Anomaly Detection with Score functions based on Nearest Neighbor Graphs

Authored by:

Manqi Zhao
Venkatesh Saligrama

Abstract

We propose a novel non-parametric adaptive anomaly detection algorithm for high dimensional data based on score functions derived from nearest neighbor graphs on n-point nominal data. Anomalies are declared whenever the score of a test sample falls below q, which is supposed to be the desired false alarm level. The resulting anomaly detector is shown to be asymptotically optimal in that it is uniformly most powerful for the specified false alarm level, q, for the case when the anomaly density is a mixture of the nominal and a known density. Our algorithm is computationally efficient, being linear in dimension and quadratic in data size. It does not require choosing complicated tuning parameters or function approximation classes and it can adapt to local structure such as local change in dimensionality. We demonstrate the algorithm on both artificial and real data sets in high dimensional feature spaces.

1 Paper Body

Anomaly detection involves detecting statistically significant deviations of test data from nominal distribution. In typical applications the nominal distribution is unknown and generally cannot be reliably estimated from nominal training data due to a combination of factors such as limited data size and high dimensionality. We propose an adaptive non-parametric method for anomaly detection based on score functions that maps data samples to the interval [0, 1]. Our score function is derived from a K-nearest neighbor graph (K-NNG) on n-point nominal data. Anomaly is declared whenever the score of a test sample falls below ? (the desired false alarm error). The efficacy of our method rests upon its close connection to multivariate p-values. In statistical hypothesis testing, p-value is any transformation of the feature space to the interval [0, 1] that induces a uniform distribution on the nominal data. When test samples with p-values smaller than ? are declared as anomalies, false alarm error is less than ?. We develop a novel notion of p-values based on measures of level sets of likelihood ratio functions. Our notion provides a characterization of the optimal

anomaly detector, in that, it is uniformly most powerful for a specified false alarm level for the case when the anomaly density is a mixture of the nominal and a known density. We show that our score function is asymptotically consistent, namely, it converges to our multivariate p-value as data length approaches infinity. Anomaly detection has been extensively studied. It is also referred to as novelty detection [1, 2], outlier detection [3], one-class classification [4, 5] and single-class classification [6] in the literature. Approaches to anomaly detection can be grouped into several categories. In parametric approaches [7] the nominal densities are assumed to come from a parameterized family and generalized likelihood ratio tests are used for detecting deviations from nominal. It is difficult to use parametric approaches when the distribution is unknown and data is limited. A K-nearest neighbor 1

(K-NN) anomaly detection approach is presented in [3, 8]. There an anomaly is declared whenever the distance to the K-th nearest neighbor of the test sample falls outside a threshold. In comparison our anomaly detector utilizes the global information available from the entire K-NN graph to detect deviations from the nominal. In addition it has provable optimality properties. Learning theoretic approaches attempt to find decision regions, based on nominal data, that separate nominal instances from their outliers. These include one-class SVM of Sch?olkopf et. al. [9] where the basic idea is to map the training data into the kernel space and to separate them from the origin with maximum margin. Other algorithms along this line of research include support vector data description [10], linear programming approach [1], and single class minimax probability machine [11]. While these approaches provide impressive computationally efficient solutions on real data, it is generally difficult to precisely relate tuning parameter choices to desired false alarm probability. Scott and Nowak [12] derive decision regions based on minimum volume (MV) sets, which does provide Type I and Type II error control. They approximate (in appropriate function classes) level sets of the unknown nominal multivariate density from training samples. Related work by Hero [13] based on geometric entropic minimization (GEM) detects outliers by comparing test samples to the most concentrated subset of points in the training sample. This most concentrated set is the K-point minimum spanning tree(MST) for n-point nominal data and converges asymptotically to the minimum entropy set (which is also the MV set). Nevertheless, computing K-MST for n-point data is generally intractable. To overcome these computational limitations [13] proposes heuristic greedy algorithms based on leave-one out K-NN graph, which while inspired by K-MST algorithm is no longer provably optimal. Our approach is related to these latter techniques, namely, MV sets of [12] and GEM approach of [13]. We develop score functions on K-NNG which turn out to be the empirical estimates of the volume of the MV sets containing the test point. The volume, which is a real number, is a sufficient statistic for ensuring optimal guarantees. In this way we avoid explicit high-dimensional level set computation. Yet our algorithms lead to statistically optimal solutions with the ability to control false alarm and miss error probabilities. The main features of our anomaly detector are summarized. (1) Like [13] our algorithm scales linearly with dimension and quadratic with

data size and can be applied to high dimensional feature spaces. (2) Like [12] our algorithm is provably optimal in that it is uniformly most powerful for the specified false alarm level, ?, for the case that the anomaly density is a mixture of the nominal and any other density (not necessarily uniform). (3) We do not require assumptions of linearity, smoothness, continuity of the densities or the convexity of the level sets. Furthermore, our algorithm adapts to the inherent manifold structure or local dimensionality of the nominal density. (4) Like [13] and unlike other learning theoretic approaches such as [9, 12] we do not require choosing complex tuning parameters or function approximation classes.

2 Anomaly Detection Algorithm: Score functions based on K-NNG In this section we present our basic algorithm devoid of any statistical context. Statistical analysis appears in Section 3. Let S = {x1 , x2 , ? ? ? , xn } be the nominal training set of size n belonging to the unit cube [0, 1]d . For notational convenience we use ? and xn+1 interchangeably to denote a test point. Our task is to declare whether the test point is consistent with nominal data or deviates from the nominal data. If the test point is an anomaly it is assumed to come from a mixture of nominal distribution underlying the training data and another known density (see Section 3). Let d(x, y) be a distance function denoting the distance between any two points x, y ? [0, 1]d . For simplicity we denote the distances by dij = d(xi , xj). In the simplest case we assume the distance function to be Euclidean. However, we also consider geodesic distances to exploit the underlying manifold structure. The geodesic distance is defined as the shortest distance on the manifold. The Geodesic Learning algorithm, a subroutine in Isomap [14, 15] can be used to efficiently and consistently estimate the geodesic distances. In addition by means of selective weighting of different coordinates note that the distance function could also account for pronounced changes in local dimensionality. This can be accomplished for instance through Mahalanobis distances or as a by product of local linear embedding [16]. However, we skip these details here and assume that a suitable distance metric is chosen. Once a distance function is defined our next step is to form a K nearest neighbor graph (K-NNG) or alternatively an ? neighbor graph (?-NG). K-NNG is formed by connecting each xi to the K closest 2

points {xi1 , ? ? ? , xiK } in S ? {xi }. We then sort the K nearest distances for each xi in increasing order di,i1 ? ? ? ? ? di,iK and denote RS (xi) = di,iK , that is, the distance from xi to its K-th nearest neighbor. We construct ?-NG where xi and xj are connected if and only if dij ? ?. In this case we define NS (xi) as the degree of point xi in the ?-NG. For the simple case when the anomalous density is an arbitrary mixture of nominal and uniform density1 we consider the following two score functions associated with the two graphs K-NNG and ?-NNG respectively. The score functions map the test data ? to the interval [0, 1]. n

K-LPE: p?K (?) =

1X I{RS (?)?RS (xi)} n i=1

?-LPE: p?? (?) =

1X I{NS (?)?NS (xi)} n i=1

(1)

n

(2)

where I{?} is the indicator function. Finally, given a pre-defined significance level ? (e.g., 0.05), we declare ? to be anomalous if p?K (?), p?? (?) ? ?. We call this algorithm Localized p-value Estimation (LPE) algorithm. This choice is motivated by its close connection to multivariate p-values(see Section 3). The score function K-LPE (or ?-LPE) measures the relative concentration of point ? compared to the training set. Section 3 establishes that the scores for nominally generated data is asymptotically uniformly distributed in [0, 1]. Scores for anomalous data are clustered around 0. Hence when scores below level ? are declared as anomalous the false alarm error is smaller than ? asymptotically (since the integral of a uniform distribution from 0 to ? is ?). anomaly detection via K?LPE, n=200, K=6, ?=0.05 5

4

4

3

3

2

2

1

1

0

0

?1

?1

?2

?2

?3

?3

?4

?4

?5

?5

empirical distribution of the scoring function K?LPE 12 nominal data anomaly data 10

empirical density

Bivariate Gaussian mixture distribution 5

level set at ?=0.05

8

6

4

2

labeled as anomaly labeled as nominal

?6 ?6

?4

?2

0
2
?6 ?6
4
?4
?2
0
2
4
0
0
0.2
?=0.05
0.4 0.6 value of K?LPE
0.8
1

Figure 1: Left: Level sets of the nominal bivariate Gaussian mixture distribution used to illustrate the KLPE algorithm. Middle: Results of K-LPE with K = 6 and Euclidean distance metric for m = 150 test points drawn from a equal mixture of 2D uniform and the (nominal) bivariate distributions. Scores for the test points are based on 200 nominal training samples. Scores falling below a threshold level 0.05 are declared as anomalies. The dotted contour corresponds to the exact bivariate Gaussian density level set at level ? = 0.05. Right: The empirical distribution of the test point scores associated with the bivariate Gaussian appear to be uniform while scores for the test points drawn from 2D uniform distribution cluster around zero.

Figure 1 illustrates the use of K-LPE algorithm for anomaly detection when the nominal data is a 2D Gaussian mixture. The middle panel of figure 1 shows the detection results based on K-LPE are consistent with the theoretical contour for significance level ? = 0.05. The right panel of figure 1 shows the empirical distribution (derived from the kernel density estimation) of the score function K-LPE for the nominal (solid blue) and the anomaly (dashed red) data. We can see that the curve for the nominal data is approximately uniform in the interval [0, 1] and the curve for the anomaly data has a peak at 0. Therefore choosing the threshold ? = 0.05 will approximately control the Type I error within 0.05 and minimize the Type II error. We also take note of the inherent robustness of our algorithm. As seen from the figure (right) small changes in ? lead to small changes in actual false alarm and miss levels. ? ? 1 Pn 1 1 When the mixing density is not uniform but, say f1 , the score functions must be modified to p ?K (?) = n i=1 I R (?)f1 (?) ? R (xi)f1 (xi) S S ? ? NS (?) NS (xi) 1 Pn and p ?? (?) = n for the two graphs K-NNG and ?-NNG respectively. i=1 I f (?) ? f (x)
1
1
1
i

3

To summarize the above discussion, our LPE algorithm has three steps: (1) Inputs: Significance level ?, distance metric (Euclidean, geodesic, weighted etc.). (2) Score computation: Construct K-NNG (or ?-NG) based on dij and compute the score function K-LPE from Equation 1 (or ?-LPE from Equation 2). (3) Make Decision: Declare ? to be anomalous if and only if p?K (?) ? ? (or p?? (?) ? ?). Computational Complexity: To compute each pairwise distance requires O(d) operations; and O(n2 d) operations for all the nodes in the training set. In the worst-case computing the K-NN graph (for small K) and the functions RS (?), NS (?) requires O(n2) operations over all the nodes in the training data. Finally, computing the score for each test data requires O(nd+n) operations(given that RS (?), NS (?) have already been computed). Remark: LPE is fundamentally different from non-parametric density estimation or level set estimation schemes (e.g., MV-set). These approaches involve explicit estimation of high dimensional quantities and thus hard to apply in high dimensional problems. By computing scores for each test sample we avoid high-dimensional computation. Furthermore, as we will see in the following section the scores are estimates of multivariate p-values. These turn out to be sufficient statistics for optimal anomaly detection.

3

Theory: Consistency of LPE

A statistical framework for the anomaly detection problem is presented in this section. We establish that anomaly detection is equivalent to thresholding p-values for multivariate data. We will then show that the score functions developed in the previous section is an asymptotically consistent estimator of the p-values. Consequently, it will follow that the strategy of declaring an anomaly when a test sample has a low score is asymptotically optimal. Assume that the data belongs to the d-dimensional unit cube [0, 1]d and the nominal data is sampled from a multivariate density f0 (x) supported on the d-dimensional unit cube [0, 1]d . Anomaly detection can be formulated as a composite hypothesis testing problem. Suppose test data, ? comes from a mixture distribution, namely, f (?) = (1 ? ?)f0 (?) + ?f1 (?) where f1 (?) is a mixing density supported on [0, 1]d . Anomaly detection involves testing the nominal hypotheses H0 : ? = 0 versus the alternative (anomaly) H1 : ? ¿ 0. The goal is to maximize the detection power subject to false alarm level ?, namely, P(declare H1 — H0) ? ?. Definition 1. Let P0 be the nominal probability measure and f1 (?) be P0 measurable. Suppose the likelihood ratio f1 (x)/f0 (x) does not have non-zero flat spots on any open ball in [0, 1]d . Define the p-value of a data point ? as ? ? f1 (?) f1 (x) ? p(?) = P0 x : f0 (x) f0 (?) Note that the definition naturally accounts for singularities which may arise if the support of f0 (?) is a lower dimensional manifold. In this case we encounter f1 (?) ¿ 0, f0 (?) = 0 and the p-value p(?) = 0. Here anomaly is always declared(low score). The above formula can be thought of as a mapping of ? ? [0, 1]. Furthermore, the distribution of p(?) under H0 is uniform on [0, 1]. However, as noted in the introduction there are other such transformations. To build intuition about the above transformation and its utility consider the following example. When

the mixing density is uniform, namely, f1 (?) = U (?) where U (?) is uniform over [0, 1]d , note that ?? = {? — p(?) ? ?} is a density level set at level ?. It is well known (see [12]) that such a density level set is equivalent to a minimum volume set of level ?. The minimum volume set at level ? is known to be the uniformly most powerful decision region for testing H0 : ? = 0 versus the alternative H1 : ? ¿ 0 (see [13, 12]). The generalization to arbitrary f1 is described next. Theorem 1. The uniformly most powerful test for testing H0 : ? = 0 versus the alternative (anomaly) H1 : ? ¿ 0 at a prescribed level ? of significance P(declare H1 — H0) ? ? is: ? H1 , p(?) ? ? ?(?) = H0 , otherwise 4

Proof. We provide the main idea for the proof. First, measure theoretic arguments are used to establish p(X) as a random variable over [0, 1] under both nominal and anomalous distributions. d

d

Next when X ? f0 , i.e., distributed with nominal density it follows that the random variable p(X) ? d

d

U [0, 1]. When X ? f = (1 ? ?)f0 + ?f1 with ? ¿ 0 the random variable, p(X) ? g where g(?) is a monotonically decreasing PDF supported on [0, 1]. Consequently, the uniformly most powerful test for a significance level ? is to declare p-values smaller than ? as anomalies. Next we derive the relationship between the p-values and our score function. By definition, RS (?) and RS (xi) are correlated because the neighborhood of ? and xi might overlap. We modify our algorithm to simplify our analysis. We assume n is odd (say) and can be written as n = 2m + 1. We divide training set S into two parts: S = S1 ? S2 = {x0 , x1 , ? ? ? , xm } ? {xm+1 , ? ? ? , x2m } P 1 We modify ?-LPE to p?? (?) = m ?K (?) xi ?S1 I{NS2 (?)?NS1 (xi)} (or K-LPE to p P 1 I). Now R (?) and R (x) are independent. S2 S1 i xi ?S1 {RS2 (?)?RS1 (xi)} m

=

Furthermore, we assume f0 (?) satisfies the following two smoothness conditions: 1. the Hessian matrix H(x) of f0 (x) is always dominated by a matrix with largest eigenvalue ?M , i.e., ?M s.t. H(x) ? M ?x and ?max (M) ? ?M 2. In the support of f0 (?), its value is always lower bounded by some ? ¿ 0. We have the following theorem. Theorem 2. Consider the setup above with the training data {xi }ni=1 generated i.i.d. from f0 (x). Let ? ? [0, 1]d be an arbitrary test sample. It follows that for a suitable choice K and under the above smoothness conditions, n??

—? pK (?) ? p(?)— ?? 0 almost surely, ?? ? [0, 1]d For simplicity, we limit ourselves to the case when f1 is uniform. The proof of Theorem 2 consists of two steps: n??

? We show that the expectation ES1 [? p? (?)] ?? p(?) (Lemma 3). This result is then exn?? tended to K-LPE (i.e. ES1 [? pK (?)] ?? p(?)) in Lemma 4. n??

? Next we show that p?K (?) ?? ES1 [? pK (?)] via concentration inequality (Lemma 5). q 1/15 3 d , with probability at least 1 ? e??m /2 , Lemma 3 (?-LPE). By picking ? = m? 5d 2?e lm (?) ? ES1 [? p? (?)] ? um (?)

(3)
where 1/15
/2
1/15
/2
lm (?) = P0 {x : (f0 (?) ? ?1) (1 ? ?2) ? (f0 (x) + ?1) (1 + ?2)} ? e??m
um (?) = P0 {x : (f0 (?) + ?1) (1 + ?2) ? (f0 (x) ? ?1) (1 ? ?2)} + e??m ?1 = ?M m?6/5d /(2?e(d + 2)) and ?2 = 2m?1/6 .

Proof. We only prove the lower bound since the upper bound follows along similar lines. By interchanging the expectation with the summation, " # 1 X ES1 [? p? (?)] = ES1 I{NS2 (?)?NS1 (xi)} m xi ?S1 h i 1 X = Exi ES1 xi I{NS2 (?)?NS1 (xi)} m xi ?S1
= Ex1 [PS1 x1 (NS2 (?) ? NS1 (x1))] 5
where the last inequality follows from the symmetric structure of {x0 , x1 , ? ? ? , xm }. n??

Clearly the objective of the proof is to show PS1 x1 (NS2 (?) ? NS1 (x1)) ?? I{f0 (?)?f0 (x1)} . Skipping technical details, this can be accomplishedRin two steps. (1) Note that NS (x1) is a binomial random variable with success probability q(x1) := B? f0 (x1 + t)dt. This relates PS1 x1 (NS2 (?) ? NS1 (x1)) to I{q(?)?q(x1)} . (2) We relate I{q(?)?q(x1)} to I{f0 (?)?f0 (x1)} based on the function smoothness condition. The details of these two steps are shown in the below. Note that NS1 (x1) ? Binom(m, q(x1)). By Chernoff bound of binomial distribution, we have 2
? ? 2mq(x
PS1 x1 (NS1 (x1) ? mq(x1) ? ?) ? e
1)
that is, NS1 (x1) is concentrated around mq(x1). This implies, PS1 x1 (NS2 (?) ? NS1 (x1)) ? I{NS
2
(?)?mq(x1)+?x1 }
2 ?x
1 ? 2mq(x
?e
(4)
1)
We choose ?x1 = q(x1)m? (? will be specified later) and reformulate equation (4) as PS1 x1 (NS2 (?) ? NS1 (x1)) ? I?
NS (?) q(x) 2 ? Vol(B1) mVol(B?) ?
? 2 (1+ m1??)
q(x1)m2??1 2
? e?
(5)
R Next, we relate q(x1)(or B? f0 (x1 + t)dt) to f0 (x1) via the Taylor?s expansion and the smoothness condition of f0 , ? ?R Z ? ? ? 1 ?M ?2 ? ? B? f0 (x1 + t)dt M ? f0 (x1)? ? ? ktk2 dt = (6) ? ? ? Vol(B?) 2 Vol(B?) B? 2d(d + 2) and then equation (5) becomes PS1 x1 (NS2 (?) ? NS1 (x1)) ? I?

? ? NS (?) ?M ?2 2 ? f0 (x1)+ 2d(d+2) mVol(B?)
?
(
2 1+ 1?? m
)
? e?
q(x1)m2??1 2

By applying the same steps to NS2 (?) as equation 4 (Chernoff bound) and equation 6 (Taylor?s explansion), we have with probability at least 1 ? e? Ex1 [PS1 x1 (NS2 (?) ? NS1 (x1))] ? 6

Finally, by choosing ?2 = m? 5d ?
q(?)m2??1 2
,
?? ?? ?? q(x1)m2??1 ? ? ?? ? ?M ?2 ?M ?2 2 2 2 ? f0 (x1)+ 2d(d+2) ?e 1? 1?? 1+ 1?? Px1 f0 (?)? 2d(d+2) m m
d 2?e

and ? = 5/6, we prove the lemma. ? ? Lemma 4 (K-LPE). By picking K = 1 ? 2m?1/6 m2/5 (f0 (?) ? ?1), with probability at least 1/15 1 ? e??m /2 , lm (?) ? ES1 [? pK (?)] ? um (?) (7)

Proof. The proof is very similar to the proof to Lemma 3 and we only give a brief outline here. Now n?? the objective is to show PS1 x1 (RS2 (?) ? RS1 (x1)) ?? I{f0 (?)?f0 (x1)} .The basic idea is to use the result of Lemma 3. To accomplish this, we note that {RS2 (?) ? RS1 (x1)} contains the events {NS2 (?) ? K} ? {NS1 (x1) ? K}, or equivalently {NS2 (?) ? q(?)m ? K ? q(?)m} ? {NS1 (x1) ? q(x1)m ? K ? q(x1)m} (8) By the tail probability of Binomial distribution, the probability of the above two events converges to 1 exponentially fast if K ? q(?)m ¡ 0 and K ? q(x1)m ¿ 0. By using the same two-step bounding techniques developed in the proof to Lemma 3, these two inequalities are implied by K ? m2/5 (f0 (?) ? ?1) ¡ 0 and K ? m2/5 (f0 (x1) + ?1) ¿ 0 ? ? Therefore if we choose K = 1 ? 2m?1/6 m2/5 (f0 (?) ? ?1), we have with probability at least ?1/15 /2 1 ? e??m , ?1/15

PS1 x1 (RS2 (?) ? RS1 (x1)) ? I{(f0 (?)??1)(1??2)?(f0 (x1)+?1)(1+?2)} ? e??m
6
/2

Remark: Lemma 3 and Lemma 4 were proved with specific choices for ? and K. However, they can be chosen in a range of values, but will lead to different lower and upper bounds. We will show in Section 4 via simulation that our LPE algorithm is generally robust to choice of parameter K. P 1 Lemma 5. Suppose K = cm2/5 and denote p?K (?) = m xi ?S1 I{RS2 (?)?RS1 (xi)} . We have 2 m1/5 c2 ? 2 d
? 2?
P0 (—ES1 [? pK (?)] ? p?K (?)— ¿ ?) ? 2e

where ?d is a constant and is defined as the minimal number of cones centered at the origin of angle ?/6 that cover Rd . Proof. We can not apply Law of Large Number in this case because I{RS2 (?)?RS1 (xi)} are correlated. Instead, we

need to use the more generalized concentration-of-measure inequality such P 1 as MacDiarmid?s inequality[17]. Denote F (x0 , ? ? ? , xm) = m xi ?S1 I{RS2 (?)?RS1 (xi)} . From Corollary 11.1 in [18], sup x0 ,??? ,xm ,x0i

—F (x0 , ? ? ? , xi , ? ? ? , xm) ? F (x0 , ? ? ? , x0i , ? ? ? , xn)— ? K?d /m

(9)

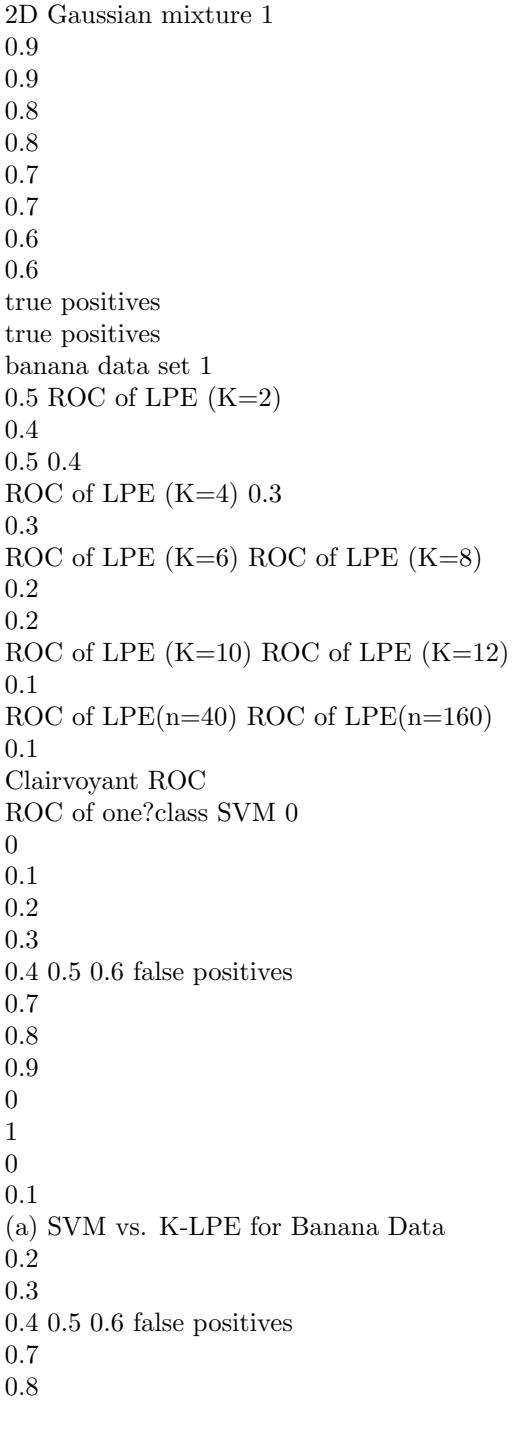
Then the lemma directly follows from applying McDiarmid?s inequality. Theorem 2 directly follows from the combination of Lemma 4 and Lemma 5 and a standard application of the first Borel-Cantelli lemma. We have used Euclidean distance in Theorem 2. When the support of f0 lies on a lower dimensional manifold (say d0 ¡ d) adopting the geodesic metric leads to faster convergence. It turns out that d0 replaces d in the expression for ?1 in Lemma 3.

4

Experiments

First, to test the sensitivity of K-LPE to parameter changes, we run K-LPE on the benchmark dataset Banana [19] with K varying from 2 to 12. We randomly pick 109 points with ?+1? label and regard them as the nominal training data. The test data comprises of 108 ?+1? points and 183 ??1? points (ground truth) and the algorithm is supposed to predict ?+1? data as nominal and ??1? data as anomalous. Scores computed for test set using Equation 1 is oblivious to true f1 density (??1? labels). Euclidean distance metric is adopted for this experiment. To control false alarm at level ?, points with score smaller than ? are predicted as anomaly. Empirical false alarm and true positives are computed from ground truth. We vary ? to obtain the empirical ROC curve. The above procedure is followed for the rest of the experiments in this section. As shown in 2(a), the LPE algorithm is insensitive to K. For comparison we plot the empirical ROC curve of the one-class SVM of [9]. For our OC-SVM implementation, for a fixed bandwidth, c, we obtain the empirical ROC curve by varying ?. We then vary the bandwidth, c, to obtain the best (in terms of AUC) ROC curve. The optimal bandwidth turns out to be c = 1.5. In LPE if we set ? = 0.05 we get empirical F A = 0.06 and for ? = 0.08, empirical F A = 0.09. For OC-SVM we are unaware of any natural way of picking c and ? to control FA rate based on training data. Next, we apply our K-LPE to the problem where the nominal and anomalous data are generated in the following way: ?? ? ? ?? ?? ? ? ?? ? ? ?? 1 1 8 1 0 ?8 1 0 49 0 f0 ? N , + N , , f1 ? N 0, (10) 0 0 9 0 0 9 0 49 2 2 We call ROC curve corresponding to the optimal Bayesian classifier as the Clairvoyant ROC (the red dashed curve in Figure 2(b)). The other two curves are averaged (over 15 trials) empirical ROC curves via LPE. Here we set K = 6 and n = 40 or n = 160. We see that for a relatively small training set of size 160 the average empirical ROC curve is very close to the clairvoyant ROC curve. Finally, we ran LPE on three real-world datasets: Wine, Ionosphere[20] and MNIST US Postal Service (USPS) database of handwritten digits. If there are more than 2 labels in the data set, we artificially regard points with one particular label as nominal and regard the points with other labels as anomalous. For example, for the USPS dataset, we

regard instances of digit 0 as nominal and instances of digits 1, ? ? ? , 9 as anomaly. The data points are normalized to be within [0, 1]d and we 7



(a) SVM vs. K-LPE for Banana Data

0.9
1
(b) Clairvoyant vs. K-LPE

Figure 2: (a) Empirical ROC curve of K -LPE on the banana dataset with K = 2, 4, 6, 8, 10, 12 (with n = 400) vs the empirical ROC curve of one class SVM developed in [9]; (b) Empirical ROC curves of K -LPE algorithm vs clairvoyant ROC curve (f0 is given by Equation 10) for K = 6 and for n = 40 or 160.

1
1
0.9
0.9
0.8
0.8
0.8
0.7
0.7
0.7
0.6
0.6
0.6
0.5 0.4
true positive
1 0.9
true positive
true positive

use geodesic distance [14]. The ROC curves are shown in Figure 3. The feature dimension of Wine is 13 and we apply the ?-LPE algorithm with ? = 0.9 and n = 39. The test set is a mixture of 20 nominal points and 158 anomaly points. The feature dimension of Ionosphere is 34 and we apply the K-LPE algorithm with K = 9 and n = 175. The test set is a mixture of 50 nominal points and 126 anomaly points. The feature dimension of USPS is 256 and we apply the K-LPE algorithm with K = 9 and n = 400. The test set is a mixture of 367 nominal points and 33 anomaly points. In USPS, setting ? = 0.5 induces empirical false-positive 6.1% and empirical false alarm rate 5.7% (In contrast F P = 7% and F A = 9% with ? = 5% for OC-SVM as reported in [9]). Practically we find that K-LPE is more preferable to ?-LPE and as a rule of thumb setting K ? n2/5 is generally effective.

0.5 0.4
0.5 0.4
0.3
0.3
0.3
0.2
0.2
0.2
0.1

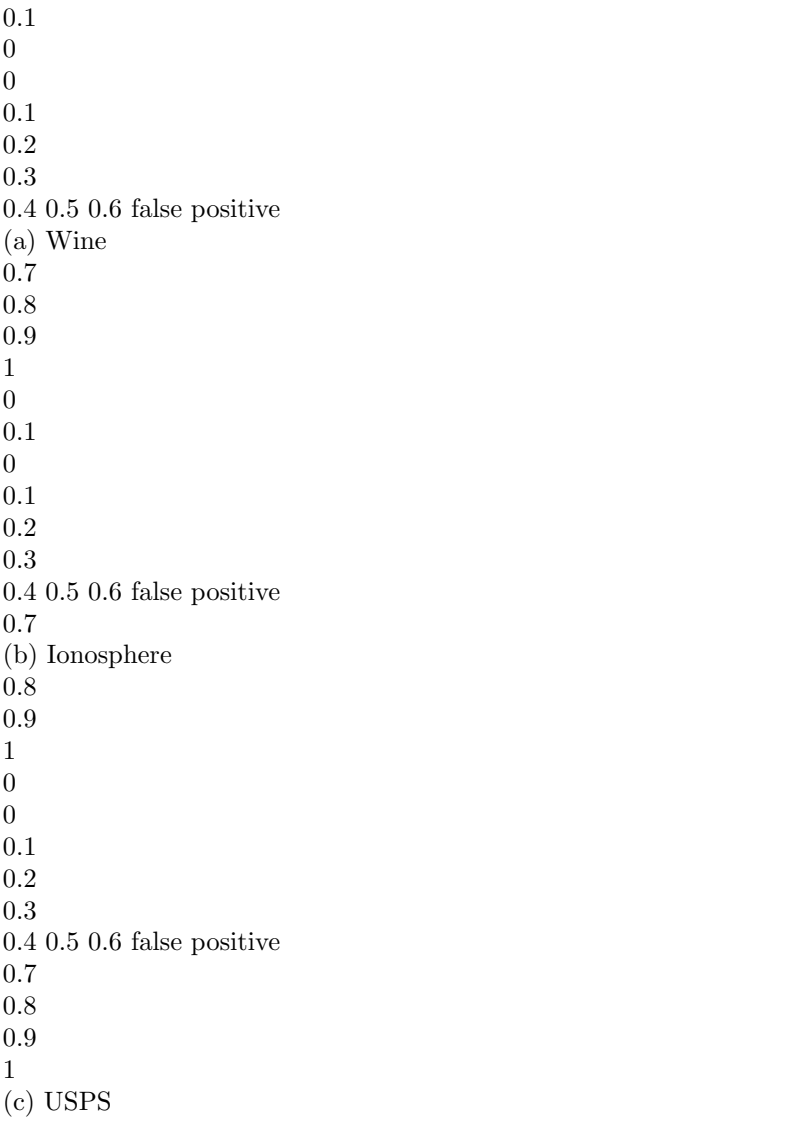


Figure 3: ROC curves on real datasets via LPE; (a) Wine dataset with D = 13, n = 39, ? = 0.9; (b) Ionosphere dataset with D = 34, n = 175, K = 9; (c) USPS dataset with D = 256, n = 400, K = 9.

5 Conclusion

In this paper, we proposed a novel non-parametric adaptive anomaly detection algorithm which leads to a computationally efficient solution with provable optimality guarantees. Our algorithm takes a K-nearest neighbor graph as an input and produces a score for each test point. Scores turn out to be empirical estimates of the volume of minimum volume level sets containing the test point. While minimum volume level sets provide an optimal characterization for

anomaly detection, they are high dimensional quantities and generally difficult to reliably compute in high dimensional feature spaces. Nevertheless, a sufficient statistic for optimal tradeoff between false alarms and misses is the volume of the MV set itself, which is a real number. By computing score functions we avoid computing high dimensional quantities and still ensure optimal control of false alarms and misses. The computational cost of our algorithm scales linearly in dimension and quadratically in data size. 8

2 References

[1] C. Campbell and K. P. Bennett, ?A linear programming approach to novelty detection,? in Advances in Neural Information Processing Systems 13. MIT Press, 2001, pp. 395?401. [2] M. Markou and S. Singh, ?Novelty detection: a review ? part 1: statistical approaches,? Signal Processing, vol. 83, pp. 2481?2497, 2003. [3] R. Ramaswamy, R. Rastogi, and K. Shim, ?Efficient algorithms for mining outliers from large data sets,? in Proceedings of the ACM SIGMOD Conference, 2000. [4] R. Vert and J. Vert, ?Consistency and convergence rates of one-class svms and related algorithms,? Journal of Machine Learning Research, vol. 7, pp. 817?854, 2006. [5] D. Tax and K. R. M? uller, ?Feature extraction for one-class classification,? in Artificial neural networks and neural information processing, Istanbul, TURQUIE, 2003. [6] R. El-Yaniv and M. Nisenson, ?Optimal singl-class classification strategies,? in Advances in Neural Information Processing Systems 19. MIT Press, 2007. [7] I. V. Nikiforov and M. Basseville, Detection of abrupt changes: theory and applications. Prentice-Hall, New Jersey, 1993. [8] K. Zhang, M. Hutter, and H. Jin, ?A new local distance-based outlier detection approach for scattered real-world data,? March 2009, arXiv:0903.3257v1[cs.LG]. [9] B. Sch? olkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. Williamson, ?Estimating the support of a high-dimensional distribution,? Neural Computation, vol. 13, no. 7, pp. 1443?1471, 2001. [10] D. Tax, ?One-class classification: Concept-learning in the absence of counter-examples,? Ph.D. dissertation, Delft University of Technology, June 2001. [11] G. R. G. Lanckriet, L. E. Ghaoui, and M. I. Jordan, ?Robust novelty detection with single-class MPM,? in Neural Information Processing Systems Conference, vol. 18, 2005. [12] C. Scott and R. D. Nowak, ?Learning minimum volume sets,? Journal of Machine Learning Research, vol. 7, pp. 665?704, 2006. [13] A. O. Hero, ?Geometric entropy minimization(GEM) for anomaly detection and localization,? in Neural Information Processing Systems Conference, vol. 19, 2006. [14] J. B. Tenenbaum, V. de Silva, and J. C. Langford, ?A global geometric framework fo nonlinear dimensionality reduction,? Science, vol. 290, pp. 2319?2323, 2000. [15] M. Bernstein, V. D. Silva, J. C. Langford, and J. B. Tenenbaum, ?Graph approximations to geodesics on embedded manifolds,? 2000. [16] S. T. Roweis and L. K. Saul, ?Nonlinear dimensionality reduction by local linear embedding,? Science, vol. 290, pp. 2323?2326, 2000. [17] C. McDiarmid, ?On the method of bounded differences,? in Surveys in Combinatorics. University Press, 1989, pp. 148?188. [18] L. Devroye, L. Gy?orfi, and G.

Lugosi, A Probabilistic Theory of Pattern Recognition. New York, Inc., 1996.
Cambridge
Springer Verlag
[19] ?Benchmark repository.? [Online]. Available: http://ida.first.fhg.de/projects/bench/benchmarks.ht
[20] A. Asuncion and D. J. Newman, ?UCI machine learning repository,? 2007.
[Online]. Available: http://www.ics.uci.edu/?mlearn/MLRepository.html
9