Kernel-Based Nonparametric Anomaly Detection

Shaofeng Zou
Dept of EECS
Syracuse University
Email: szou02@syr.edu

Yingbin Liang
Dept of EECS
Syracuse University
Email: yliang06@syr.edu

H. Vincent Poor
Dept of EE
Princeton University
poor@princeton.edu

Xinghua Shi
Dept of Bioinformatics and Genomics
University of North Carolina at Charlotte
xshi3@uncc.edu

***Abstract*—An anomaly detection problem is investigated, in which there are totally n sequences, with s anomalous sequences to be detected. Each normal sequence contains m independent and identically distributed (i.i.d.) samples drawn from a distribution p, whereas each anomalous sequence contains m i.i.d. samples drawn from a distribution q that is distinct from p. The distributions p and q are assumed to be unknown a priori. The scenario with a reference sequence generated by p is studied. Distribution-free tests are constructed using maximum mean discrepancy (MMD) as the metric, which is based on mean embeddings of distributions into a reproducing kernel Hilbert space (RKHS). It is shown that as the number n of sequences goes to infinity, if the value of s is known, then the number m of samples in each sequence should be of order $\mathcal{O}(\log n)$ or larger in order for the developed tests to consistently detect s anomalous sequences. If the value of s is unknown, then m should be of order strictly larger than $\mathcal{O}(\log n)$. The computational complexity of all developed tests is shown to be polynomial. Numerical results demonstrate that these new tests outperform (or perform as well as) tests based on other competitive traditional statistical approaches and kernel-based approaches under various cases.**

I. Introduction

In this paper, we study an anomaly detection problem (see Figure 1), in which there are n sequences in total, out of which s sequences are anomalous. Each normal sequence consists of m independent and identically distributed (i.i.d.) samples drawn from a distribution p, whereas each anomalous sequence contains i.i.d. samples drawn from a distribution q that is distinct from p. The distributions p and q are assumed to be unknown a priori. Instead, a reference data sequence consisting of i.i.d. samples generated from p is available. This is reasonable because a normal sequence of samples from p is easy to collect in typical applications. (The study of the scenario without a reference sequence is treated in an extended version of this work [1]). The goal is to build distribution-free tests to detect the s anomalous data sequences generated by q out of all data sequences.

Such a problem is very useful in many applications. For example, as studied in [2], in cognitive wireless networks, output signals follow different distributions p or q depending on whether the channel is busy or vacant. A major issue in such a network is to identify vacant channels out of a large number of busy channels based on their corresponding output signals in order to utilize vacant channels for improving spectral efficiency.

We note that in the model here each data point contains a sequence of data samples drawn from one distribution. This is different from the typical anomaly or outlier detection problems studied in machine learning [3], [4], in which each data point contains only one sample. The parametric model of the problem has been well studied, e.g., [2], which assumes that the distributions of p and q are known a priori and can be exploited for detection. However, the nonparametric model which assumes that the distributions p and q are unknown and arbitrary, has been less well explored. Recently, Li, Nitinawarat, and Veeravalli proposed divergence-based generalized likelihood tests in [5], and characterized the error decay exponents of these tests. However, their tests utilize empirical distributions of p and q, and hence are applicable only to discrete distributions with finite alphabets.

In this paper, we study the nonparametric model, in which the distributions p and q can be continuous and arbitrary. A number of statistical approaches and tools may be applied to solve this problem. A natural approach, e.g., the FR-Smirnov test [6], is to first estimate the distributions based on data samples, and then compare the estimated distributions for anomaly detection. Such an approach typically does not perform very well, because the error in estimating the distributions can propagate to the anomaly detection step. Some traditional statistical approaches such as the t-test, FR-Wolf test [6], and Hall test [7] do not require distribution estimation as an intermediate step, and can be applied to solve this problem. However, the t-test and FR-wolf test do not perform well for arbitrary distributions. The Hall test has high computational complexity. More recently, kernel-based approaches such as the kernel density ratio (KDR) test [8] and kernel Fisher discriminant analysis (KFDA) test [9] have been developed, which use kernels to estimate certain distance metrics between two distributions. In particular, the KDR test uses kernels to estimate the ratio between two probability densities and then further estimates the divergence between the two probability distributions. In this paper, our approach introduced below falls into the class of kernel-based approaches.We demonstrate that our tests outperform or equal those tests mentioned above under various test cases.

More specifically, our approach adopts the emerging technique based on mean embedding of distributions into a reproducing kernel Hilbert space (RKHS) [10], [11]. The idea is to map probability distributions into an RKHS with an associated kernel such that distinguishing the two probabilities can be carried out by distinguishing their corresponding embeddings in the RKHS. Such an approach is justified by the fact that mapping of distributions into an RKHS is injective (i.e., one-to-one) for certain kernels including Gaussian and Laplace kernels as shown in [12] and [11]. Since an RKHS naturally carries a distance metric, mean embeddings of distributions can be compared easily based on their distances in the RKHS using the samples of distributions. Such a metric is referred to as the *maximum mean discrepancy (MMD)* as introduced in [13]. A major advantage of MMD-based approaches is that MMD can be easily estimated based on samples, and hence leads to low complexity tests.

In this paper, we apply MMD as a metric to construct our tests for detecting data sequences generated by the anomalous distribution. We are interested in the large data regime, in which the total number n of data sequences goes to infinity. It is clear that as the total number n of sequences becomes large (and possibly the number s of anomalous data sequences also becomes large), it becomes increasingly challenging to consistently detect all anomalous sequences. It is then necessary that the number m of samples in each data sequence correspondingly enlarge in order to more accurately detect anomalous sequences. Hence, there is a tension between (n, s) and m in the asymptotic regime as n goes to infinity for guaranteeing correct detection. This also differentiates our work from the study in [5], in which n and s are assumed to be fixed and only the number of samples becomes large.

We summarize our main contributions as follows. We construct MMD-based distribution-free consistent tests, which enjoy low computational complexity (which is polynomial as n increases) and superior performance (as compared to other tests). We study the scenario with a reference data sequence generated by p, and build distribution-free tests for the two cases with and without knowledge of the number s of anomalous sequences. From the performance of the tests, lack of information about s results in an order-level increase in m needed for consistent detection. We not only provide numerical results to demonstrate our theoretical assertions but also compare our MMD-based tests with other competitive tests. Our numerical results demonstrate that the MMD-based test is the best performing test (or among the best performing tests) under various experimental cases. We note that due to space limitations, we omit the proofs of the theorems here. The details can be found in [1].

II. Problem Statement and Preliminaries on MMD

A. Problem Statement

We study an anomaly detection problem (see Figure 1), in which there are in total n data sequences denoted by Y_k for $1 \leq k \leq n$. Each data sequence Y_k consists of m i.i.d. samples $y_{k1}, \ldots, y_{km}$ drawn from either a distribution p or an anomalous distribution q, where $p \neq q$. In the sequel, we use the notation $Y_k := (y_{k1}, \ldots, y_{km})$. We assume that the distributions p and q are arbitrary and unknown a priori. Instead, a reference data sequence X is assumed to be available a priori, which contains i.i.d. samples $(x_1, \ldots, x_m)$ generated from the distribution p. We use the notation $X := (x_1, \ldots, x_m)$.

We assume that s out of n data sequences are anomalous, i.e., are generated by the anomalous distribution q. We study both cases with the value of s known and unknown a priori, respectively. We are interested in the asymptotic regime, in which the number n of data sequences goes to infinity. We assume that the number s of anomalous sequences satisfies $\frac{s}{n} \to \alpha$ as $n \to \infty$, where $0 \leq \alpha \leq 1$. This includes the following three cases: (1) s is fixed as $n \to \infty$; (2) $s \to \infty$, but $\frac{s}{n} \to 0$ as $n \to \infty$; and (3) $\frac{s}{n}$ approaches a positive constant, which is less than or equal to 1. Some of our results are applicable to the case with $s = 0$, i.e., the null hypothesis in which there is no anomalous sequence. We will comment on such a case when the corresponding results are presented. In this paper, $f(n) = \mathcal{O}(g(n))$ denotes that $f(n)/g(n)$ converges to a constant as $n \to \infty$.

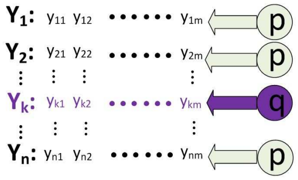

Y_1: y_{11} y_{12} ...... y_{1m} p
Y_2: y_{21} y_{22} ...... y_{2m} p
Y_k: y_{k1} y_{k2} ...... y_{km} q
Y_n: y_{n1} y_{n2} ...... y_{nm} p

Fig. 1. An anomaly detection model with data sequences generated by distribution p and anomalous distribution q.

We next define the probability of error as the performance measure of tests. We let $\mathcal{I}$ denote the set that contains indices of all anomalous data sequences. Hence, the cardinality $|\mathcal{I}| = s$. We let $\hat{\mathcal{I}}^n$ denote a sequence of index sets that contain indices of all anomalous data sequences claimed by a corresponding sequence of tests.

Definition 1. *A sequence of tests are said to be consistent if*

$$\lim_{n\to\infty} P_e = \lim_{n\to\infty} P\{\hat{\mathcal{I}}^n \neq \mathcal{I}\} = 0. \tag{1}$$

We note that the limit in the above definition in fact corresponds to the asymptotic regime, in which m scales fast enough as n goes to infinity in order to guarantee asymptotically small probability of error. Such a regime is also applicable to the following definition.

Definition 2. *A sequence of tests are said to be exponentially consistent if*

$$\liminf_{m\to\infty} -\frac{1}{m}\log P_e = \liminf_{m\to\infty} -\frac{1}{m}\log P\{\hat{\mathcal{I}}^n \neq \mathcal{I}\} > 0. \tag{2}$$

In this paper, our goal is to construct distribution-free tests for detecting anomalous sequences, and characterize the scaling behavior of m with n (and possibly s) so that the developed tests are consistent (and possibly exponentially consistent).

B. Introduction of MMD

In this subsection, we briefly introduce the idea of mean embedding of distributions into an RKHS [10], [11] and the metric of MMD. Suppose $\mathcal{P}$ is a class of probability distributions, and suppose $\mathcal{H}$ is an RKHS with an associated kernel $k(\cdot, \cdot)$ (see [14] for an introduction to RKHS theory). We define a mapping from $\mathcal{P}$ to $\mathcal{H}$ such that each distribution $p \in \mathcal{P}$ is mapped into an element in $\mathcal{H}$ as follows:

$$\mu_p(\cdot) = \mathbb{E}_p[k(\cdot, x)] = \int k(\cdot, x)dp(x).$$

Here, $\mu_p(\cdot)$ is referred to as the *mean embedding* of the distribution p into the Hilbert space $\mathcal{H}$.

It has been shown in [12] and [11] that for many RKHSs such as those associated with Gaussian and Laplace kernels, the mean embedding is injective, such that each $p \in \mathcal{P}$ is mapped to a unique element $\mu_p \in \mathcal{H}$. In order to distinguish between two distributions p and q, [13] introduced the following measure of MMD based on the mean embeddings μ_p and μ_q of p and q in an RKHS $\mathcal{H}$:

$$\mathrm{MMD}[p,q] := \|\mu_p - \mu_q\|_{\mathcal{H}}. \tag{3}$$

Due to the reproducing property of the kernel, it can be easily shown that

$$\mathrm{MMD}^2[p,q]{=}\mathbb{E}_{x,x'}[k(x,x')] - 2\mathbb{E}_{x,y}[k(x,y)] + \mathbb{E}_{y,y'}[k(y,y')],$$

where x and x' are independent but have the same distribution p, and y and y' are independent but have the same distribution q. An unbiased estimator of $\mathrm{MMD}^2[p,q]$ based on l_1 samples of X and l_2 samples of Y is given as follows:

$$\begin{aligned}\mathrm{MMD}_u^2[X,Y] &= \frac{1}{l_1(l_1-1)}\sum_{i=1}^{l_1}\sum_{j\neq i}^{l_1} k(x_i,x_j) \\ &+ \frac{1}{l_2(l_2-1)}\sum_{i=1}^{l_2}\sum_{j\neq i}^{l_2} k(y_i,y_j) - \frac{2}{l_1 l_2}\sum_{i=1}^{l_1}\sum_{j=1}^{l_2} k(x_i,y_j). \end{aligned}\tag{4}$$

III. Main Results

We first consider the case with only one anomalous sequence. With the reference sequence X, we compute $\mathrm{MMD}_u^2[X,Y_k]$ for each sequence Y_k for $1 \leq k \leq n$ using (4). It is clear that if Y_k is generated by p, then $\mathrm{MMD}_u^2[X,Y_k]$ is an estimator of $\mathrm{MMD}^2[p,p]$ and hence should be close to zero. If Y_k is anomalous, $\mathrm{MMD}_u^2[X,Y_k]$ is an estimator of $\mathrm{MMD}^2[p,q]$, which is a positive constant. This understanding naturally leads to the following distribution-free test when $s = 1$. The sequence k^* is the index of the anomalous data sequence if

$$k^* = \arg\max_{1\leq k\leq n} \mathrm{MMD}_u^2[X,Y_k]. \tag{5}$$

The following theorem characterizes the condition under which the above test is consistent.

Theorem 1. *Consider the anomaly detection model with a reference sequence generated by p and with one anomalous sequence. Suppose the test* (5) *applies a bounded kernel with $0 \leq k(x,y) \leq K$ for any (x,y). Then the test* (5) *is consistent if*

$$m > \frac{24K^2(1+\eta)}{\mathit{MMD}^4[p,q]}\log n, \tag{6}$$

where η is any positive constant. Furthermore, under the above condition, the test (5) *is also exponentially consistent.*

We note that the boundedness condition $0 \leq k(x,y) \leq K$ on the kernel function is satisfied by many kernels such as the Gaussian kernel and Laplace kernel. Theorem 1 implies that it is sufficient to have $\mathcal{O}(\log n)$ samples in each data sequence in order to guarantee consistency of the test (5). This is desirable in practice because as the number of sequences gets large (and hence detection becomes more challenging), the number of sample needed for building a distribution-free and consistent test can still be much smaller, i.e., of the logarithmic order of the number of sequences.

We now consider the general case with s anomalous sequences, where $1 \leq s \leq n-1$. Here, we allow $s \geq \frac{n}{2}$ for generality of our result. We first consider the case when the value of s is known a priori, and build the following test. We compute $\mathrm{MMD}_u^2[X,Y_k]$ for each $1 \leq k \leq n$, and choose sequences with the largest s values of $\mathrm{MMD}_u^2[X,Y_k]$ to be anomalous. More specifically, the test outputs the following set that contains indices of anomalous sequences:

$$\begin{aligned}\hat{\mathcal{I}} = \{k :\ & \mathrm{MMD}_u^2[X,Y_k] \text{ is among the } s \text{ largest} \\ & \text{values of } \mathrm{MMD}_u^2[X,Y_i] \text{ for } i = 1,\ldots,n\}. \end{aligned}\tag{7}$$

The following theorem characterizes a condition under which the above test is consistent.

Theorem 2. *Consider the anomaly detection model with a reference sequence generated by p and with s anomalous sequences, where $1 \leq s \leq n-1$. Assume the value of s is known. Further assume that the test* (7) *applies a bounded kernel with $0 \leq k(x,y) \leq K$ for any (x,y). Then the distribution-free test* (7) *is consistent if*

$$m > \frac{24K^2(1+\eta)}{\mathit{MMD}^4[p,q]}\log((n-s)s), \tag{8}$$

where η is any positive constant. Furthermore, under the above condition, the test (7) *is also exponentially consistent. The computational complexity of the test* (7) *is* $\max\{\mathcal{O}(nm^2), \mathcal{O}(ns - \frac{s^2}{2})\}$.

Since s has an order less than or equal to $\mathcal{O}(n)$, Theorem 2 implies that it is sufficient to have $\mathcal{O}(\log n)$ samples in each data sequence in order to guarantee consistency of the test. Although s does not affect the order of m, it is still interesting to further understand how s affects the exact value of the threshold in (8). It can be seen that if $s \leq \frac{n}{2}$, the threshold on m to guarantee consistent detection increases as s increases, which is reasonable. It is somewhat surprising that if $s > \frac{n}{2}$, the threshold on m decreases as s increases. This is in fact also intuitive, because if $s > \frac{n}{2}$, the number of normal sequences is less than $\frac{n}{2}$, and it is hence more convenient to detect normal sequences (and consequently anomalous sequences are also identified). As s increases, the number of normal sequences decreases, and thus detection is easier.

Theorem 2 also implies that computational complexity of the test (7) is at most $\mathcal{O}(n^2)$, because the order $\mathcal{O}(\log n)$ for m is sufficient for the test to be consistent, and s is at most n.

We next consider the case in which the value of s is unknown. For this case, the test (7) is not applicable, because it depends on the value of s. In order to build a test now, we observe that for large enough m, $\mathrm{MMD}_u^2[X,Y_k]$ should be close to zero if Y_k is drawn from p, and should be far away enough from zero (in fact, close to $\mathrm{MMD}^2[p,q]$) if Y_k is drawn from the anomalous distribution q. Based on this understanding, we build the following test:

$$\widehat{\mathcal{I}} = \{k : \mathrm{MMD}_u^2[X,Y_k] > \delta_n\} \tag{9}$$

where the threshold $\delta_n \to 0$ as $n \to \infty$. This requirement on δ_n is to guarantee that the threshold is asymptotically less than $\mathrm{MMD}^2[p,q]$, which is positive but unknown. The following theorem characterizes a condition that m should satisfy in order for the test (9) to be consistent.

Theorem 3. *Consider the anomaly detection model with a reference sequence generated by p and with s anomalous sequences, where $0 \leq s \leq n-1$. Assume that the value of s is unknown a priori. Further assume that the test* (9) *adopts a threshold δ_n that converges to 0 as $n \to \infty$, and applies a bounded kernel with $0 \leq k(x,y) \leq K$ for any (x,y). Then the test* (9) *is consistent if*

$$m > 16(1+\eta)K^2 \max\left\{\frac{\log(\max\{s,1\})}{(MMD^2[p,q]-\delta_n)^2}, \frac{\log(n-s)}{\delta_n^2}\right\} \quad (10)$$

where η is any positive constant. The computational complexity of the test (9) *is $\mathcal{O}(nm^2)$.*

Remark 1. *Theorem 3 is also applicable for the case with $s=0$, i.e., the null hypothesis when there is no anomalous sequence.*

Theorem 3 implies that if $\frac{s}{n} < 1$ as $n \to \infty$, the threshold on m in (10) is dominated by the second term. Since $\delta_n \to 0$ as $n \to \infty$, m should scale strictly faster than $\mathcal{O}(\log n)$ in order to guarantee consistent detection. Compared to the case with the value of s known (for which it is sufficient for m to scale at the order $\mathcal{O}(\log n)$), the threshold on m has an order-level increase due to lack of the knowledge of s. An extreme case occurs if $\frac{s}{n} = 1$ as $n \to \infty$, in which having the order $\mathcal{O}(\log n)$ for m is sufficient. This is reasonable, because now anomalous sequences dominate so that errors caused by the asymptotically small threshold δ_n do not dominate the performance, and hence do not enlarge the requirement on the order of m.

We also note that the test (9) is not exponentially consistent. In fact, when there is no null hypothesis (i.e., $s > 1$), an exponentially consistent test can be built as follows. For each subset $\mathcal{S}$ of $\{1,\ldots,n\}$, we compute $\frac{1}{|\mathcal{S}|}\sum_{k\in\mathcal{S}} \mathrm{MMD}_u^2[X,Y_k]$, and the test finds the set of indices corresponding to the largest average value. However, for such a test to be consistent, m needs to scale linearly with n and the computational complexity is exponential with n, which is not desirable.

IV. Numerical Results

A. Demonstration of Theorems

We choose the distribution p to be Gaussian with mean zero and variance one, i.e., $\mathcal{N}(0,1)$, and choose the anomalous distribution q to be the Laplace distribution with mean one and variance one. We use the Gaussian kernel $k(x,x') = \exp(-\frac{|x-x'|^2}{2\sigma^2})$ with $\sigma = 1$ for all experiments.

For the case with $s = 1$, we run the test (5) for five cases with $n = 100, 200, 300, 400$ and 500, respectively. In Figure 2, we show how the probability of error changes with m. For illustrational convenience, we normalize m by $\log n$, i.e., the horizontal axis represents $\frac{m}{\log n}$. It is clear from the figure that when $\frac{m}{\log n}$ is above a certain threshold, the probability of error converges to zero, which is consistent with our theoretical results. Furthermore, for different values of n, all curves drop to zero almost at the same threshold. This observation confirms Theorem 1, which states that the threshold on $\frac{m}{\log n}$ depends only on the bound K of the kernel and the MMD of the two distributions. Both quantities are constant for all values of n.

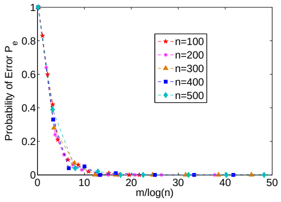


Fig. 2. Performance when $s = 1$.

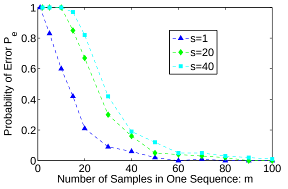


Fig. 3. Performance when $s \geq 1$.

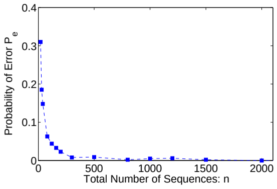


Fig. 4. Performance when s is unknown.

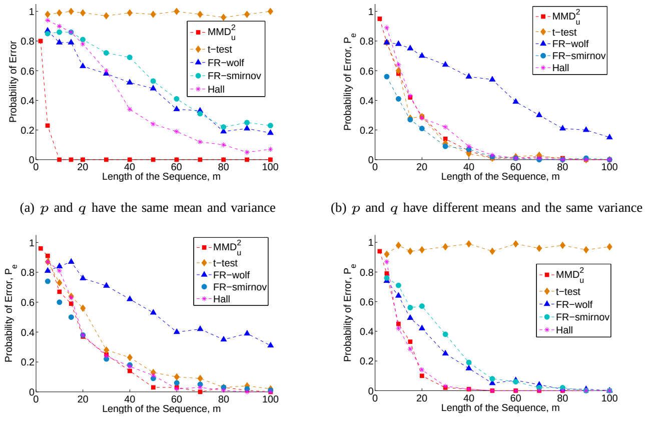
For the case with $s \geq 1$, we set $n = 100$, and run the test (7) for three cases with the numbers of anomalous sequences being $s = 1, 20$ and 40, respectively. In Figure 3, we plot the probability of error as a function of m. It can be seen that for each value of s, when m is above a certain threshold, the probability of error converges to zero, confirming that our test is consistent. Furthermore, the threshold on m at which a curve drops to zero increases with s. This observation is consistent with Theorem 2, which suggests that the threshold on m increases with s if $s < \frac{n}{2}$.

We next study the case with s anomalous sequences, but the value of s is unknown a priori. For this simulation, we set $s = 10$, but our test does not exploit such information. We choose the distribution p to be $\mathcal{N}(0, \frac{1}{2})$, and choose q to be a mixture of two Laplace distributions with equal probability: one with mean -3 and variance $\frac{1}{2}$ and the other with mean 3 and variance $\frac{1}{2}$. We apply the test (9), and set the threshold $\delta_n = \frac{1}{(\log n)^{0.7}}$, which converges to zero as $n \to \infty$. We set $m = \lceil 0.28(\log n)^{1.4}\log(n-s)\rceil$. In Figure 4, we plot the probability of error as a function of n. It can be seen that as n increases, the probability of error converges to zero. This clearly confirms Theorem 3, which implies that the chosen scaling behavior of m should guarantee consistency of the test.

B. Comparison with Other Tests

We first compare the MMD-based test with four other tests based on traditional statistical approaches: the t-test, FR-Wolf test, FR-Smirnov test, and Hall test. We focus on the scenario with $s = 1$, and set $n = 100$, and compare the five tests for the following four cases:

- p and q have same mean and same variance: p is a Laplacian distribution with mean 1 and variance 5, and q is a mixture of two Laplacian distributions with equal probability: one with mean -1 and variance 1, and the other with mean 3 and variance 1.
- p and q have different means and same variance: p is a Gaussian distribution with mean 0 and variance 1, and q is a Laplacian distribution with mean 1 and variance 1.
- p and q have different mean and different variance: p is Gaussian distribution with mean 0 and variance 1, and q



(a) p and q have the same mean and variance

(b) p and q have different means and the same variance

(c) p and q have different means and variances

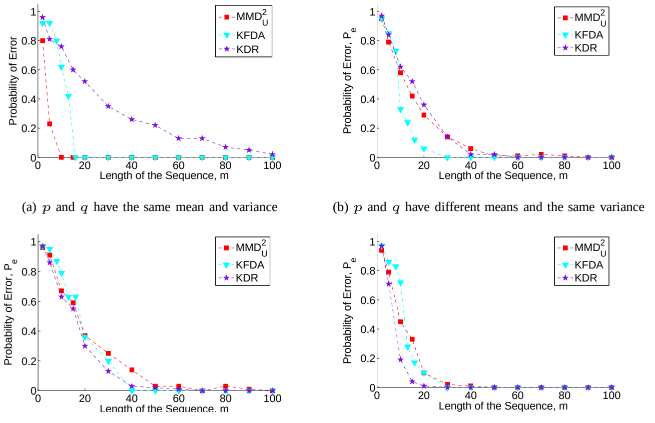
(d) p and q have the same mean and different variances

Fig. 5. Comparison of MMD-based test with other four tests

is a Laplacian distribution with mean 1 and variance 3.

- p and q have same mean and different variance: p is a Gaussian distribution with mean 0 and variance 1, and q is a mixture of two Laplacian distributions with equal probability: one with mean -2 variance 3 and the other with mean 2 and variance 3.

In Figure 5, we plot the probability of error as a function of m for the five tests for the above four cases. It can be seen that for all cases, the MMD-based test is either the best or one of the best tests among the five tests. In particular, the MMD-based test performs much better than other tests for the case when p and q have the same mean and variance, which suggests that the MMD-based test is especially useful for capturing differences in higher order moments between two distributions compared to other tests. We also note that although the Hall test sometimes yields comparable performance with the MMD-based test, its complexity is much larger than that of the MMD-based test.



(a) p and q have the same mean and variance

(b) p and q have different means and the same variance

(c) p and q have different means and variances

(d) p and q have the same mean and different variances

Fig. 6. Comparison of MMD-based test with two other kernel-based tests

We next compare our MMD-based test with two other kernel-based tests, KFDA and KDR (KDR uses divergence as the metric). We focus on the scenario with $s = 1$, and set $n = 100$. For all tests, we use the Gaussian kernel with $\sigma = 1$. We compare the three kernel-based tests for the same four cases as in the previous comparison. In Figure 6, we plot the probability of error as a function of m for the three tests. It can be seen that the MMD-based test performs much better than the other two tests when p and q have the same mean and variance, and performs as well as the other two tests in the remaining three cases.

V. Conclusion

In this paper, we have investigated a nonparametric anomaly detection problem, and have built MMD-based distribution-free tests to detect anomalous sequences. We have characterized the scaling behavior of the number m of samples as the total number n of sequences goes to infinity in order to guarantee consistency of the developed tests. We have demonstrated the performance of our tests by comparing them to other appealing tests. Our study of this problem demonstrates a useful application of the mean embedding of distributions and MMD, and we believe that such an approach can be applied to solving various other nonparametric problems.

Acknowledgment

The work of S. Zou and Y. Liang was supported by a NSF CAREER Award under Grant CCF-10-26565 and by the NSF under Grant CNS-11-16932. The work of H. V. Poor was supported in part by the NSF under Grants DMS-1118605 and ECCS-1343210.

References

[1] S. Zou, Y. Liang, H. V. Poor, and X. Shi. Nonparametric detection of anomalous data via kernel mean embedding, Preprint. http://szou02.mysite.syr.edu/Journal/SequenceDetection_IT2014_arxiv.pdf, 2014.

[2] L. Lai, H. V. Poor, Y. Xin, and G. Georgiadis. Quickest search over multiple sequences. *IEEE Trans. Inform. Theory*, 57(8):5375–5386, August 2011.

[3] A. Patcha and J.-M. Park. An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Comput. Networks*, 51(12):3448–3470, August 2007.

[4] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):1–58, July 2009.

[5] Y. Li, S. Nitinawarat, and V.V. Veeravalli. Universal outlier hypothesis testing. Submitted to *IEEE Trans. Inform. Theory*, May 2013.

[6] J. H. Friedman and L. C. Rafsky. Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics*, 7(4):pp. 697–717, 1979.

[7] P. Hall and N. Tajvidi. Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89(2):pp. 359–374, 2002.

[8] T. Kanamori, T. Suzuki, and M. Sugiyama. Divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. *IEEE Trans. Inform. Theory*, 58(2):708–720, Feb 2012.

[9] Z. Harchaoui, F. Bach, and E. Moulines. Testing for homogeneity with kernel Fisher discriminant analysis. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2008.

[10] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.

[11] B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Scholköpf. Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.*, 11:1517–1561, 2010.

[12] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2008.

[13] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012.

[14] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT press, 2002.