UNIVERSAL OUTLIER HYPOTHESIS TESTING: APPLICATION TO ANOMALY DETECTION

Yun Li, Sirin Nitinawarat, Yu Su and Venugopal V. Veeravalli*

Department of Electrical and Computer Engineering
and
Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
*Qualcomm Technologies, Inc.

ABSTRACT

In outlier hypothesis testing, multiple observation sequences are collected, a small subset of which are outliers. Observations in an outlier sequence are generated by a mechanism different from that generating the observations in the majority of sequences. The goal is to best discern all the outlier sequences without any knowledge of the underlying generating mechanisms. A generalized likelihood test is considered in the fixed sample size setting. In the sequential setting, a test based on the Multihypothesis Sequential Probability Ratio Test and the repeated significance test is considered. The sequential test outperforms the generalized likelihood test when the lengths of the observation sequences exceed certain values. Applied to a real data set for spam detection, the performance of the proposed tests is shown to be superior to those based on the maximum mean discrepancy for large sample size.

***Index Terms*—** anomaly detection, universal outlier hypothesis testing, generalized likelihood test, multihypothesis sequential probability ratio test, maximum mean discrepancy

1. INTRODUCTION

Consider the following inference problem of outlier hypothesis testing. Among M independent and memoryless observation sequences, it is assumed that there is a small subset (possibly empty) of outlier sequences. The observations in an outlier sequence are distributed according to an "outlier" distribution, distinct from the common "typical" distribution that governs the observations in the majority of sequences. The goal is to design a test to best discern all the outliers. We are interested in a universal setting of the problem, where the test has to be designed without any knowledge of the outlier and typical distributions. Outlier hypothesis testing arises in fraud and anomaly detection in large data sets studied here, environmental monitoring in sensor networks, spectrum sensing and high frequency trading.

It is to be noted that outlier hypothesis testing is distinct from statistical *outlier detection* [1, 2], where the goal is to efficiently winnow out a few outlier observations from a single sequence of observations. In statistical outlier detection, the outlier observations constitute a much smaller fraction of the entire observations than in outlier hypothesis testing, and they can be arbitrarily spread out among all observations.

The fixed sample size setting of universal outlier hypothesis testing was studied in [3]. The main finding therein was that the *generalized likelihood* (GL) test is far more efficient for universal outlier hypothesis testing than for the other inference problems studied in a universal setting, such as homogeneity testing and classification [4–6]. In particular, for outlier hypothesis testing, the GL test was shown to achieve *universally exponential consistency* under every non-null hypothesis, and consistency under the null hypothesis. In addition, when there is at most one outlier, as M goes to infinity, the achievable error exponent of the GL test converges to the absolutely optimal one achievable when both the outlier and typical distributions are known.

In the sequential setting, the goal is to identify all the outlier sequences using the fewest number of observations on average. A universal sequential test based on the principles underlying the Multihypothesis Sequential Probability Ratio Test (MSPRT) [7] and the GL test [8] was proposed in [9]. The proposed test also adopts a time-dependent threshold, which is inspired by the repeated significance test [10, 11]. The proposed test was shown to achieve *universally exponential consistency* under every non-null hypothesis, and yield *consistency* un-

This work was supported by the Air Force Office of Scientific Research (AFOSR) under the Grant FA9550-10-1-0458 through the University of Illinois at Urbana-Champaign, and by the National Science Foundation under Grant NSF CCF 11-11342.

der the null hypothesis. In addition, when there is at most one outlier, as M goes to infinity, the achievable error exponent of the proposed test approaches the optimal error exponent achievable when both the typical and outlier distributions are known.

In a recent work by Zou, et al. [12], the authors proposed a universal test for the fixed sample size setting, which is based on mean embedding of distributions into a reproducing kernel Hilbert space (RKHS) [13]. The test is constructed using estimates of the *maximum mean discrepancy* (MMD) between the distributions underlying each pair of observation sequences. This test was shown to be universally consistent, and sometimes universally exponentially consistent for various models. However, it is not known whether the MMD-based test can be generalized to the sequential setting.

In this paper, we evaluate the performance of the various proposed tests on a spam detection data set. Multiple sequences of emails are collected. One of the sequences contains only spams, while the rest non-spams. The goal is to identify the outlier sequence that consists of only spams. It is shown that for large enough sample size, the sequential test outperforms the GL test, which again yields better performance than the MMD-based test for this data set.

2. PRELIMINARIES

Throughout the paper, random variables are denoted by capital letters, and their realizations are denoted by the corresponding lower-case letters. All random variables are assumed to take values in *finite* alphabets.

For a finite set $\mathcal{Y}$, let $\mathcal{Y}^m$ denote the m Cartesian product of $\mathcal{Y}$, and $\mathcal{P}(\mathcal{Y})$ denote the set of all probability mass functions (pmfs) on $\mathcal{Y}$. The empirical distribution of a sequence $\boldsymbol{y} = y^m = (y_1, \ldots, y_m) \in \mathcal{Y}^m$, denoted by $\gamma = \gamma_{\boldsymbol{y}} \in \mathcal{P}(\mathcal{Y})$, is defined at each $y \in \mathcal{Y}$ as

$$\gamma(y) \triangleq \frac{1}{m}\Big|\{k = 1, \ldots, m : y_k = y\}\Big|.$$

In the rest of the paper, we restrict our attention to models with *at most one* outlier, where the outlier distribution does not depend on the identity of the outlier. Results on models with multiple and possibly distinctly distributed outliers can be found in [3].

Consider $M \geq 3$ independent sequences, each of which consists of i.i.d. observations. Denote the k-th observation of the i-th sequence by $Y_k^{(i)} \in \mathcal{Y}$. It is assumed that there is either one or no outlier among the M sequences. In particular, if the i-th sequence is the outlier, the observations in that sequence are uniquely distributed according to an "outlier" distribution $\mu \in \mathcal{P}(\mathcal{Y})$, while all the other sequences are distributed according to a "typical" distribution $\pi \in \mathcal{P}(\mathcal{Y})$. *Nothing is known about μ and π except that $\mu \neq \pi$, and that each of them has a full support.*

Conditioned on the i-th sequence being the outlier, $i = 1, \ldots, M$, the joint distribution of the first n observations is

$$\begin{aligned} p_i\left(y^{Mn}\right) &= \prod_{k=1}^{n}\left\{\mu\left(y_k^{(i)}\right)\prod_{j\neq i}\pi\left(y_k^{(j)}\right)\right\} \\ &\triangleq L_i(y^{Mn}, \mu, \pi). \end{aligned} \tag{1}$$

Under the *null* hypothesis with no outlier, the joint distribution of the first n observations is given by

$$p_0\big(y^{Mn}\big) = \prod_{k=1}^{n}\prod_{i=1}^{M}\pi\big(y_k^{(i)}\big).$$

3. FIXED SAMPLE SIZE SETTING

We first consider the setting where the number of observations in each sequence is fixed at the outset. A fixed sample size test for the outlier is done based on a *universal* rule $\delta : \mathcal{Y}^{Mn} \to \{0, 1, \ldots, M\}$, where "0" corresponds to a decision in favor of the null hypothesis. Specifically, the test δ is not allowed to be a function of the unknown distributions (μ, π).

The accuracy of a test is gauged using the maximal probability of error, defined as

$$P_{\max} \triangleq \max_{i=0,1,\ldots,M}\mathbb{P}_i\Big\{\delta\big(Y^{Mn}\big) \neq i\Big\}.$$

We say a test is *universally consistent* if the maximal probability of error vanishes for any (μ, π), $\mu \neq \pi$, as $n \to \infty$. Further, it is termed *universally exponentially consistent* if the exponent for the maximal probability of error, defined as

$$\alpha \triangleq \lim_{n\to\infty} -\frac{1}{n}\log P_{\max},$$

is strictly positive for any (μ, π), $\mu \neq \pi$.

3.1. Universal Test

For each $i = 1, \ldots, M$, denote the empirical distribution of $\boldsymbol{y}^{(i)}$ by γ_i. In the universal setting with (μ, π) being unknown, we compute the generalized likelihood (GL) of y^{Mn} by replacing the μ and π in (1) with their maximum likelihood (ML) estimates $\hat{\mu}_i \triangleq \gamma_i$, and $\hat{\pi}_i \triangleq \frac{\sum_{k\neq i}\gamma_k}{M-1}$, $i = 1, \ldots, M$, as

$$\hat{p}_i^{\text{univ}}\left(y^{Mn}\right) = L_i\left(y^{Mn}, \hat{\mu}_i, \hat{\pi}_i\right).$$

The test can be described as

$$\delta(y^{Mn}) = \begin{cases} \arg\max\limits_{i=1,\ldots,M} \hat{p}_i^{\text{univ}}(y^{Mn}), & \text{if } \max\limits_{j\neq k}\frac{1}{n}\big(\log\hat{p}_j^{\text{univ}}(y^{Mn}) \\ & \quad -\log\hat{p}_k^{\text{univ}}(y^{Mn})\big) > \lambda_n, \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where $\lambda_n = 2(M-1)|\mathcal{Y}|\frac{\log n}{n}$, and the ties in the first case of (2) are broken arbitrarily.

3.2. Results

The performance of the GL test in (2) is characterized in the following theorem, and its proof can be found in [3].

Theorem 1. *The test in (2) is universally consistent, i.e., for any $\mu, \pi, \mu \neq \pi$, it holds that*

$$P_{\max} \ \to 0$$

as $n \to \infty$. The test also yields a positive exponent for the conditional error probability under every non-null hypothesis (cf. Theorem 2 and 5 in [3]).

In addition, under every non-null hypothesis, as $M \to \infty$, the achievable error exponent of the test in (2) approaches the absolutely optimal one achievable when both μ, π are known (cf. Theorem 3 and 5 in [3]).

4. SEQUENTIAL SETTING

A sequential test for the outlier consists of a stopping rule and a final decision rule. The stopping rule defines a random stopping time, denoted by N, which is the number of observations taken until a final decision is made. At the stopping time $N = n$, a decision is made based on a decision rule $\delta : \mathcal{Y}^{Mn} \to \{0, 1, \ldots, M\}$. The overall goal of sequential testing is to achieve a certain level of accuracy for the final decision using the fewest number of observations on average.

We say a sequence of tests is *universally consistent* if the maximal error probability $P_{\max}$, defined as

$$P_{\max} \ \triangleq \ \max_{i=0,1,\ldots,M} \ \mathbb{P}_i \left\{ \delta \left(Y^{MN} \right) \neq i \right\},$$

vanishes for any $\mu, \pi, \mu \neq \pi$. Further, we say it is *universally exponentially consistent* if under each hypothesis, the exponent for the maximal error probability with respect to the expected stopping time is strictly positive, i.e., there exists $\alpha_i > 0$ such that

$$\mathbb{E}_i \left[N \right] \ \leq \ \frac{-\log P_{\max}}{\alpha_i} \left(1 + o(1) \right) \tag{3}$$

for any $\mu, \pi, \mu \neq \pi$ as $P_{\max} \to 0$.

4.1. Universal Sequential Test

The proposed sequential test stops when the GL for the most likely hypothesis is larger than those for all the competing hypotheses by a time-dependent threshold, if that happens not too late. In particular, the test can be described by the following stopping and final decision rules

$$N \ = \ \min \left(\tilde{N}, \lfloor T \log T \rfloor \right), \tag{4}$$

$$\delta \ = \ \begin{cases} \hat{i} \left(Y^{MN} \right) & \text{if } \tilde{N} \leq T \log T \\ 0 & \text{if } \tilde{N} > T \log T, \end{cases} \tag{5}$$

where

$$\tilde{N} \ \triangleq \ \operatorname*{argmin}_{n \geq 1} \left[\frac{\hat{p}_i^{\text{univ}} \left(y^{Mn} \right)}{\max\limits_{j \neq \hat{i}} \hat{p}_j^{\text{univ}} \left(y^{Mn} \right)} \ > \ T(n+1)^{M|\mathcal{Y}|} \right], \tag{6}$$

and $\hat{i}(y^{Mn}) \ \triangleq \ \operatorname*{argmax}_{i=1,\ldots,M} \hat{p}_i^{\text{univ}} \left(y^{Mn} \right)$ is the instantaneous estimate of the non-null hypothesis.

4.2. Result

The performance of the sequential test in (4) – (6) is characterized in the following theorem. We provide only a sketch of the proof due to space limitations.

Theorem 2. *The proposed sequential test in (4) – (6) is universally consistent, i.e., for any $\mu, \pi, \mu \neq \pi$, it holds that*

$$P_{\max} \ \to 0$$

as $T \to \infty$. The test also yields a positive exponent (cf. (3)) for the maximal probability of error under every non-null hypothesis (cf. Theorem 3 in [14]).

In addition, under every non-null hypothesis, as $M \to \infty$, the achievable error exponent of the test approaches the absolutely optimal one achievable when both μ, π are known (cf. Proposition 1 and (17) in [14]).

Proof. For each $i = 1, \ldots, M$, it follows from the convergence of the empirical distributions and the uniform integrability of the sequence of rvs $\{N/\log T\}$ that

$$\lim_{T \to \infty} \mathbb{E}_i \left[\left| \frac{N}{\log T} - \frac{1}{\alpha(\mu, \pi)} \right| \right] = 0, \tag{7}$$

where $\alpha(\mu, \pi) > 0$ for any $\mu, \pi, \mu \neq \pi$. For each $i = 1, \ldots, M$, we then establish using Sanov's theorem and the Markov inequality that

$$\mathbb{P}_i \left\{ \delta \neq i \right\} \leq \frac{C(\mu, \pi, |\mathcal{Y}|, M)}{T}, \tag{8}$$

where $C(\mu, \pi, |\mathcal{Y}|, M)$ is a constant indepenf of T. In addition, it follows from the definition of the test that

$$\mathbb{P}_0 \{ \delta \neq 0 \} \leq \frac{C'(|\mathcal{Y}|, M)}{T}. \tag{9}$$

The claim of universal consistency now follows from (8) and (9); and the claim of universal exponential consistency under each non-null hypothesis now follows from the combination of (7) – (9). □

Remark 1. *An interesting question here is whether one can set the value of the threshold T to satisfy a predefined level of test accuracy in the completely universal setting. It can be shown that although an arbitrarily small probability of error can be achieved with T sufficiently large, the exact value of T cannot be set unless we have an a priori lower bound for the distance between μ and π.*

5. APPLICATION TO SPAM DETECTION

We design an experiment relevant to spam detection to evaluate the performance of our tests on a real data set. The data set contains information from 4610 emails (each being labeled as a spam or non-spam) addressed to an employee at Hewlett-Packard (HP) [15]. The information for each email consists of relative frequencies of a set of 48 words and 6 punctuation marks. We shall refer to the relative frequencies of such words and punctuation marks as *features*. There are 1813 spams among the 4601 emails.

The specific application that we envision pertains to identifying spam sources of an individual email account. Consider the situation where an email account may be spammed by a few vicious IP addresses, which constitute a small fraction of all possible IP addresses. Cast into the formulation of outlier hypothesis testing, each sequence consists of emails from a certain IP address. When an account is compromised, a small subset of the sequences are outliers that contain only spams, while the majority of the sequences are typical with non-spams. The goal is to decide whether an email account is compromised, and if so, which are the sources of spams.

The experiment is designed such that there is exactly one outlier sequence among $M = 6$ number of sequences. The outlier sequence contains only spams, and typical sequences non-spams. It is known that the values of certain features, such as the relative frequencies of "RE", "FREE", the name of the recipient, and the name of the company where the recipient is employed ("HP" and HP laboratory ("HPL")), tend to vary greatly between spams and non-spams [15]. In this experiment, we choose the relative frequencies of "HP", "HPL" and "RE" as the observations. Specifically, the k-th observation of sequence i, $i = 1, \ldots, M$, is $y_k^{(i)} = \left(y_{k,1}^{(i)}, y_{k,2}^{(i)}, y_{k,3}^{(i)}\right)$, where $y_{k,1}^{(i)}$ is the relative frequency of "HP" in the correponding email, $y_{k,2}^{(i)}$ of "HPL", and $y_{k,3}^{(i)}$ of "RE." It is assumed that the coordinates of an observation are mutually independent, and identically distributed across the observations.

In the original data set, the features take continuous values in the finite interval of $[0, 100]$. The tests described in Section 3 and 4 are only applicable when the observations take values in finite alphabets. In order to apply our proposed tests, the observations are first quantized, where the quantization intervals of a certain feature are appropriately chosen based on the distribution of the feature values over all emails, regardless of their labels. Specifically, for a certain feature, the region in [0, 100] which finds the majority of the values of said feature is quantized more finely than other regions. There are 5 levels in the quantizations for "HP" and "HPL", and 6 levels for "RE". The value of each quantization interval is chosen to be the midpoint of that interval.

We apply the sequential test in (4) – (6) to the quantized data with a series of increasing thresholds T. For each T, the sequential test is repeated a number of trials using bootstrap samples (we randomly permute the emails when we run out of data, and reuse the permuted data). For comparison, we also evaluate the performance of various fixed sample size tests including the GL test in (2), and the MMD-based tests in [12]. One advantage of the MMD-based test is that it is applicable when the underlying distributions are continuous. In this experiment, we implement the MMD-based test using the original data (continuous), the quantized data, and the indices of the quantization intervals, respectively. The numerical results are obtained by averaging over a number of trials. It is shown in Figure 1 that the sequential test starts to outperform all the fixed sample size tests when the average stopping time exceeds 30. And the GL test outperforms all three MMD-based tests for large enough n, which agrees with the optimality result of the GL test in Theorem 1. In particular, the GL test outperforms the MMD-based tests when the length of the sequences n is larger than 20. This is due to the fact that an intermediate step of the GL test is to estimate the underlying distributions (cf. Section 3.1), which becomes more accurate as n increases.

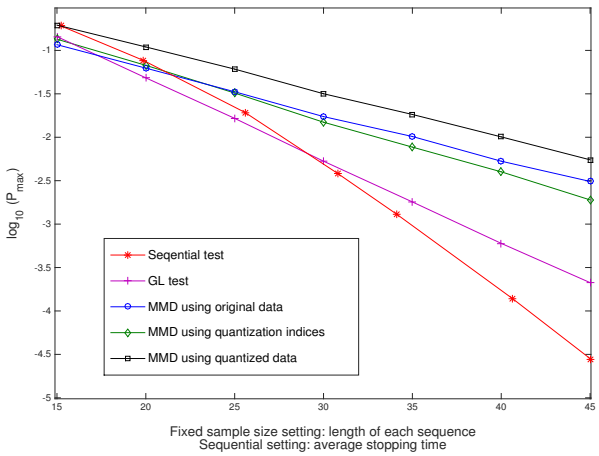

Fig. 1. Comparison between the sequential test and various fixed sample size tests.

6. REFERENCES

[1] V. Barnett, "The study of outliers: purpose and model," *Appl. Stat.*, vol. 27, no. 3, pp. 242–250, 1978.

[2] D. Hawkins, *Identification of Outliers*, Chapman and Hall, 1980.

[3] Y. Li, S. Nitinawarat, and V. V. Veeravalli, "Universal outlier hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 60, pp. 4066 – 4082, 2014.

[4] K. Pearson, "On the probability that two independent distributions of frequency are really samples from the same population," *Biometrika*, vol. 8, pp. 250–254, 1911.

[5] J. Ziv, "On classification with empirically observed statistics and universal data compression," *IEEE Trans. Inf. Theory*, vol. 34, pp. 278–286, 1988.

[6] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Inf. Theory*, vol. 35, pp. 401–408, 1989.

[7] V. P. Dragalin, A. G. Tartakovsky, V. V. Veeravalli, "Multihypothesis sequential probability ratio tests—part I: Asymptotic optimality," *IEEE Trans. Inf. Theory*, vol. 45, pp. 2448–2461, Nov. 1999.

[8] V. H. Poor, *An Introduction to Signal Detect and Estimation*, Springer, 1994.

[9] Y. Li, S. Nitinawarat and V. V. Veeravalli, "Universal sequential outlier hypothesis testing," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 7-12 2014, pp. 2666–2670.

[10] M. Woodroofe, *Nonlinear Renewal Theory in Sequential Analysis*, CBMS-NSF regional conference series in applied mathematics. SIAM, 1982.

[11] D. Siegmund, *Sequential Analysis: Tests and Confidence Intervals*, Springer series in statistics. Springer-Verlag, 1985.

[12] S. Zou, Y. Liang, V. H. Poor and X. Shi, "Nonparametric detection of anomalous data via kernel mean embedding," *IEEE Trans. Inf. Theory*, submitted, 2014.

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

[14] Y. Li, S. Nitinawarat, and V. V. Veeravalli, "Universal sequential outlier hypothesis testing," presented at the 48th Annual Asilomar Conf., Pacific Grove, CA, USA, Nov. 2-5, 2014.

[15] T. Hastie, R. Tibshirani, J. H. Friedman, *The elements of Statistical Learning*, Springer-Verlag, 2009.