OUTLYING SEQUENCE DETECTION IN LARGE DATASETS: COMPARISON OF UNIVERSAL HYPOTHESIS TESTING AND CLUSTERING

Yun Li[†] *and Venugopal V. Veeravalli*[*]

[†] Walmart Labs, San Bruno, CA 94066
[*] Department of ECE and Coordinated Science Laboratory
University of Illinois at Urbana-Champaign, Urbana, IL 61801

ABSTRACT

Multiple observation sequences are collected, among which there is a small subset of outliers. A sequence is considered an outlier if the observations therein are generated by a mechanism different from that generating the observations in the majority of sequences. In the universal setting, the goal is to identify all the outliers without any knowledge about the underlying generating mechanisms. In prior work, this problem was studied as a universal hypothesis testing problem, and a generalized likelihood test was constructed and its asymptotic performance characterized. Here a connection is made between the generalized likelihood test and clustering algorithms from machine learning. It is shown that the generalized likelihood test is equivalent to combinatorial clustering over the probability simplex with the Kullback-Leibler divergence being the dissimilarity measure. Applied to synthetic data sets for outlier hypothesis testing, the performance of the generalized likelihood test is shown to be superior to that of a number of other clustering algorithms for sufficiently large sample sizes.

***Index Terms*—** outlying sequence detection, universal outlier hypothesis testing, generalized likelihood test, cluster analysis, spectral clustering, combinatorial clustering

1. INTRODUCTION

The problem of interest in this paper is to identify a small subset (possibly empty) of outliers among multiple observation sequences. It is assumed that the observations in the majority of sequences are distributed according to a common "typical" distribution. A sequence is considered an outlier if the distribution underlying it is different from the common typical distribution. We are interested in a universal setting of the problem, where nothing is known about the outlier and typical distributions except that each outlier distribution is different from the typical distribution, and all of them have full support over a finite alphabet. The goal is to design a test, which does not depend on the outlier and typical distributions, to best discern all the outliers. Outlying sequence detection finds possible applications in anomaly detection in large data sets, spectrum sensing and high frequency trading.

The problem of outlying sequence detection was studied as a universal outlier hypothesis testing problem in both fixed sample size [1] and sequential [2] settings. For the fixed sample size setting, the main contribution in [1] was to show 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 [3–5]. The *exponential consistency* of the GL test under various universal settings was established in [1]. When there is a known number of identically distributed outliers, the GL test was also shown to be *asymptotically optimal*, i.e., as the number of sequences approaches infinity, the achievable error exponent of the GL test converges to the optimal one achievable when all underlying distributions are known.

The machine learning approach to the problem of outlying sequence detection treats it as a *clustering* problem. The goal of cluster analysis is to segment a collection of data objects into homogeneous subsets or "clusters", such that objects assigned to the same cluster are more closely related to one another than objects assigned to different clusters [6–9]. Similar to classification, cluster analysis creates labeling of the objects with class (cluster) labels. The labels are derived from the data in cluster analysis, whereas for classification, unlabeled objects are assigned a class label using a model developed from training objects with known labels.

The notion of similarity or dissimilarity is of central importance to a majority of clustering algorithms [6–9]. The dissimilarity between two data objects is a quantitative measurement that characterizes how closely related

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.

the two objects are. In cluster analysis, an object is usually described by a set of measurements. The dissimilarity between a pair of objects is often given by an appropriately chosen distance metric, which can be computed using such measurements. For instance, a popular choice of the distance metric is the Euclidean distance for continuous measurements, and Jaccard coefficient for binary measurements [10, 11]. Having chosen an appropriate dissimilarity measure, a clustering algorithm seeks to either maximize the similarity within clusters, or to minimize the proximity between clusters, or a combination of both [6–9].

In this paper, we study the connection between the approaches of universal outlier hypothesis testing and cluster analysis for outlying sequence detection. We show that the GL test [1] for universal outlier hypothesis testing is equivalent to combinatorial clustering over the probability simplex, with the Kullback-Leibler (KL) divergence as the dissimilarity measure. We compare the performance of the GL test and that of a number of clustering algorithms against a synthetic data set for outlying sequence detection. We show that the GL test outperforms the other clustering algorithms for large enough sample size.

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

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

The following technical facts will be useful; their derivations can be found in [12, Theorem 11.1.2]. Consider random variables Y^n which are i.i.d. according to $p \in \mathcal{P}(\mathcal{Y})$. Let $y^n \in \mathcal{Y}^n$ be a sequence with an empirical distribution $\gamma \in \mathcal{P}(\mathcal{Y})$. It follows that the probability of such sequence y^n, under p and under the i.i.d. assumption, is

$$p(y^n) = \exp\Big\{-n\big(D(\gamma\|p) + H(\gamma)\big)\Big\}, \quad (1)$$

where $D(\gamma\|p)$ and $H(\gamma)$ are the KL divergence of γ and p, and entropy of γ, defined as

$$D(\gamma\|p) \triangleq \sum_{y\in\mathcal{Y}} \gamma(y)\log\frac{\gamma(y)}{p(y)},$$

and

$$H(\gamma) \triangleq -\sum_{y\in\mathcal{Y}} \gamma(y)\log\gamma(y),$$

respectively.

3. UNIVERSAL OUTLIER HYPOTHESIS TESTING

Consider $M \geq 3$ independent sequences, each of which consists of n i.i.d. observations. The majority of the sequences are distributed according to a "typical" distribution $\pi \in \mathcal{P}(\mathcal{Y})$ except for a subset $S \subset \{1, \ldots, M\}$, $|S| < \frac{M}{2}$, of outlier sequences. Outlier sequences are distributed according to a common "outlier" distribution $\mu \in \mathcal{P}(\mathcal{Y})$. *Nothing is known about μ and π except that $\mu \neq \pi$, and that both of them have full support over a finite alphabet $\mathcal{Y}$.*

Let $Y_k^{(i)} \in \mathcal{Y}$ denote the k-th observation of the i-th sequence, and let $\mathcal{S}$ be the set comprising all possible outlier subsets. For the hypothesis corresponding to an outlier subset $S \in \mathcal{S}$, the joint distribution of all the observations is given by

$$\begin{aligned} p_S\left(y^{Mn}\right) &= L_S\left(y^{Mn}, \mu, \pi\right) \\ &= \prod_{k=1}^{n}\left\{\prod_{i\in S}\mu\left(y_k^{(i)}\right)\prod_{j\notin S}\pi\left(y_k^{(j)}\right)\right\}, \end{aligned} \quad (2)$$

where $L_S\left(y^{Mn}, \mu, \pi\right)$ denotes the likelihood, which is a function of the observations y^{Mn}, and μ and π.

In this paper, we consider models with at least one and up to $K, 1 \leq K < \frac{M}{2}$, identically distributed outliers, where K is known at the outset. Models with an unknown number (possibly zero) of distinctly distributed outliers are studied in [1]. We also restrict our attention to the fixed sample size setting where the number of observations in each sequence is fixed at the outset. The results for the sequential setting can be found in [2].

A test for the outlier subset is done based on a *universal* rule $\delta : \mathcal{Y}^{Mn} \to \mathcal{S}$. In particular, the test δ is not allowed to be a function of the unknown distributions μ and π. The accuracy of a universal test is gauged using the maximal probability of error

$$P_{\max} \triangleq \max_{S\in\mathcal{S}} \mathbb{P}_S\left\{\delta\left(y^{Mn}\right) \neq S\right\},$$

which is a function of the test δ, and the underlying distributions μ and π. 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. Generalized Likelihood Test

For each $i = 1, \ldots, M$, denote the empirical distribution of $\boldsymbol{y}^{(i)}$ by γ_i. In the universal setting with μ and π being unknown, conditioned on the set of outliers being $S \in \mathcal{S}$, we compute the generalized likelihood (GL) of y^{Mn} by replacing μ and π in (2) with their maximum likelihood (ML) estimates $\hat{\mu}_S \triangleq \frac{\sum_{i \in S} \gamma_i}{|S|}$, and $\hat{\pi}_S \triangleq \frac{\sum_{j \notin S} \gamma_j}{M - |S|}$, as

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

The GL test then selects the hypothesis under which the GL is maximized, i.e.,

$$\delta\big(y^{Mn}\big) \;=\; \underset{S \subset \{1,\ldots,M\},\ |S| \leq K}{\operatorname{argmax}} \hat{p}_S^{\text{univ}}, \tag{3}$$

where ties are broken arbitrarily.

3.2. Performance of GL Test

When the number of outliers is known, i.e., when $|S| = K$ for all $S \in \mathcal{S}$, the GL test in (3) is shown in [1] to be exponentially consistent, and it also achieves the absolutely optimal error exponent asymptotically as the number of sequences M approaches infinity.

Theorem 1. *For each $M \geq 3$, when the number of outliers is known, the GL test δ in (3) is universally exponentially consistent for any $\mu, \pi, \mu \neq \pi$ (cf. Theorem 9 in [1]).*

Furthermore, as $M \to \infty$, the achievable error exponent of the GL test in (3) converges to the optimal one achievable when both μ and π are known (cf. Theorem 10 in [1]).

For more general models with at most K outliers, $1 < K < M/2$, the GL test achieves exponential consistency universally as long as the outlier distribution is distinct from the typical one.

Theorem 2. *For each $M \geq 3$, when there is at least one and up to K outliers, the GL test in (3) is universally exponentially consistent for any $\mu, \pi, \mu \neq \pi$ (cf. Theorem 11 in [1]).*

4. CONNECTION TO CLUSTER ANALYSIS

Another approach to outlying sequence detection is to treat it as a clustering problem in the domain of machine learning [6–9]. In outlying sequence detection, an entire sequence can be considered an object. Typical sequences are more closely related to one another than to any outlier sequence in the sense that the observations therein are distributed according to the same typical distribution. The same holds for outlier sequences when the outliers are identically distributed. Under such assumption, outliers can be identified by clustering the sequences into two clusters, where the cluster with more members contains all typical sequences, and the other outliers. Interestingly, we shall show that the GL test in (3) can be interpreted as combinatorial clustering over the probability simplex that has the KL divergence as the dissimilarity measure.

4.1. Combinatorial Clustering

Consider vector observations $x_i, i = 1, \ldots, m$, taking values in $\mathbb{R}^p$, $p \geq 1$. The goal is to cluster these vector observations into T clusters where T is known at the outset. We index each cluster by a unique integer $t = 1, \ldots, T$. A cluster assignment is completely specified by a many-to-one mapping C such that for each observation $x_i, i = 1, \ldots, m$, the cluster label of x_i is given by $C(i), C(i) \in \{1, \ldots, T\}$.

Combinatorial clustering often starts with a "loss" function, which characterizes the extend to which the clustering goal is *not* met [7]. A natural way to construct such a loss function is to use a pairwise dissimilarity measure. Let $d(x_i, x_j)$ denote the dissimilarity between x_i and x_j, $i, j = 1, \ldots, m$. Having chosen an appropriate dissimilarity measure, a natural loss function is given by

$$\begin{aligned} W(C) &= \sum_{t=1}^{T} w(t) \\ &= \sum_{t=1}^{T} \sum_{i:C(i)=t} \sum_{j:C(j)=t} d(x_i, x_j), \end{aligned} \tag{4}$$

where $w(t)$ is called the *point scatter* of the t-th cluster, and $W(C)$ the *within-cluster pointer scatter* associated with the cluster assignment C.

Let $\mathcal{C}_{m,T}$ denote the set comprising *all* possible cluster assignments of the m observations into T clusters. One then selects the cluster assignment $C^\star$ that minimizes the value of the loss function, i.e.,

$$C^\star = \underset{C \in \mathcal{C}_{m,T}}{\arg\min}\ W(C).$$

4.2. GL Test as Combinatorial Clustering

It is straightforward to show using (1) that the GL test in (3) is equivalent to

$$\begin{aligned} \delta\big(y^{Mn}\big) \;=\; \underset{S \subset \{1,\ldots,M\},\ |S| \leq K}{\operatorname{argmin}} & \sum_{i \in S} D\left(\gamma_i \,\middle\|\, \tfrac{\sum_{l \in S} \gamma_l}{|S|}\right) \\ & + \sum_{j \notin S} D\left(\gamma_j \,\middle\|\, \tfrac{\sum_{k \notin S} \gamma_k}{M - |S|}\right). \end{aligned} \tag{5}$$

The GL test can be interpreted as follows. Each sequence of observations is represented by its corresponding empirical distribution on the probability simplex. When the outliers are identically distributed, it suffices to cluster the empirical distributions into two clusters, where the larger cluster contains the empirical distributions of all typical sequences, and the smaller cluster outliers. A cluster assignment is completely specified by the set of outliers S. For a particular cluster, define the point scatter as the sum of the KL divergences between each individual cluster member (an empirical distribution) and the cluster center (the average of all empirical distributions, that is, a mixture distribution). Given a cluster assignment S, the objective function of the optimization problem in (5) corresponds to the within-cluster point scatter defined in (4). The GL test decides on the cluster assignment that minimizes the within-cluster point scatter among all possible cluster assignments, each specified by $S \in \mathcal{S}, |S| \leq K$. We now see that the GL test is equivalent to combinatorial clustering over the probability simplex that has the KL divergence as the dissimilarity measure.

Remark 1. *The original data objects (sequences of length n) are each represented by its empirical distribution on the probability simplex. The dimension of these empirical distributions is equal to the size of the alphabet, which does not increase as the length of the sequences, n, increases.*

It is possible to replace the KL divergence in (5) with an alternative dissimilarity measure such as the l_2 distance. Our theoretical results in Theorem 1 suggest that the KL divergence is the asymptotically optimal dissimilarity measure for certain settings of outlying sequence detection. Specifically, when there is a known number of identically distributed outliers, as M approaches infinity, the achievable error exponent of the GL test in (5) converges to the absolutely optimal one achievable when both μ and π are known (cf. Theorem 1). It is not known if the same asymptotic optimality will continue to hold with other choices of the dissimilarity measure.

5. NUMERICAL RESULTS

We evaluate the performance of two universal tests and a number of clustering algorithms against synthetic data sets for outlying sequence detection. We apply four different clustering algorithms: two combinatorial clustering algorithms using the l_2 distance, and the Hamming distance, respectively, and two spectral clustering algorithms using the Hamming distance, and a Hamming-like distance defined for any pair of sequences, respectively. For comparison, we also apply the GL test in (5), and another universal test based on the l_2 distance. For each data set, the GL test in (5) outperforms all other algorithms for large enough sample sizes.

Due to the limited space, we focus on the results relevant to the GL test and two other algorithms. In particular, we compare the GL test with a spectral clustering algorithm due to Ng, Jordan, and Weiss [13]. We adopt a pairwise Hamming-like distance to measure the similarity between two sequences (cf. Section 2 in [13]), i.e., the similarity between sequences i and j, $i, j \in \{1, \ldots, M\}$, is

$$d(i,j) \triangleq \sum_{k=1}^{n} \sum_{l=1}^{n} \mathbb{I}(Y_k^{(i)} = Y_l^{(j)}).$$

In additional, we consider another universal test using the l_2 distance as the dissimilarity measure. Specifically, this test solves the same optimization problem as in (5), but with the KL divergence being replaced by the l_2 distance in the objective function in (5).

In this comparison, the particular choice of typical and outlier distributions are $\pi = (0.25, 0.41, 0.34)$ and $\mu = (0.1, 0.55, 0.35)$. There is exactly one outlier among $M = 5$ sequences. For different sample size n, we evaluate the maximum probability of error incurred by various algorithms. As we can see from Figure 1, for this synthetic data set, the spectral clustering using the Hamming-like distance outperforms the GL test when n is small. For sufficiently large n, the GL test outperforms the other two algorithms. These results suggest that for outlying sequence detection, it may be beneficial to use spectral clustering when the number of observations is limited. For n sufficiently large, the simulation results corroborate our theoretical findings in Theorem 1, which establishes the asymptotic optimality of the KL divergence as a dissimilarity measure for this particular setting of outlying sequence detection (cf. Remark 1).

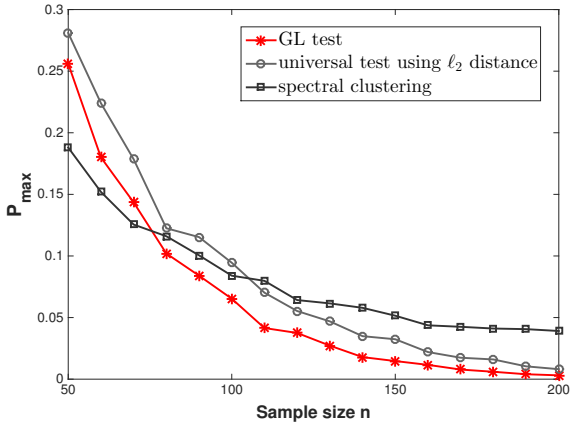

Fig. 1. Compare the GL test with two other algorithms

6. REFERENCES

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

[2] 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.

[3] 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.

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

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

[6] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Cluster Analysis*, John Wiley and Sons, Inc., fifth edition, 2011.

[7] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements in Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2009. Available at `http://statweb.stanford.edu/tibs/ElemStatLearn/`.

[8] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, The MIT Press, 2012.

[9] P. Than, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Pearson, 2005. Available at `http://www-users.cs.umn.edu/~kumar/dmbook/index.php`.

[10] J. C. Gower and P. Legendre, "Metric and Euclidean properties of dissimilarity coefficients," *J. Classification*, vol. 5, pp. 5–48, 1986.

[11] M. Schwaiger and O. Opitz, *Exploratory Data Analysis in Empirical Research*, Springer, 2002.

[12] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, New York: John Wiley and Sons, Inc., 2006.

[13] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Adv. Neural Inf. Process. Syst*, vol. 14, pp. 849–856, 2002.