible since we do not have access to p or q but rather only to samples from both distributions. Hence one needs regularization. That is, instead of minimizing (7) one solves the following optimization problem.

$$\underset{h}{\operatorname{minimize}} \frac{1}{n}\sum_{i=1}^{n} f^*(h(y_i)) - \frac{1}{m}\sum_{i=1}^{m} h(x_i) + \lambda\Omega[h] \quad (15)$$

Here $\Omega[h]$ is a regularization term which ensures good generalization properties. Unfortunately the constrained convex optimization problem arising from (15) is heavily constrained. This means that it can be costly to find a feasible initialization. Moreover, it is a *constrained* convex program which makes optimization more costly. An alternative is to change variables in analogy to Nguyen et al. [2008]. We perform the substitution:

$$h(x) = -\exp\left(l(x) - \rho\right) \quad (16)$$

$$f^*_{\mathrm{nv}}(l(x)) = \begin{cases} -e^{l(x)} & \text{if } l(x) \leq 0 \\ -1 - l(x) & \text{if } l(x) > 0 \end{cases} \quad (17)$$

$$(f^*_{\mathrm{nv}})'(l(x)) = \begin{cases} e^{\rho} & \text{if } l(x) \leq 0 \\ e^{\rho - l(x)} & \text{if } l(x) > 0 \end{cases} \quad (18)$$

In analogy to (13) we infer that whenever the risk minimizer $\bar{l}(x) \geq 0$ we have $q(x) = p(x)e^{\rho - \bar{l}(x)}$ whereas for $\bar{l}(x) < 0$ we can only conclude that $q(x) > p(x)e^{\rho}$. This is exactly what is desired for a novelty detector — it should ignore the particular shape of a density in regions where $q(x)/p(x)$ is high and provide a faithful representation of regions where $q(x)/p(x)$ is low. In summary, the overall optimization problem becomes

$$\underset{l}{\operatorname{minimize}} \frac{1}{n}\sum_{i=1}^{n} f^*_{\mathrm{nv}}(l(y_i)) + \frac{1}{m}\sum_{i=1}^{m} e^{l(x_i)-\rho} + \lambda\Omega[l]. \quad (19)$$

The main difference to the reparametrization of Nguyen et al. [2008] is that instead of a linear function $f^*_{\mathrm{KL}}(l(y_i)) = -1 - l(x)$ and $\rho = 0$ we now have an exponential downweighting for observations with $l(y_i) < 0$, while the weight of the observations in x has been rescaled by $e^{-\rho}$.

4 Optimization

Unlike the optimization problem arising in KL divergence estimation, (19) is no longer convex for novelty detection. Nonetheless, we show how a globally optimal solution can be obtained efficiently. Subsequently we describe a kernel expansion and we address the issue of adjusting ρ automatically such that a given fraction of observations is flagged as novel relative to $p(x)$. We will do this by proposing automatic rescaling for ρ similar in spirit to the ν-trick [Schölkopf et al., 1999] in SV regression.

DC Programming Assuming convexity in $\Omega[l]$ the optimization problem (19) is nonetheless nonconvex in l overall due to the term $e^{l(x)}$ for all $l(x) \leq 0$. Such problems can be readily alleviated by means of DC Programming [Dinh and An, 1988], also known as the concave convex procedure in machine learning [Yuille and Rangarajan, 2003]. The basic idea is that for functions which can be decomposed into a sum of convex and concave parts, we can upper bound the latter via

$$F_{\mathrm{concave}}(x) \leq F_{\mathrm{concave}}(x_0) + \langle x - x_0, \partial_x F_{\mathrm{concave}}(x_0)\rangle .$$

This is used to minimize an upper bound on the objective function and successively one obtains better convex upper bounds on the nonconvex objective until the optimization procedure converges to a local minimum.

While in general no convergence guarantees can be given, the procedure is sufficient to guarantee convergence to a *global* minimum for the problem of relative novelty detection (19): the mapping $h(x) = -\exp(l(x) - \rho)$ is strictly *monotonic*. Moreover, the original optimization problem is convex, hence it has only a global minimum. Consequently DC programming will converge to the same global optimum.

Stochastic Gradient Descent Note that instead of DC programming we could also resort to stochastic gradient descent directly on the objective function. In particular, there is no need to receive observations x_i and y_j in a specific order. In this case the updates derived from DC programming and stochastic gradient descent on the *non*convex objective function coincide, as DC programming provides an accurate approximation at the point of expansion.

Kernel expansion We may choose an RKHS norm as regularizer $\Omega[l]$. This allows us to apply to Representer Theorem and to expand l in terms of

$$l(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x) + \sum_{i=1}^{n} \beta_i k(y_i, x).$$

This expansion can be plugged into (19) to obtain an unconstrained optimization problem in α and β.

Adjusting ρ It is not clear, a priori, which value of the threshold ρ we should choose. Picking a large value of ρ corresponds to a rather aggressive choice of novelty which may miss almost all observations. Choosing a threshold that is too small may include an overly large subset of observations y_i which are used in the reference set. Note that $f^*_{\mathrm{nv}}(l(y_i))$ does *not* explicitly depend on ρ. Instead, only the terms dependent on x do via $e^{l(x_i)-\rho}$. That is, we adjust the relative weight of the observations x_i.

Note that the objective function is *convex* in ρ. Since we would like to limit the number of detected items, a large value of ρ is desirable. This is achieved by adding a penalty of

$$\nu\rho$$

to the optimization problem (19). Unfortunately, no direct equivalent of the role of ρ in single class SVM can be found. That said, ρ is effectively adjusted such that the average influence of all observations x_i is limited to ν. This can be seen through the optimality conditions with respect to ρ yielding

$$\nu - \frac{1}{m}\sum_{i=1}^{m} e^{l(x_i)-\rho} = 0. \qquad (20)$$

This condition can be enforced simultaneously to the overall optimization as it only adds an additional variable to a convex subproblem.

5 Experiments

5.1 Image Mosaics

In this experiment, we demonstrate the ability of our relative novelty detector (TKL) in identifying duplicate patches in image mosaics.[1] We constructed two image mosaics consisting of the same set of 256 distinct patches (of size 20 × 20 pixels); these patches are arranged in different orders in the *background* and the *target* images to form mosaics with 16 × 16 blocks (Figure 1(a,b)). Then we created relative novelty in the *target* mosaic by duplicating one patch 4 times. More specifically, we duplicate the first patch of the target mosaic) and place these “novel” patches on the 1st, 11th , 13th, and 15th diagonal entries (Figure 1(b)).

Both mosaics have a resolution of 320 × 320 pixels. We treated each non-overlapping 4 × 4 pixel block as a data point, and used their pixel values (i.e. RGB triples) as the input vector. Then we applied an RBF kernel $k(x, x') = \exp(-s \|x - x'\|^2)$ on the input vectors.[2] The scale parameter s of the kernel was chosen to be the reciprocal of the median distance between x and x'.

We compared the performance of our method (TKL), KL divergence where $p(x)/q(x)$ is estimated (KL), SVM$^{\mathsf{perf}}$ where Precision@k is directly optimized [Joachims, 2005][3], and 1-class SVM [Schölkopf et al., 2001].

Figure 1(c,d,e,f) shows the detection results of various methods. Pixels marked in red indicate the detected novelty (fixed at 1% of the image size) whereas the duplicated patches are highlighted in green frames. We can see that TKL, KL, and SVM$^{\mathsf{perf}}$ successfully identify a significant number of pixels from the duplicated patches whereas conventional 1-class SVM failed totally. Among TKL, KL, and SVM$^{\mathsf{perf}}$, the former two are more preferable since the marked pixels all lie inside the duplicated patches.

5.2 Satellite Image

The experiment in Section 5.1 shows that relative novelty detection works in well-controlled setting. In this experiment, we consider a *real-world* application of relative novelty detection. We examine the performance of the relative novelty detectors in identifying novel objects in a *target* satellite image relative to another *background* satellite image. Both the *target* and *background* images are of different housing areas in Canberra, Australia (Figure 2(a,b)). In the *target* image, the greenish fields (especially the dark green oval) are absent from the *background* image, and hence they are considered as relatively novel. With traditional novelty/outlier detector, the number of greenish pixels in the *target* image may be overly large to make them novel. The aim here is to identify these fields by using relative novelty detection.

Both images have a resolution of 200×200. We treated each non-overlapping 2 × 2 pixel block as a data point, and used its pixel values (i.e. RGB triples) as the input vector. Similar to previous section, we applied an RBF kernel to the input vectors and used the same RBF kernel scale parameter selection procedure.

Again, we compared the performance of TKL, KL, SVM$^{\mathsf{perf}}$, and 1-class SVM. Figure 2(c,d,e,f) shows the novelty (in red; fixed at 5% of the image size) detected by TKL, KL, SVM$^{\mathsf{perf}}$, and 1-class SVM, respectively. In this case, TKL, KL, and SVM$^{\mathsf{perf}}$ detect the oval and some trees in dark green color whereas 1-class SVM produces qualitatively different result.

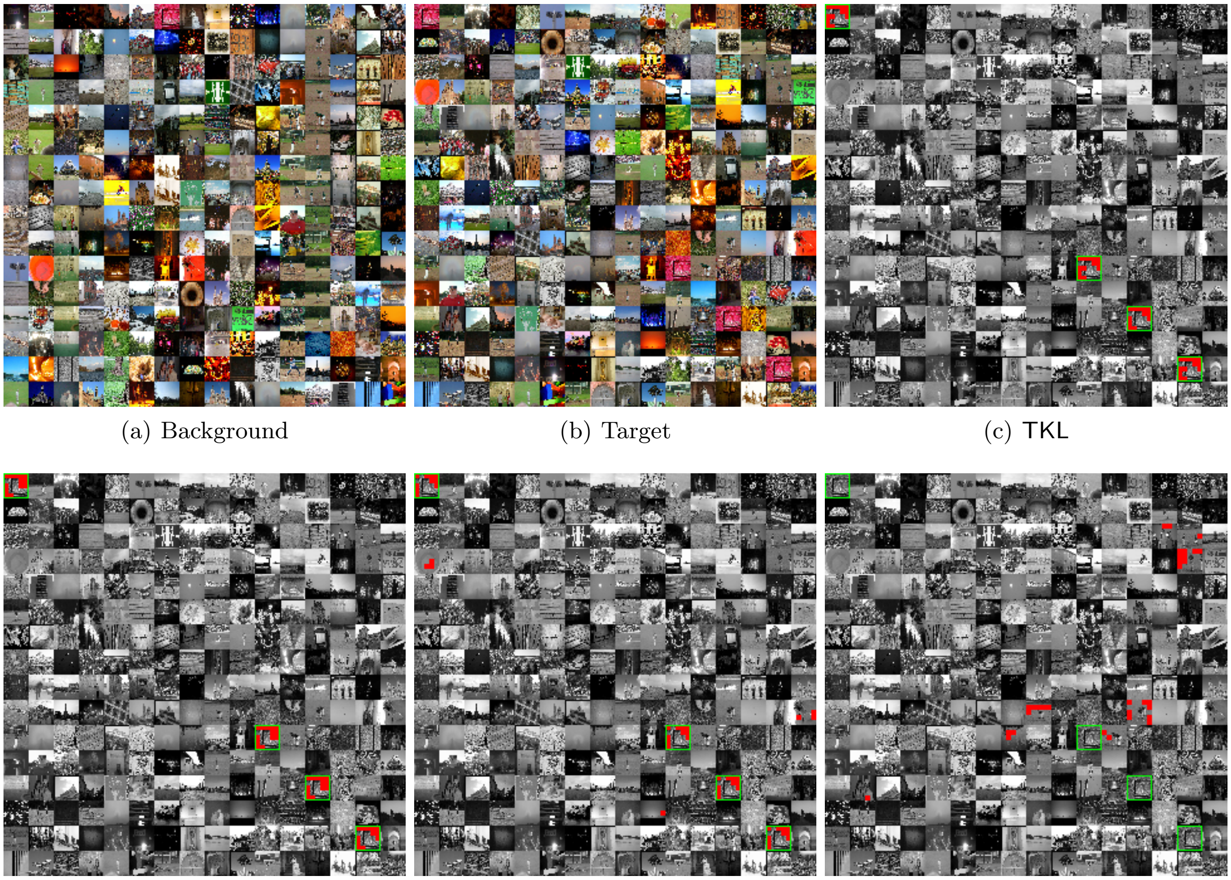
5.3 Novel Digit Detection (USPS)

In Section 5.1 and 5.2, TKL, KLand SVM$^{\mathsf{perf}}$ produces qualitatively similar results. Therefore, in this section, we used the USPS digit dataset to quantitatively ex-

[1] This example is for demonstration purpose; as exact detection of these duplicates can be solved by hashing the patches.

[2] We used random features for the RBF kernel to speed up the computation [Rahimi and Recht, 2008].

[3] The option we used for SVM$^{\mathsf{perf}}$: `-l 4 --b 0 --p r` where r is the fraction of novelty. Note that k is equal to r times the number of observations drawn from $p(x)$



(a) Background

(b) Target

(c) TKL

(d) KL

(e) SVM$^{\mathsf{perf}}$

(f) 1-class SVM

Figure 1: (a) Background and (b) target image mosaic pairs for relative novelty detection. Novel pixels detected by (c) TKL, (d) KL, (e) SVM$^{\mathsf{perf}}$, and (f) 1-class SVM are highlighted in red. True novel patches (or duplicated patches) in the target image are also marked by green frames.

amine the difference between them. The USPS digit dataset contains 7291 training data points and 2007 test data points, each of which is a 16×16 gray level image of handwritten digit (0–9). We further split the training set into two parts: 2/3 as a new training set, and 1/3 as a validation set. The test set remains unchanged.

We will now further synthesize datasets suitable for relative novelty detection. Basically, our procedure first mimics drawing observations from two distributions $s(x)$ and $q(x)$ where $s(x) = q(x)$; then it adds some novelty into the sample from $s(x)$ such that the sample looks as if it was from a new distribution $p(x)$. With this manipulation, we will have $p(x)/q(x)$ large for the novel points, and we try to detect novelty in the *target* disbution $p(x)$ with respect to the *background* distribution $q(x)$.

More specifically, we will treat each of the ten digits (0–9) as novelty once. The detail for creating samples from $s(x)$ and $q(x)$ and adding novelty to obtain $p(x)$ is described below. Note that we will use digit 0 as example and the procedure also applies to other digits:

1. First, put aside all images with digit 0. These are the novel data points.
2. Randomly split the remaining images into 2 equal halves, one half as observations from $s(x)$ and another half from $q(x)$.
3. Add k novel points (digit 0) into the sample from $s(x)$, such that the fraction of novel points is r. This is treated as a sample from $p(x)$.

In our experiments, we used a set of 5 different r ($r = \{2, 4, 6, 8, 10\}$ in percentage). Furthermore, we searched regularization parameter over the range of $10^{[-3:1:3]}$. We chose the best regularization parameter according to the validation set and reported the performance on the test set. The numbers reported in Table 1 are the average Precision@k over 10 repeats of each experiment.

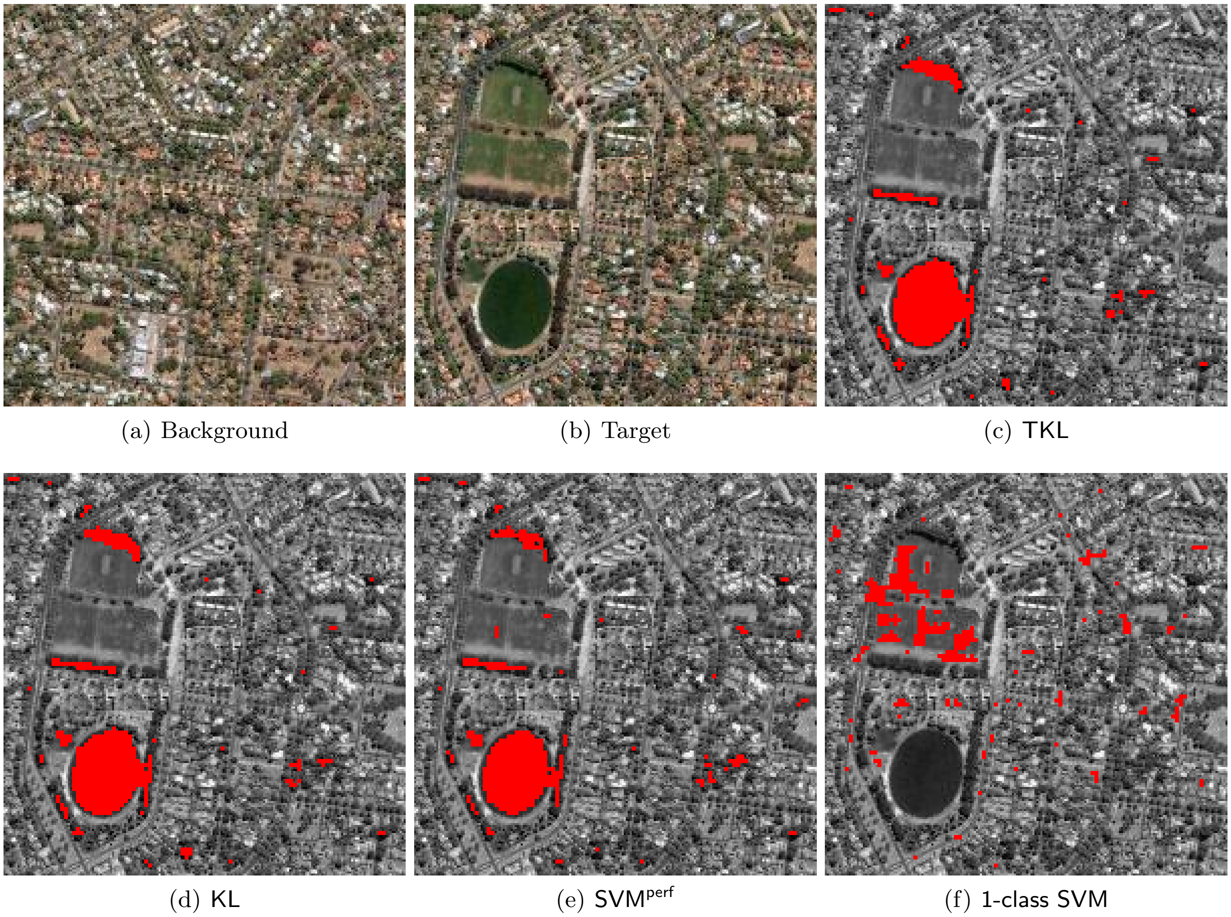


Figure 2: (a) Background and (b) target satellite image pairs for relative novelty detection. Novel pixels detected by (c) TKL, (d) KL, (e) SVM$^{\mathsf{perf}}$, and (f) 1-class SVM are highlighted in red.

A common trend for all three methods is that as the fraction of novelty increases, the accuracy of detecting them also increases. However, TKL and KL are noticably better than SVM$^{\mathsf{perf}}$ in all experiments. In many cases, such as when digit 0 is the novelty and $r = 2\%$, TKL and KL is more than 2 times better than SVM$^{\mathsf{perf}}$.

Between TKL and KL, TKL also wins in a majority of the experiments. Especially when the fraction of novelty is small which makes the novelty hard to detect, the leading margin of TKL over KL is more obvious. For instance, TKL is better than KL for only 3% when digit 8 is the novelty and $r = 10\%$. However, for the same novlety but $r = 2\%$, TKL is nearly 9% better than KL.

6 Discussion

Relation to Classification It may seem surprising that truncated Kullback-Leibler estimation outperforms estimators such as Precision@k which are specifically designed for good retrieval performance of the most relevant terms. We believe that this is due to the fact that structured estimation to compute Precision@k scores [Joachims, 2005] is likely not statistically consistent [Tewari and Bartlett, 2007].

On the other hand, it is easy to see that the Bayes-optimal Precision@k estimate will decide to accept an observation when the log-likelihood ratio between $p(x)$ and $q(x)$ exceeds a given threshold, or in other words, when $\log p(x|y = 1)/p(x|y = -1) \geq c$ for some constant c.[4] Moreover, as established in (10), the minimizer of the variational optimization problem is a rescaled log-likelihood ratio, at least in the region of high values. This is exactly what we need for retrieval. Hence our procedure is consistent.

We believe that this is the main reason why truncated KL estimation outperforms custom designed estimators in retrieval settings. A detailed analysis of the statistical efficiency is subject to future research.

[4] With some abuse of notation we identified $p(x)$ with $p(x|y = 1)$ and $q(x)$ with $p(x|y = -1)$.

Table 1: Relative novelty detection results on USPS digits. The first column is the fraction r of novel data points. The main part of this table reports the average Precision@k (%) when digits 0–9 are treated as novelty respectively. The last column reports the number of times one method outperforms the others for a given r.

r	Method	0	1	2	3	4	5	6	7	8	9	#win
2	KL	86.9±2.5	87.6±1.6	46.7±2.9	45.0±2.9	28.9±3.5	24.4±3.2	43.9±2.5	78.3±1.9	12.8±2.2	15.6±5.0	1
	TKL	92.5±1.6	77.6±3.9	57.8±1.7	51.1±1.4	38.9±3.5	32.2±3.4	57.8±2.4	83.3±1.9	21.7±2.7	22.2±3.7	9
	SVM$^{\text{perf}}$	43.8±4.2	51.2±4.6	24.4±3.7	22.8±2.3	16.1±2.8	10.6±2.8	25.0±3.0	42.8±6.5	10.0±2.0	7.8±4.2	0
4	KL	89.4±0.9	92.8±0.5	60.0±1.3	68.7±1.4	60.3±1.8	55.3±2.2	69.5±1.6	86.3±0.9	44.7±1.6	48.7±3.4	0
	TKL	90.6±1.2	93.1±0.6	66.8±1.3	75.8±1.7	68.4±1.2	60.0±2.7	75.3±1.8	87.1±0.5	52.9±1.8	52.1±3.6	10
	SVM$^{\text{perf}}$	58.2±2.7	83.6±2.0	33.2±1.6	42.1±2.8	38.9±3.3	25.5±2.3	36.1±2.6	64.2±3.8	25.8±2.4	29.2±2.1	0
6	KL	85.0±0.8	93.6±0.3	67.2±0.9	74.7±1.1	73.3±0.7	66.6±1.5	78.6±1.3	86.1±0.5	62.4±1.0	67.9±1.4	0
	TKL	87.1±0.9	94.2±0.4	70.5±0.8	79.8±0.5	78.8±0.9	69.5±1.3	81.7±0.9	88.0±0.6	64.8±0.9	71.9±1.6	10
	SVM$^{\text{perf}}$	61.2±2.2	90.2±1.1	43.9±1.4	52.8±1.5	50.9±1.6	42.2±1.5	53.1±1.6	72.7±1.3	39.7±2.6	47.9±1.6	0
8	KL	84.8±0.8	94.5±0.1	72.4±0.8	76.0±1.1	77.7±0.5	74.0±1.2	82.9±1.0	86.6±0.4	68.0±0.9	75.3±1.1	1
	TKL	87.7±1.0	93.3±0.4	74.7±0.6	78.4±1.2	80.8±0.5	74.8±0.6	84.8±0.7	86.9±0.8	69.2±1.1	78.4±0.9	9
	SVM$^{\text{perf}}$	71.0±1.0	90.1±0.7	54.7±1.1	61.0±0.9	60.0±2.1	52.1±1.9	63.5±1.3	72.6±1.0	50.2±1.7	57.7±2.3	0
10	KL	88.1±0.7	95.7±0.1	78.4±0.5	78.7±0.7	80.2±0.6	77.8±0.8	87.2±0.6	88.3±0.2	71.7±0.6	79.9±1.0	2
	TKL	90.4±0.8	95.3±0.2	79.4±0.7	80.1±0.9	81.8±0.5	78.7±0.8	87.9±0.7	87.1±0.6	74.6±1.0	83.4±0.8	8
	SVM$^{\text{perf}}$	78.1±1.5	92.0±0.4	61.8±1.1	65.4±0.7	65.3±1.6	58.5±1.2	73.0±1.3	75.9±0.7	55.3±1.3	64.6±1.8	0

Stability Our setting may also have advantages in terms of uniform convergence over the plain KL estimation procedure. They arise mainly from the flexible adjustment of the threshold parameter ρ via ν. While for the plain KL estimation procedure the weight of individual observations may grow unbounded very rapidly via $\exp l(x_i)$, the automatic adjustment of ρ ensures that the average weight of all observations is limited to ν as can be seen in (20). A detailed analysis is also a subject of future research.

Implementation We have shown that the method can be easily implemented using existing variational KL estimation code. Moreover, it is also possible to use stochastic gradient descent on the objective function directly, thereby allowing us to apply the method to large collections of data.

7 Conclusion

In this paper, we identify the problem of relative novelty detection and propose a truncated KL divergence formulation for solving this problem. Based on a variational decomposition of trucated KL divergence, we also design an efficient algorithm. We show that our algorithm outperforms KL divergence, SVM$^{\text{perf}}$ and 1-class SVM in a range of different experiments.

References

S.M. Ali and S.D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(1):131–142, 1966.

T. Pham Dinh and L. Hoai An. A D.C. optimization algorithm for solving the trust-region subproblem. *SIAM Journal on Optimization*, 8(2):476–505, 1988.

T. Joachims. A support vector method for multivariate performance measures. In *Proc. Intl. Conf. Machine Learning*, pages 377–384, San Francisco, California, 2005. Morgan Kaufmann Publishers.

X.L. Nguyen, M. Wainwright, and M. Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *NIPS 20*. MIT Press, 2008.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*. MIT Press, Cambridge, MA, 2008.

B. Schölkopf, P. L. Bartlett, A. J. Smola, and R. C. Williamson. Shrinking the tube: a new support vector regression algorithm. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems 11*, pages 330–336, Cambridge, MA, 1999. MIT Press.

B. Schölkopf, J. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13 (7):1443–1471, 2001.

A. J. Smola, S. V. N. Vishwanathan, and T. Hofmann. Kernel methods for missing variables. In R.G. Cowell and Z. Ghahramani, editors, *Proceedings of International Workshop on Artificial Intelligence and Statistics*, pages 325–332. Society for Artificial Intelligence and Statistics, 2005.

D. Tax and R. Duin. Support vector domain description. *Pattern Recognition Letters*, 20(11-13):1191–1199, 1999.

A. Tewari and P.L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–1025, 2007.

A.L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 15:915–936, 2003.