Fifty-second Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
October 1 - 3, 2014

Unsupervised Nonparametric Anomaly Detection: A Kernel Method

Shaofeng Zou[1] Yingbin Liang[1] H. Vincent Poor[2] Xinghua Shi[3]

***Abstract*— An anomaly detection problem is investigated, in which s out of n sequences are anomalous and need to be detected. Each sequence consists of m independent and identically distributed (i.i.d.) samples drawn either from a nominal distribution p or from an anomalous distribution q that is distinct from p. Neither p nor q is known a priori. Two scenarios respectively with s known and unknown are studied. Distribution-free tests are constructed based on the metric of the maximum mean discrepancy (MMD). It is shown that if the value of s is known, as n goes to infinity, the number m of samples in each sequence should be of order $\mathcal{O}(\log n)$ or larger to guarantee that the constructed test is exponentially consistent. On the other hand, if the value of s is unknown, the number m of samples in each sequence should be of the order strictly greater than $\mathcal{O}(\log n)$ to guarantee the constructed test is consistent. The computational complexity of all tests are shown to be polynomial. Numerical results are provided to confirm the theoretic characterization of the performance. Further numerical results on both synthetic data sets and real data sets demonstrate that the MMD-based tests outperform or perform as well as other approaches.**

I. Introduction

An anomaly detection problem (also referred to as outlier hypothesis testing problem) has recently attracted considerable research interest. In this problem, s out of n sequences are anomalous and need to be detected. Each sequence consists of m independent and identically distributed (i.i.d.) samples drawn either from a nominal distribution p or from an anomalous distribution q that is distinct from p. We note that each data point in this problem contains multiple samples from a distribution. This is different from the anomaly or outlier detection problem generally considered in machine learning [1], [2], in which each data point is only one realization of a certain distribution.

Such a problem has significant potential for practical application. For example, for a group of people with one certain genetic disease, it is likely that expression levels of a few genes responsible for the disease follow distributions different from those of genes not related to the disease. It is thus important to identify those genes that are responsible for the disease out of a large number of genes based on their expression levels. Another potential application is in cognitive wireless networks, in which signals follow different distributions depending on whether the channels are busy or not. The major task is to identify the vacant channels such that users can transmit over those vacant channels to improve the spectral efficiency. Other applications include detecting virus infected computers from other virus free computers and detecting slightly modified images from other untouched images.

The parametric case of the problem has been well studied previously, e.g., in [3], in which the distributions p and q are assumed to be known a priori, and a maximum likelihood ratio test can be applied. Recently, the nonparametric model has been studied in [4] and [5], in which p and q are assumed to be unknown, and both the distributions are assumed to be discrete. In particular, [4] proposed a nonparametric divergence-based generalized likelihood test, and characterized error exponents in the asymptotic regime as the sample size goes to infinity. In [5], tests based on l_1 distance are constructed and analyzed for the non-asymptotic regime (e.g., with finite sample size). Both studies exploit the fact that the distributions are discrete, and hence empirical distributions based on data samples are used for constructing tests.

In [6], the nonparametric model with p and q being arbitrary was studied. In particular, p and q can be continuous. It is assumed that a reference sequence containing samples generated from the distribution p is available. A distance metric referred to as the *maximum mean discrepancy (MMD)* [7], [8] is adopted for constructing the nonparametric tests and conditions for the tests to be consistent are characterized. The MMD-based approach uses mean embedding of distributions into a reproducing kernel Hilbert space (RKHS) [9], [10], i.e., mapping probability distributions into an RKHS. It can be shown [10]–[13] that such a mapping is injective for characteristic kernels (such as Gaussian and Laplace kernels). Thus the distance between two mean embeddings of two distributions in the RKHS provides a natural distance measure between two distributions. There are a few advantages of the MMD-based metric: (1) it is computationally efficient to estimate the MMD from samples, particularly for vector distributions; and (2) MMD-based approaches do not need to estimate probability density functions as intermediate steps, and hence can avoid error propagation. Our anomaly detection tests are constructed utilizing the MMD metric.

In this paper, we further study the nonparametric model, in which the distributions p and q are unknown, arbitrary, and can be continuous. In contrast to [6], we assume that no reference sequence generated from the distribution p is available. As in [6], we also adopt the MMD as the

*The work of S. Zou and Y. Liang was supported by an 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.

[1]S. Zou and Y. Liang are with the Department of Electrical Engineering and Computer Science, Syracuse University, USA {szou02,yliang06}@syr.edu

[2]H. V. Poor is with the Department of Electrical Engineering, Princeton University, USA poor@princeton.edu

[3] X. Shi is with Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, xshi3@uncc.edu

distance measure between distributions to construct our tests. Since a reference sequence is not available, our construction of tests cannot be based on the distance (i.e., the MMD) between each sequence and the reference to detect anomalous sequences. Instead, we exploit the fact that the MMD between a nominal sequence and the remaining sequences is different from the MMD between an anomalous sequence for constructing the tests. In this paper, we focus on the same asymptotic regime as in [6], i.e., the total number n of data sequences goes to infinity, and possibly the number s of anomalous sequences can also become large. It then requires that the number m of samples in each sequence increase to guarantee the asymptotically small detection error probability. We are interested in characterizing the sufficient conditions for m to scale with (n, s) in order to guarantee that the tests are consistent.

In summary, our main contributions in this paper are as follows. (1) We construct computationally efficient distribution-free MMD-based tests for two scenarios: the number s of anomalous sequences is known and unknown, respectively. (2) We characterize how the number m of samples in each sequence should scale with (n, s) to guarantee consistency of the tests. We show that m can be much smaller than n (i.e., on the order of $\mathcal{O}(\log n)$ if s is known, and on the order greater than $\mathcal{O}(\log n)$ if s is unknown). Therefore, lack of the knowledge of s results in an order level increase in the sample size m to guarantee consistency of the tests. (3) We provide numerical results on synthetic data to demonstrate our theoretical assertions. We further demonstrate that our MMD-based approach slightly outperforms nonparametric generalized likelihood tests in [4] on discrete distributions. Finally, we develop/implement traditional statistical approaches together with our tests on a real data set to demonstrate the consistency of our tests and its competitive performance with other approaches. We note that in this paper, we omit the proofs. The details can be found in [14].

II. Problem Statement and Approach

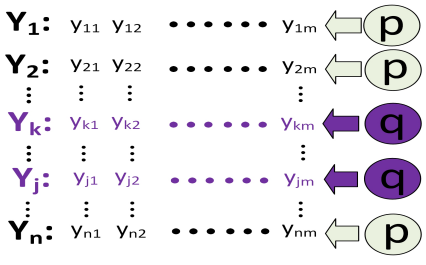

Fig. 1. An anomaly detection model with data sequences from distribution p and anomalous distribution q.

The problem we study in this paper is depicted in Fig. 1. There are n data sequences in total, out of which s sequences are anomalous. The k-th sequence is denoted by $Y_k := (y_{k1}, \ldots, y_{km})$ for $1 \leq k \leq n$, in which $y_{k1}, \ldots, y_{km}$ are m i.i.d. samples drawn from either a nominal distribution p or an anomalous distribution q. We assume that p and q are unknown a priori. We also assume that p and q can be arbitrary, and are distinct, i.e., $p \neq q$. In contrast to [6], we focus on the unsupervised case, in which no reference sequence from p is available. Our goal here is to build distribution-free tests to detect s data sequences generated by the anomalous distribution q.

We study two cases with the value of s being known and unknown a priori, respectively. Our focus is on 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 $0 \leq \alpha < \frac{1}{2}$. By symmetry, the case that $\frac{1}{2} < \alpha \leq 1$ is equivalent with the roles of p and q being exchanged. The test for the unknown s and the corresponding analysis are also applicable to the case in which $s = 0$, i.e., the null hypothesis in which there is no anomalous sequence. We will comment on this 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 $n \to \infty$.

The problem is a multi-hypothesis testing problem. We adopt the following risk function as the performance measure of the tests:

$$P_e^{(n)} = \max_{|\mathcal{I}|=s} P\{\hat{\mathcal{I}}^n \neq \mathcal{I} | \mathcal{I}\}, \tag{1}$$

where $\mathcal{I}$ denotes the index set of all anomalous data sequences, and $\mathcal{I}n$ denotes a sequence of index sets of anomalous sequences claimed by a corresponding sequence of tests. Thus, our definition of a test being consistent is given as follows.

***Definition** 1:* A sequence of tests is said to be consistent if

$$\lim_{n\to\infty} P_e^{(n)} = 0. \tag{2}$$

We note that although the limit in the above definition is taken with respect to n, since consistency of tests requires that m increase with n, the limit can also be equivalently viewed to be taken as m goes to infinity. Furthermore, for a consistent test, we are interested in whether the risk decays exponentially fast with respect to the number m of samples.

***Definition** 2:* A sequence of tests is said to be exponentially consistent if

$$\liminf_{m\to\infty} -\frac{1}{m} \log P_e^{(n)} > 0. \tag{3}$$

In this paper, we adopt the following technique of mean embedding of distributions into an RKHS [9], [10] and the metric of MMD for constructing the tests.

As developed in [8], the MMD as the distance between the mean embeddings μ_p and μ_q of the distributions p and q is given by

$$\begin{aligned}\mathrm{MMD}^2[p, q] &:= \|\mu_p - \mu_q\|_{\mathcal{H}} \\ &= \mathbb{E}_{x,x'}[k(x, x')] - 2\mathbb{E}_{x,y}[k(x, y)] + \mathbb{E}_{y,y'}[k(y, y')],\end{aligned}$$

where x and x' are independent but have the same distribution p, and y and y' are independent but have the same distribution q. Naturally, an unbiased estimator of

$\text{MMD}^2[p,q]$ based on l_1 samples of X and l_2 samples of Y is given as follows:

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

III. Main Results

In this section, we start with the case in which the value of s is known, and then study the case with s unknown.

A. Known s

We first use a simple case with $s = 1$ to introduce the idea of our tests, and then study the more general case for arbitrary s, in which $\frac{s}{n} \to \alpha$ as $n \to \infty$, where $0 \leq \alpha \leq \frac{1}{2}$.

Suppose that $s = 1$. For each sequence Y_k, we use $\overline{Y}_k$ to denote the $(n-1)m$ dimensional sequence that stacks all other sequences together, as given by

$$\overline{Y}_k = \{Y_1, \ldots, Y_{k-1}, Y_{k+1}, \ldots, Y_n\}.$$

We then compute $\text{MMD}_u^2[Y_k, \overline{Y}_k]$ for $1 \leq k \leq n$. It is clear that if Y_k is an anomalous sequence, then $\overline{Y}_k$ is fully composed of sequences from p. Hence, $\text{MMD}_u^2[Y_k, \overline{Y}_k]$ is a good estimator of $\text{MMD}^2[p,q]$, which is a positive constant. On the other hand, if Y_k is a sequence from p, $\overline{Y}_k$ is composed of $n-2$ sequences generated by p and only one sequence generated by q. As n increases, the impact of the anomalous sequence on $\overline{Y}_k$ is negligible, and $\text{MMD}_u^2[Y_k, \overline{Y}_k]$ should be close to zero. Based on the above understanding, we construct the following test when $s = 1$. The sequence k^* is the index of the anomalous data sequence if

$$k^* = \arg\max_{1\leq k\leq n} \text{MMD}_u^2[Y_k, \overline{Y}_k]. \quad (5)$$

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

Theorem 1: Consider the anomaly detection model 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

$$m \geq \frac{16K^2(1+\eta)}{\text{MMD}^4[p,q]}\log n, \quad (6)$$

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

We now consider the more general case in which $\frac{s}{n} \to \alpha$ as $n \to \infty$, where $0 \leq \alpha < \frac{1}{2}$. Our test is a natural generalization of the test (5) except now the test chooses the sequences with the largest s values of $\text{MMD}_u^2[Y_k, \overline{Y}_k]$, which is given by

$$\hat{\mathcal{I}} = \{k : \text{MMD}_u^2[Y_k, \overline{Y}_k] \text{ is among the } s \text{ largest values of } \text{MMD}_u^2[Y_i, \overline{Y}_i] \text{ for } i = 1, \ldots, n\}. \quad (7)$$

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

Theorem 2: Consider the anomaly detection model with s anomalous sequences, where $\frac{s}{n} \to \alpha$ as $n \to \infty$ and $0 \leq \alpha < \frac{1}{2}$. 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 test (7) is consistent if

$$m \geq \frac{16K^2(1+\eta)}{(1-2\alpha)^2\text{MMD}^4[p,q]}\log(s(n-s)), \quad (8)$$

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 $\mathcal{O}(n^3m^2)$.

Since $s \leq \mathcal{O}(n)$, we have $\log s(n-s) \sim \mathcal{O}(\log n)$. Hence, Theorem 1 and Theorem 2 imply that, our MMD based tests (5) and (7) require only $\mathcal{O}(\log n)$ samples in each data sequence in order to guarantee consistency of the tests.

Remark 1: For the case with $\frac{s}{n} \to 0$, as $n \to \infty$, we can build a test that has reduced computational complexity. For each Y_k, instead of using $n-1$ sequences to build $\overline{Y}_k$ as in the test (7), we take any l sequences out of the remaining $n-1$ sequences to build a sequence $\widetilde{Y}_k$, such that $\frac{l}{n} \to 0$ and $\frac{s}{l} \to 0$ as $n \to \infty$. Such an l exists for any s and n satisfying $\frac{s}{n} \to 0$ (e.g., $l = \sqrt{sn}$). It can be shown that using $\widetilde{Y}_k$ to replace $\overline{Y}_k$ in the test (7) still leads to consistent detection under the same condition given in Theorem 2. The computational complexity of such a test becomes $\mathcal{O}(nl^2m^2)$, which is substantially smaller than $\mathcal{O}(n^3m^2)$ of the test (7), considering that l is less than n in the order sense.

B. Unknown s

In this subsection, we consider the case in which the value of s is unknown a priori. For this case, the previous test (7) is not applicable due to the lack of knowledge of s. We observe that for large value of m, $\text{MMD}_u^2[Y_k, \overline{Y}_k]$ should be close to 0 if Y_k is drawn from p, and should be close to $\text{MMD}^2[p,q]$ if Y_k is drawn from q. Based on this understanding, we build the following test:

$$\widehat{\mathcal{I}} = \{k : \text{MMD}_u^2[Y_k, \overline{Y}_k] > \delta_n\} \quad (9)$$

where $\delta_n \to 0$ and $\frac{s^2}{n^2\delta_n} \to 0$ as $n \to \infty$. We note that the above requirements on δ_n implies that the test (9) is applicable only when $\frac{s}{n} \to 0$ as $n \to \infty$. This includes two cases: (1) s is fixed; and (2) $s \to \infty$ and $\frac{s}{n} \to 0$ as $n \to \infty$. Furthermore, the scaling behavior of s as n increases needs to be known in order to choose δ_n for the test. This is reasonable to assume because mostly in practice the scale of anomalous data points can be estimated based on domain knowledge.

The following theorem characterizes conditions under which the test (9) is consistent.

Theorem 3: Consider the anomaly detection model with s anomalous sequences, where $\lim_{n\to\infty}\frac{s}{n} = 0$. Assume that the value of s is unknown a priori. Further assume that the test (9) adopts a threshold δ_n such that $\delta_n \to 0$ and $\frac{s^2}{n^2\delta_n} \to 0$,

as $n \to \infty$, and the test applies a bounded kernel with $0 \leq k(x,y) \leq K$ for any (x,y). Then the test (9) is consistent if

$$m \geq 16(1+\eta)K^2 \max\left\{ \frac{\log(\max\{1,s\})}{(\mathrm{MMD}^2[p,q]-\delta_n)^2}, \frac{\log(n-s)}{(\delta_n - \mathbb{E}\left[\mathrm{MMD}_u^2[Y,\overline{Y}]\right])^2} \right\}, \quad (10)$$

where η is any positive constant, and $E[\mathrm{MMD}_u^2[Y,\overline{Y}]]$ is a constant, where Y is a sequence generated by p and $\overline{Y}$ is a stack of $(n-1)$ sequences with s sequences generated by q and the remaining sequences generated by p. The computational complexity of the test (9) is $\mathcal{O}(n^3m^2)$.

We note that Theorem 3 is also applicable to the case with $s = 0$, i.e., the null hypothesis when there is no anomalous sequence. We further note that the test (9) is not exponentially consistent. If there is no null hypothesis (i.e., $s > 0$ and unknown), an exponentially consistent test can be built as follows. For each subset $\mathcal{S}$ of $\{1,\ldots,n\}$ we compute the average $\frac{1}{|\mathcal{S}|}\sum_{k\in\mathcal{S}} \mathrm{MMD}_u^2[Y_k,\overline{Y}_k]$, and the test finds the set of indices corresponding to the largest average value. However, now m needs to scale linearly with n for the test to be consistent, and the computational complexity is exponential with n, which is not desirable.

Theorem 3 implies that the threshold on m to guarantee consistent detection has an order strictly larger than $\mathcal{O}(\log n)$, because $\frac{s}{n} \to 0$ and $\delta_n \to 0$ as $n \to \infty$. This is the price paid due to not knowing s.

IV. Numerical Results

In this section, we provide numerical results to demonstrate our theoretical assertions, and compare our MMD-based tests with a number of tests based on other approaches on both synthetic and real data sets.

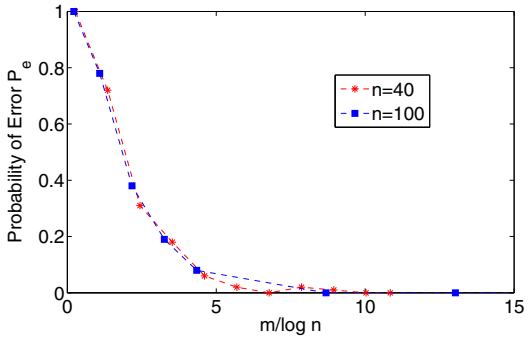

Fig. 2. Performance of the test with $s = 1$.

A. Demonstration of Theorems

We choose the distribution p to be Gaussian with mean zero and variance one, 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 $\sigma = 1$. We choose $s = 1$. We run the test for cases with the numbers of sequences being $n = 40$ and 100, respectively. In Fig. 2, we plot the probability of error as a function of $\frac{m}{\log n}$. It can be seen that, as m increases, the probability of error converges to zero. In particular, both curves drop to zero at almost the same threshold, which agrees with Theorem 1.

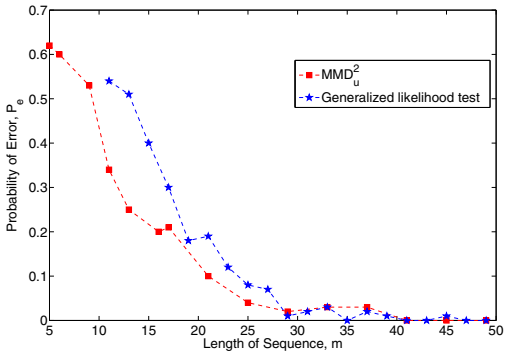

Fig. 3. Comparison of the MMD-based test with divergence-based generalized likelihood test.

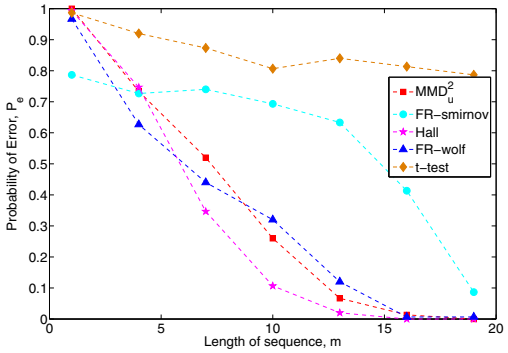

Fig. 4. Comparison of the MMD-based test with four other tests on a real data set.

B. Comparison with Other Tests

In this subsection, we compare our MMD-based tests with tests based on other nonparametric approaches. We first compare our test with the divergence-based generalized likelihood approach developed in [4]. Since the test in [4] is applicable only when the distributions p and q are discrete and have finite alphabets, we set the distributions p and q to be binary with p having probability 0.3 to take "0" (and probability 0.7 to take "1"), and q having probability 0.7 to take "0" (and probability 0.3 to take "1"). We let $s = 1$ and assume that s is known. We let $n = 50$.

In Fig. 3, we plot the probability of error as a function of m. It can be seen that the MMD-based test has a slightly better performance than the divergence-based generalized likelihood test in both cases. We note that it has been shown in [4] that the divergence-based test has optimal convergence rate in the limiting case when n is infinite, which suggests that such a test should also perform well for the case with finite n. Thus, the comparison demonstrates that the MMD-based test can provide comparable or even better performance than the well-performing divergence-based test.

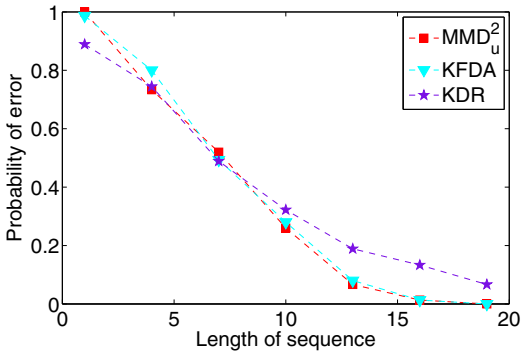

Fig. 5. Comparison of the MMD-based test with two other kernel-based tests on a real data set.

C. Application to Real Data Set

In this subsection, we study how the MMD-based test performs on a real data set. We choose the collection of daily maximum temperature of Syracuse (New York, USA) in July from 1993 to 2012 as the nominal data sequences, and the collection of daily maximum temperature of Makapulapai (Hawaii, USA) in May from 1993 to 2012 as anomalous sequences. Here, each data sequence contains daily maximum temperatures of a certain day across twenty years from 1993 to 2012. In our experiment, the data set contains 32 sequences in total, including one temperature sequence of Hawaii and 31 sequences of Syracuse. The probability of error is averaged over all cases with each using one sequence of Hawaii as the anomalous sequence. Although it seems easy to detect the sequence of Hawaii out of the sequences of Syracuse, the temperatures we compare for the two places are in May for Hawaii and July for Syracuse, during which the two places have approximately the same mean in temperature. In this way, it may not be easy to detect the anomalous sequence (in fact, some tests do not perform well as shown in Fig. 4).

We apply the MMD-based test and compare its performance with the t-test, FR-Wolf test [15], FR-Smirnov test [15], and Hall test [16]. For the MMD-based test, we use the Gaussian kernel with $\sigma = 1$. In Fig. 4, we plot the probability of error as a function of m for all tests. It can be seen that the MMD-based test, Hall test, and FR-wolf test have the best performance, and all of the three tests are consistent with the probability of error converging to zero as m increases. Furthermore, comparing to Hall and FR-wolf tests, the MMD-based test has the lowest computational complexity.

We also compare the performance of the MMD-based test with the kernel-based tests KFDA [17] and KDR [18]. We use a Gaussian kernel with $\sigma = 1$. In Fig. 5, we plot the probability of error as a function of m. It can be seen that all tests are consistent with the probability of error converging to zero as m increases, and the MMD-based test has the best performance among the three tests.

V. Conclusion

In this paper, we have studied a nonparametric anomaly detection problem, in which s anomalous sequences need to be detected out of n sequences. We have built MMD-based distribution-free tests and characterized how the sample size m (in each sequence) should scale with n and s to guarantee consistency (or exponentially consistency) of tests for both the case with known s and the case with unknown s. If s is known, we have shown that m should scale with an order of $\mathcal{O}(\log n)$. If s is unknown, we have shown that m should scale with the order greater than $\mathcal{O}(\log n)$. Furthermore, we have demonstrated our theoretical results by numerical results on synthetic data and have demonstrated the performance of our tests by comparing them to other appealing tests on a real data set. Our study of this problem demonstrates that the MMD metric can be used as a powerful tool for distinguishing between distributions based on data samples, and can thus be used for solving other nonparametric problems in the future.

References

[1] 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.

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

[3] 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.

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

[5] J. Acharya, A. Jafarpour, A. Orlitsky, and A.T. Suresh. Sublinear algorithms for outlier detection and generalized closeness testing. In *Proc. IEEE Int. Symp. Information Theory (ISIT)*, June 2014.

[6] S. Zou, Y. Liang, H. V. Poor, and X. Shi. Kernel-based nonparametric anomaly detection. In *Proc. IEEE Int. Workshop on Signal Processing Advances for Wireless Communications (SPAWC)*, June 2014.

[7] A. Gretton, K. Borgwardt, M. Rasch, B. Schökopf, and A. Smola. A kernel method for the two-sample-problem. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2007.

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

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

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

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

[12] B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Injective Hilbert space embeddings of probability measures. In *Proc. Annual Conference on Learning Theory (COLT)*, 2008.

[13] K. Fukumizu, B. Sriperumbudur, A. Gretton, and B. Schölkopf. Characteristic kernels on groups and semigroups. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2009.

[14] S. Zou, Y. Liang, H. V. Poor, and X. Shi. Nonparametric detection of anomalous data via kernel mean embedding. *http://arxiv.org/abs/1405.2294*, 2014.

[15] J. H. Friedman and L. C. Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *Ann. Statist.*, 7(4):pp. 697–717, 1979.

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

[17] 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.

[18] 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.