A NEW ONE-CLASS SVM FOR ANOMALY DETECTION

Yuting Chen *Jing Qian* *Venkatesh Saligrama*

Boston University
Boston, MA, 02215

ABSTRACT

Given n i.i.d. samples from some unknown nominal density f_0, the task of anomaly detection is to learn a mechanism that tells whether a new test point η is nominal or anomalous, under some desired false alarm rate α. Popular non-parametric anomaly detection approaches include one-class SVM and density-based algorithms. One-class SVM is computationally efficient, but has no direct control of false alarm rate and usually gives unsatisfactory results. In contrast, some density-based methods show better statistical performance but have higher computational complexity at test time. We propose a novel anomaly detection framework that incorporates statistical density information into the discriminative Ranking SVM procedure. At training stage a ranker is learned based on rankings R of the average k nearest neighbor (k-NN) distances of nominal nodes. This rank $R(x)$ is shown to be asymptotically consistent, indicating how extreme x is with respect to the nominal density. In test stage our scheme predicts the rank $R(\eta)$ of test point η, which is then thresholded to report anomaly. Our approach has much lower complexity than density-based methods, and performs much better than one-class SVM. Synthetic and real experiments justify our idea.

Index Terms— Anomaly Detection, One-class SVM, p-value, Ranking SVM

1. INTRODUCTION

Anomaly detection [1] refers to the problem of identifying statistically significant deviations of data from an expected nominal distribution. These deviated data patterns are often referred to as anomalies or outliers. Anomaly detection has found various applications in many domains such as credit fraud detection, cyber intrusion detection, and video surveillance. Typically based on a training set of nominal examples, anomaly detection techniques design a decision rule such that the detection power is maximized while the false alarm rate is controlled under some prescribed significance level α.

Classical parametric methods [2] for anomaly detection assume some family of the unknown nominal density followed by estimating the parameters from training data. While these methods provide a statistically justifiable solution when the assumptions hold true, they are likely to suffer from model mismatch and lead to poor performance.

During recent years non-parametric approaches have been widely applied to anomaly detection tasks. Such methods make fewer assumptions on the data and tend to be more stable. Typical approaches includes one-class SVM [3] and density-based methods [4, 5, 6, 7, 8]. One-class SVM attempts to find decision boundaries by mapping nominal data to a high-dimensional kernel space and separate them from the origin with maximum margin. While attaining computational efficiency, there is no direct way known to control the false alarm rate. The unawareness of the underlying density could also lead to unsatisfactory performance. Density-based methods such as minimum volume (MV) set estimation [4] and geometric entropy minimization (GEM) [5] involve approximating high-dimensional quantities such as multivariate density or MV set boundaries, which is computationally prohibitive and unreliable. [6, 8, 7] propose to estimate the p-value function based on k nearest neighbor (k-NN) distances within the graph constructed from nominal points. While providing better performance than one-class SVM, these approaches need expensive computations at test stage such as calculating k-NN distance of test point, which makes them inapplicable for tasks requiring real time processing.

In this paper, we propose an novel anomaly detection framework that incorporates statistical density information into the discriminative Ranking SVM procedure. A ranker is learned through ranking algorithms such as Ranking SVM, based on pair-wise comparison information of nominal data points. An input preference pair (x_i, x_j) represents "x_j is more likely to be anomalous than x_i" with respect to the nominal density. These pairs are obtained from the ranking of the average k-NN distance of each sample. We present the asymptotic consistency of this ranking to justify the reliability of preference pairs that are input to a Ranking SVM. During test stages our method estimates the ranks of test points, which are thresholded to report anomaly. Our scheme not only performs much better than one-class SVM, but has much lower complexity than density-based methods.

The rest of the paper is organized as follows. Section 2 describes our anomaly detection algorithm. Section 3 provides

This work was partially supported by NSF Grant 0932114, ONR grant N000141010477, NGA grant HM1582-09-1-0037, NSF grant CCF-0905541, and DHS grant 2008-ST-061-ED0001.

asymptotic analysis. Synthetic and real-world experiments are reported in Section 4. Section 5 concludes the paper.

2. ANOMALY DETECTION ALGORITHM

Let $S = \{x_i, x_2, ..., x_n\}$ be the nominal training set, sampled i.i.d from some unknown multivariate nominal density $f_0(\cdot)$ of d-dimension. Assume that a test sample η is drawn from a mixture of the nominal density $f_0(\cdot)$ and some known anomalous density $f_1(\cdot)$: $f(\eta) = (1-\pi)f_0(\eta) + \pi f_1(\eta)$. For simplicity, we assume that anomaly could happen everywhere with the same probability, i.e., $f_1(\cdot)$ is the uniform distribution. Anomaly detection task can be formulated as a composite hypothesis testing problem: $H_0 : \pi = 0$ (nominal data) versus $H_1 : \pi > 0$ (anomaly).

The aim is to maximize the detection power under a desired false alarm level: $P_F \triangleq \mathcal{P}(decision = H_1|H_0) \leq \alpha$. In [6], it is proven that the uniformly most powerful test for the above detection problem is:

$$D(\eta) = \begin{cases} H_1 & p(\eta) \leq \alpha \\ H_0 & \text{otherwise} \end{cases} \tag{1}$$

where $p(\cdot)$ is the $p-$value function defined as:

$$p(\eta) = \mathcal{P}_0\left(x : \frac{f_1(x)}{f_0(x)} \geq \frac{f_1(\eta)}{f_0(\eta)}\right) = \int_{\{x: f_0(x) \leq f_0(\eta)\}} f_0(x)dx \tag{2}$$

[6, 7] propose to use the rank $R(x)$ of x among all nominal points as an estimate of p-value $p(x)$. This rank $R(x)$ is based on some statistic G, which involves density information at x. During test stage this rank of test point is computed and then thresholded to report anomaly. However, computing $R(\eta)$ has high computational complexity, which can be prohibitive for real-time applications. Motivated by this fact, we propose to learn a discriminative ranker based on the ranks R of all training nominal points. While still maintaining statistical properties of R, the plug-in of discriminative ranking scheme greatly reduces the complexity in test stage.

2.1. Ranking Based Anomaly Detection Algorithm

Our algorithm contains the following steps:

1. Rank Computation

For each training sample x, its rank in set S can be computed as follows:

$$R(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{G(x_i) \geq G(x)\}} \tag{3}$$

where $\mathbb{I}_{\{\cdot\}}$ is the indicator function, and $G(x)$ is the average k-NN distance statistic introduced in [7]:

$$G(x) = \frac{1}{K}\sum_{i=K+1}^{2K} D_{(i)}(x) \tag{4}$$

where $D_{(i)}(x)$ is the i-th nearest neighbor distance of x among $\{x_i, x_2, ..., x_n\}$. It is shown that this statistic outperforms other forms of G such as single k-NN distance or ϵ-neighborhood density in [6]. To compute G, the U-statistic bootstrapping technique [7] can be adopted to reduce variance.

2. Ranker Learning

From Step 1 our training set is now $\{(x_i, R(x_i))\}$. The goal is to learn a ranker r that outputs an ordinal value $r(x_i)$ for x_i. A ranker can be trained by procedures such as Ranking SVM [9] or Ordinal Regression [10].

We adopt the Ranking SVM to train our ranker r. The Ranking SVM takes as input preference pairs (x_i, x_j), generates the following constraints and outputs weight w.

$$(x_i, x_j) \Longleftrightarrow w^T\Phi(x_i) > w^T\Phi(x_j) \tag{5}$$

where $\Phi(x_i)$ is the mapping from X to a high dimensional kernel space. Details about Ranking SVM can be found in [9].

Given n ranks, instead of generating $\binom{n}{2}$ preference pairs, we quantize the ranks to m rank levels $R_q(\cdot) \in \{1, .., m\}$. A preference pair (x_i, x_j) is generated for every $R_q(x_i) > R_q(x_j)$, indicating that x_j is "more anomalous" than x_i. These preference pairs are then fed into Ranking SVM. A ranker r is output along with n sorted ordinal values $r(x_i)$ for all nominal points.

3. Prediction

At test time, the ordinal value for η can be computed:

$$r(\eta) = w^T\Phi(\eta) \tag{6}$$

Then the rank $\hat{R}(\eta)$ is estimated using Eq.(3) by replacing $G(\cdot)$ with $r(\cdot)$. If $\hat{R}(\eta)$ falls under the false alarm level α, anomaly is declared.

Our algorithm is summarized as follows:

Ranking Based Anomaly Detection Algorithm

---

i. Input:

Nominal training data $S = \{x_1, x_2, ..., x_n\}$, desired false alarm level α, and test point η

ii. Training Stage:

(a) Calculate ranks $R(x_i)$ for nominal data x_i, using Eq.(3).

(b) Quantize the ranks $R(x_i)$ into m levels: $R_q(x_i) \in \{1, .., m\}$. Generate preference pairs (x_i, x_j) whenever their quantized levels satisfy $R_q(x_i) > R_q(x_j)$.

(c) Train a ranker $r(\cdot)$ through Ranking SVM.

iii. Testing Stage:

(a) Calculate $r(\eta)$ for test point η according to Eq.(6).

(b) Estimate the rank $\hat{R}(\eta)$ according to Eq.(3), replacing $G(\cdot)$ with $r(\cdot)$.

(c) Declare η as anomalous if $\hat{R}(\eta) \leq \alpha$, otherwise nominal.

2.2. Comparison With State-of-the-art Algorithms

The main advantages of our approach are summarized below:
(1) False alarm control: one-class SVM approach does not have any natural control over the false alarm rate. In fact, as shown in Fig.1 in Sec.4, some part of the receiver operating characteristic (ROC) curve is missing. Our approach outputs the rank of the test point scaling in $[0, 1]$, which is a direct estimate of false alarm probability and is compared with the prescribed level α to report anomaly.
(2) Incorporating density information: Instead of somewhat heuristically separating nominal points from the origin in the kernel space for one-class SVM, our approach learns the ranker based on the ranks of points. These empirical ranks $R(x)$ are shown to converge asymptotically to the p-value function $p(x)$ in Sec.3, whose value indicates to what extent x is likely to be an anomaly. So our approach explicitly incorporates statistical density information, and outperforms one-class SVM as will be shown in Sec.4.
(3) Complexity Reduction over density-based methods: Computing distances from one point to n points takes $O(dn)$. Sorting takes $O(n \log n)$. So our training stage needs $O(n^2 (d + \log n))$, the same as aK-LPE, plus the time to train an SVM ranker. However at test stage, our algorithm only requires $O(sd + \log n)$ ($s \ll n$ being the number of sparse support vectors), a big reduction from $O(n(d + \log n))$ of aK-LPE [7]. Meanwhile, our experiments show little performance degradation from aK-LPE [7] which has full access to k-NN distance information at test stage.

3. ANALYSIS

To justify the reliability of the ranks, we establish the asymptotic consistency of the ranks in this section. Specifically we show that the rank $R(\eta)$ converges to the p-value at η: $p(\eta)$, as $n \to \infty$.

Suppose the nominal density $f = f_0$ satisfies some regularity conditions: f is continuous and lower-bounded on a compact support C: $f(x) \geq f_{min} > 0$. It is smooth, i.e. $||\nabla f(x)|| \leq \lambda$, where $\nabla f(x)$ is the gradient of $f(\cdot)$ at x. Flat regions are not allowed, i.e. $\forall x \in C$, $\forall \sigma > 0$, $\mathcal{P}\{y : |f(y) - f(x)| < \sigma\} \leq M\sigma$, where M is a constant.

Theorem 1. *By choosing K properly, as $n \to \infty$, we have,*

$$|R(\eta) - p(\eta)| \to 0. \tag{7}$$

The proof involves two parts:
(1) Show concentration of $R(\eta)$ around its expectation $\mathbb{E}[R(\eta)]$ through concentration of measure inequality.
(2) Show that $\mathbb{E}[R(\eta)]$ tends to $p(\eta)$ as $n \to \infty$.
We do not present detailed proofs here and refer the readers to [7].
Remark: The p-value function $p(x)$ is the volume outside the level set of f containing x. Thresholding $p(x)$ has been identified as the uniformly most powerful rule for anomaly detection [6]. For any two points x_i and x_j, $p(x_i) > p(x_j)$ indicates the level set containing x_j completely covers that of x_i, or statistically, x_j is more likely to be an anomaly than x_i. Based on this fact and the above theorem, it is reliable that we feed pairs (x_i, x_j) with $R(x_i) > R(x_j)$ into Ranking SVM, to indicate that x_j is "more anomalous" than x_i.

4. EXPERIMENTS

In this section, we compare our approach with the density-based method aK-LPE [7] and the one-class SVM [3] on both synthetic and real-world data sets.

4.1. Implementation Details

In our simulations, the one-class SVM code in lib-SVM [11] and the Ranking SVM package in SVM^{light} [12] are used.

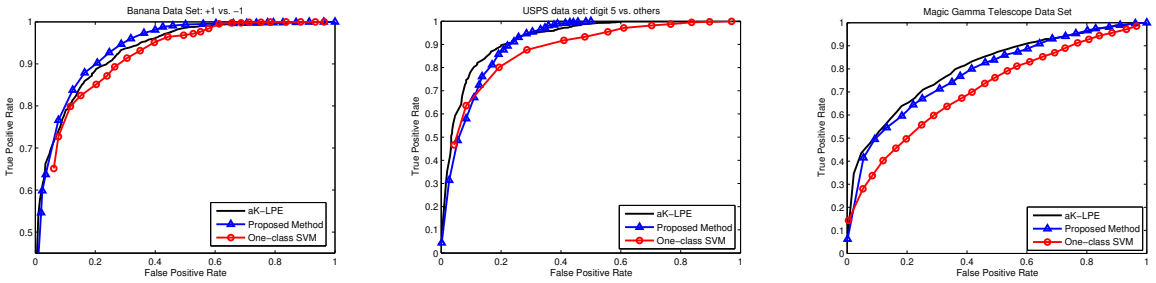
In this section, the Euclidean distance is used as distance metric. The G statistic we adopt is the average k-NN distance with k ranging from 4 to 20. Here we quantize ranks into m=3 levels and generate preference pairs (x_i, x_j) whenever $R_q(x_i) > R_q(x_j)$. In practice, to reduce training time, we can only select preference pairs with significant rank differences $(R_q(x_i) - R_q(x_j) > \tau)$. The RBF kernel is used for one-class SVM and our approach. For the kernel parameter γ, we first vary the regularization parameter ν (One-class SVM) or C (Ranking SVM) for a fixed γ to obtain an empirical ROC curve, and then vary γ to choose the best ROC curve in terms of the maximum area under curve (AUC) principle.

4.2. Synthetic Data sets

We first apply our method for a Gaussian toy problem, where the nominal density is: $f_0 \sim 0.2\mathcal{N}([5; 0], [1, 0; 0, 9]) + 0.8\mathcal{N}([-5; 0], [9, 0; 0, 1])$. To control false alarm at level α, points with $R(\eta)$ no bigger than α is claimed as anomaly. We vary α to obtain the empirical ROC curve. The above procedure is followed for the rest of this section.

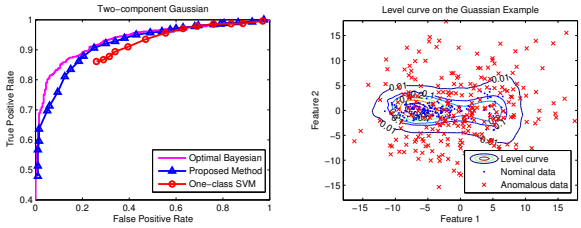
The empirical ROC curves of our method and one-class SVM along with the optimal Bayesian classifier is shown in Fig.1 (a). We can see that our algorithm performs fairly close to the optimal Bayesian classifier and much better than one-class SVM, of which some part of the ROC curve is missing due to lack of false alarm rate control. Fig.1 (b) shows the level curves for the estimated ranks on the test data. We can see that the empirical level curves represent the level sets of the underlying density quite well.

The performance on another synthetic data set "Banana" [13] is shown in Fig.2 (a). As shown in the figure, our detector dominates the one-class SVM on this data set while performs even better than the density-based aK-LPE. The testing time for aK-LPE, oc-SVM and our method are 0.45s, 0.02s and 0.01s (with 67/1598 support vectors) respectively.



(a) (b) (c)

Fig. 2. The ROC curves for one-class SVM and the proposed method on different data sets: (a) Banana, "+1" (nominal) vs. "-1", 2-dim, 1598 training points, 3702 test points (778 nominal). (b) USPS, digit "5" (nominal) vs. others, reduced 64-dim, 600 training points, 1500 test points (300 nominal). (c) Magic, gamma particles (nominal) vs. background, 10-dim, 1500 training points, 4000 test points (1000 nominal)



(a) (b)

Fig. 1. Performance on synthetic data sets: (a) ROC curve on a two-component Gaussian Mixture data. (b) Level sets for the estimated ranks. Here 600 training points, 200 nominal and 1000 anomalous test points are used.

4.3. Real-world data sets

We also apply our anomaly detection method to the USPS digit data set [14] and the Magic gamma telescope data set [14]. For the USPS digit data set, we down-sampled the data to a 64 dimensional space. Here instances of digit 5 are regarded as nominal and instances of other digits as anomaly. The testing time for aK-LPE, oc-SVM and our method are 0.21s, 0.02s and 0.04s (with 254/600 support vectors) respectively. As shown in Fig.2 (b), our algorithm clearly outperforms one-class SVM. In fact, our method achieves 100% true detection rate at a false positive rate of around 50%, while one-class SVM cannot achieve 100% true detection rate until a false positive rate of 80%.

The Magic gamma telescope data set is used to classify high energy gamma particles from cosmic rays in an atmospheric telescope. Images of gamma-initiated photons are recorded by the telescope. 10 attributes of the observed images are used as input features. Here we regard all gamma particles as nominal data and background cosmic rays as anomaly. The testing time for aK-LPE, oc-SVM and our method are 0.42s, 0.02s and 0.01s (with 41/1500 support vectors) respectively. Fig.2 (c) demonstrate that our method outperforms one-class SVM by a large margin, and performs comparable to aK-LPE.

4.4. Discussion

When generating pairwise preference constraints, only 3 rank levels are assigned to the training data, which leads to relatively few preference pairs and thus short training time. However, we observe little to none performance degradation comparing to a much bigger m=9. In fact, it turns out that our algorithm is insensitive to parameters m or γ while one-class SVM degrades significantly when the kernel parameter γ is perturbed from optimum.

5. CONCLUSIONS

In this paper, we propose a novel anomaly detection framework that combines statistical density information with the discriminative ranking procedure. Our scheme learns a ranker by making use of pair-wise orderings of the rank $R(x)$ of training samples. $R(x)$ is the ranking of the average k-NN distance of x within the graph constructed from nominal points, incorporates density information at x, and is shown to be asymptotically consistent. Pairs (x_i, x_j) with $R(x_i) > R(x_j)$, indicating x_j is more likely to be anomalous than x_i, are then fed into Ranking SVM to train the ranker $r(\cdot)$. In test stage our method outputs the rank of test point, which is thresholded to report anomaly. Compared to existing non-parametric methods, our approach not only incorporates statistical density information, leading to better performance than one-class SVM, but also has much lower complexity than density-based methods at test stage due to the simple discriminative ranking scheme. Synthetic and real experiments demonstrate the superiority of our method.

6. REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15:1–15:58, July 2009.

[2] M. Basseville, I.V. Nikiforov, et al., *Detection of abrupt changes: theory and application*, vol. 104, Prentice Hall Englewood Cliffs, NJ, 1993.

[3] B. Schölkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola, and R.C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[4] C.D. Scott and R.D. Nowak, "Learning minimum volume sets," *The Journal of Machine Learning Research*, vol. 7, pp. 665–704, 2006.

[5] A.O. Hero, "Geometric entropy minimization (gem) for anomaly detection and localization," in *Neural Information Processing Systems Conference*, 2006, vol. 19.

[6] M. Zhao and V. Saligrama, "Anomaly detection with score functions based on nearest neighbor graphs," in *Neural Information Processing Systems Conference*, 2009, vol. 22.

[7] J. Qian and V. Saligrama, "New statistic in p-value estimation for anomaly detection," in *Statistical Signal Processing Workshop, IEEE*, Aug. 2012, pp. 393 –396.

[8] K. Sricharan and A.O. Hero III, "Efficient anomaly detection using bipartite k-nn graphs," in *Neural Information Processing Systems Conference*.

[9] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 2002, KDD '02, pp. 133–142, ACM.

[10] R. Herbrich, T. Graepel, and K. Obermayer, "Support vector learning for ordinal regression," in *Artificial Neural Networks, 1999. ICANN 99. Ninth International Conference on (Conf. Publ. No. 470)*. IET, 1999, vol. 1, pp. 97–102.

[11] C. Chang and C. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, May 2011.

[12] T. Joachims, "Advances in kernel methods - support vector learning," chapter Making large-scale support vector machine learning practical, pp. 169–184. MIT Press, Cambridge, MA, USA, 1999.

[13] "Benchmark repository," `http://mldata.org/repository/data/viewslug/banana-ida/`.

[14] A. Frank and A. Asuncion, "UCI machine learning repository," `http://archive.ics.uci.edu/ml`, 2010.