Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Using labeled data to evaluate change detectors in a multivariate streaming environment

Albert Y. Kim [a], Caren Marzban [a,b], Donald B. Percival [b,a,*], Werner Stuetzle [a]

[a] Department of Statistics, Box 354322, University of Washington, Seattle, WA 98195 4322, USA
[b] Applied Physics Laboratory, Box 355640, University of Washington, Seattle, WA 98195 5640, USA

ARTICLE INFO

Article history:
Received 12 May 2008
Received in revised form
8 April 2009
Accepted 13 April 2009

Keywords:
Block bootstrap
Change point detection
Time series analysis

ABSTRACT

We consider the problem of detecting changes in a multivariate data stream. A change detector is defined by a detection algorithm and an alarm threshold. A detection algorithm maps the stream of input vectors into a univariate detection stream. The detector signals a change when the detection stream exceeds the chosen alarm threshold. We consider two aspects of the problem: (1) setting the alarm threshold and (2) measuring/comparing the performance of detection algorithms. We assume we are given a segment of the stream where changes of interest are marked. We present evidence that, without such marked training data, it might not be possible to accurately estimate the false alarm rate for a given alarm threshold. Commonly used approaches assume the data stream consists of independent observations, an implausible assumption given the time series nature of the data. Lack of independence can lead to estimates that are badly biased. Marked training data can also be used for realistic comparison of detection algorithms. We define a version of the receiver operating characteristic curve adapted to the change detection problem and propose a block bootstrap for comparing such curves. We illustrate the proposed methodology using multivariate data derived from an image stream.

1. Introduction

We consider the problem of detecting changes in a multivariate data stream. We want to assess whether the most recently observed data vectors (the “current set”) differ in some significant manner from previously observed vectors (the “reference set”). Change detection is of interest in a number of applications, including neuroscience [3], surveillance [7], seismology [18], voice activity detection [8] and identification of activity periods in radar, sonar and biomedical signals using a known template (see also [16,17] and references therein).

The notion of change is often formalized in terms of distributions: vectors in the current set are assumed to be sampled from some multivariate distribution Q, whereas those in the reference set are assumed to come from a (possibly different) distribution P. The task of a change detector then is to test the hypothesis $P = Q$ given the two samples. We obtain a new value of the test statistic every time a new observation arrives. We flag a change as soon as the test statistic exceeds a chosen alarm threshold [10,12,14].

In a concrete application of this recipe we face a number of choices such as picking a two-sample test that

[*] Corresponding author at: Applied Physics Laboratory, Box 355640, University of Washington, Seattle, WA 98195 5640, USA.
E-mail addresses: albert@stat.washington.edu (A.Y. Kim), marzban@stat.washington.edu (C. Marzban), dbp@apl.washington.edu (D.B. Percival), wxs@stat.washington.edu (W. Stuetzle).
URLS: http://www.stat.washington.edu/albert (A.Y. Kim), http://faculty.washington.edu/marzban (C. Marzban), http://faculty.washington.edu/dbp (D.B. Percival), http://www.stat.washington.edu/wxs (W. Stuetzle).

0165-1684/$ - see front matter © 2009 Published by Elsevier B.V.
doi:10.1016/j.sigpro.2009.04.011

is sensitive toward changes of interest; choosing the sizes of the current and reference sets; and choosing an alarm threshold that results in the desired tradeoff between false alarms and missed changes. More complicated schemes are possible involving, e.g., multiple two-sample tests used in parallel and adoption of a more complex notion of "change". No matter what the details, ultimately we will end up with a univariate stream that we call the "detection stream". We flag a change whenever the detection stream exceeds a chosen alarm threshold. Abstracting away details, a change detector can be defined as a combination of a detection algorithm mapping the multivariate input stream $\mathbf{x}_t$ into a univariate detection stream d_t, and an alarm threshold τ. The only fundamental restriction is that d_t can only depend on input observed up to time t.

In this paper we focus on two problems: (i) choosing between different detection algorithms and (ii) selecting an alarm threshold to obtain a desired false alarm rate. We assume the existence of labeled training data, i.e., a segment of the stream where changes of interest have been marked. To quantify the performance of a detection algorithm, we propose an adaptation of the standard receiver operating characteristic (ROC) curve (Section 3). A resampling method similar to the block bootstrap lets us compare the ROC curves of different detection algorithms on the labeled data in a statistically meaningful way (Section 5). The labeled data also allow us to determine the alarm threshold for a desired false alarm rate without the usual assumption that vectors in the stream are observations of independent random variables, which is implausible when observing a time series. If the assumption is violated, estimates of the false alarm rate based on this assumption can be wildly off the mark (Section 4). We illustrate our main points using a multivariate data stream derived from a series of images of Portage Bay in Seattle (Sections 2 and 6). Section 7 concludes the paper with a summary and some ideas for future work.

2. Data

To illustrate the ideas in this paper, we created a multivariate data stream from a sequence of images recorded with a web camera operated by the Sound Recording for Education (SORFED) project at the Applied Physics Laboratory, University of Washington. The camera is mounted on a barge several feet above the water in Portage Bay, Seattle, and usually takes images at 2 s intervals. We use a sequence of 5002 images recorded on June 27, 2007, and divide the 168×280 pixels in each image into a 14×20 grid of bins, with each of the 280 bins containing 168 pixels. We summarize each bin by its average gray level, resulting in a stream of 280-dimensional data vectors.

Motivated by potential applications of change detection to surveillance, we decided to regard the appearance of boats in the image stream as changes of interest. We looked at each of the 5002 images and manually marked the bins in each image containing a boat passing through Portage Bay. Fig. 1 shows one such image, with four bins marked as containing a boat. Fig. 2 shows the number of marked bins for each image plotted against image index. We define a "boat event" as a sequence consisting of two or more consecutive images with at least one marked bin. There are 19 boat events in all, and their location and extent are indicated by the black rectangles at the bottom of Fig. 2. There are 20 quiescent periods surrounding the boat events. The images during the quiescent periods are quite variable because of light variations on the water from cloud movement, ducks moving around in the water close to the camera, wind-driven ripples in the water, wakes from boats no longer in view of the camera, and other sources of noise.

We emphasize that we use the images primarily as a means for constructing a multivariate data stream with characteristics that one would expect in actual applications of change detection, but that are not typically present in simulated data (e.g., correlated and heterogeneous noise). We do not make use of the neighborhood structure among the 280 variables; in fact, all of the results we present would be exactly the same if we were to randomly reorder the variables. In short, the methods we propose are not specific to image streams.

3. Quantifying the performance of a change detector

Defining a general measure quantifying the performance of a change detector for streams $\mathbf{x}_t$ is a nontrivial problem. Generally there are two kinds of errors, missed changes and false alarms, but appropriate definitions for these are application dependent. Consider a simple scenario where the stream consists of stretches during which the $\mathbf{x}_t$ are independent and identically distributed (IID). Suppose a change occurs at time t, but an alarm rings later on. It is not clear if this should be chalked up as a correct (but delayed) alarm or a missed change followed by a false alarm. In addition, even if the $\mathbf{x}_t$ are independent, the detection stream d_t is typically correlated, leading to false alarms that occur in bursts and forcing us to choose between counting individual false alarms or counting bursts. The piecewise IID assumption is also questionable: in our motivating example, it makes more sense to think of the stream as a concatenation of quiescent periods (no boats), interrupted by events (activity periods with boats present). During events, the distribution of $\mathbf{x}_t$ is not constant due to boat movement, and it might not be constant during quiescent periods either because of, e.g., lighting changes from the passage of clouds.

Raising an alarm soon after the start of an event is crucial for surveillance: if the alarm occurs too long after the start, the horse will have left the barn, and the alarm is useless. Changes within events or transitions from events to quiescent periods are not of interest. We define an event to be successfully detected if the detection stream exceeds the alarm threshold τ at least once within a tolerance window of width N_W after the event's onset. We define the hit rate $h(\tau)$ as the proportion of successfully detected events. The false alarm rate $f(\tau)$ is simply the proportion of times in the quiescent periods during

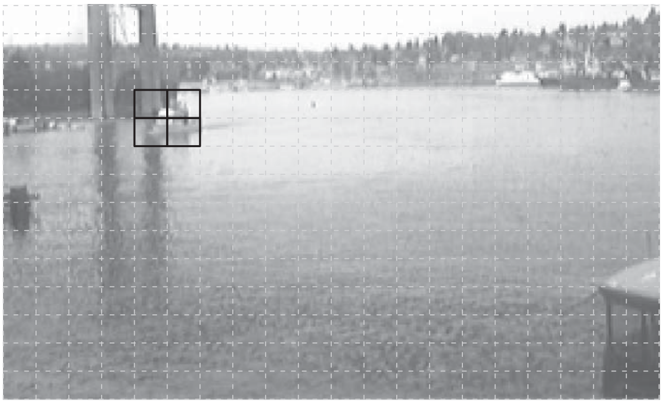

Fig. 1. Picture taken by a web camera overlooking Portage Bay, Seattle. The picture has been divided into a 14 × 20 grid of rectangular bins, four of which are highlighted and contain a boat passing through the bay.

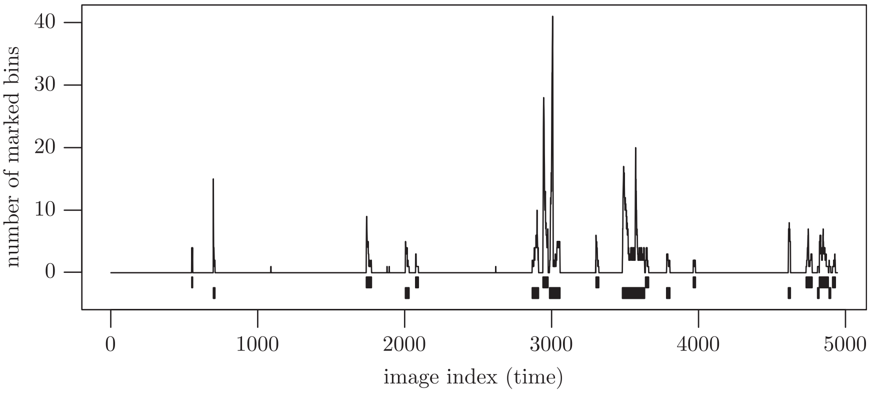

Fig. 2. Number of bins (variables) marked as containing a boat versus image index (top part of plot), along with markers for the 19 boat events (bottom).

which the detection stream exceeds the alarm threshold. There is no penalty for raising multiple alarms during an event. Our definitions for $h(\tau)$ and $f(\tau)$ are admittedly simple, and others might be better in scenarios not involving surveillance; however, our method for comparing change detectors (Section 5) is not dependent on these particular definitions.

We can summarize the performance of a change detection algorithm by plotting the hit rate $h(\tau)$ versus the false alarm rate $f(\tau)$ as we increase the alarm threshold τ. Both $h(\tau)$ and $f(\tau)$ are monotonically non-increasing functions of τ. The graph of the curve $\tau \longrightarrow (f(\tau), h(\tau))$ is a monotonically non-decreasing function of $f(\tau)$. We call this curve the ROC curve for the algorithm since it is similar to the standard ROC curve used to evaluate binary classifiers [6].

It is useful to compare the performance of a detection algorithm with a "null" algorithm that ignores the data and signals an alarm with probability $\alpha \in [0,1]$ independently at each time t. Clearly the false alarm rate for this algorithm is α. The rate at which this algorithm will successfully flag an event is given by the probability that an alarm is raised at least once within the tolerance window N_W, which is governed by a binomial distribution with parameters N_W and α. The ROC curve of the null algorithm is thus $\alpha \longrightarrow (\alpha, 1-(1-\alpha)^{N_W})$.

4. Setting the alarm threshold

A critical parameter of a change detector is the alarm threshold τ, which controls the tradeoff between false alarms and missed changes. Without training data that mark changes of interest, there is no way of realistically assessing the hit rate $h(\tau)$ for a given τ. The commonly proposed approach to setting τ is therefore to choose a false alarm rate α considered acceptable and then determine the corresponding τ. If we assume a piecewise IID model, we can sometimes analytically determine the appropriate value of τ. If an explicit calculation is not feasible, we can resort to a computational approach based on a permutation argument [1,5,9]. Assuming there is no change at or before the current time T, then $\mathbf{x}_1, \ldots, \mathbf{x}_T$ would be IID. To test the IID hypothesis, we compare the current value d_T^{orig} of the detection stream to values $d_T^1, \ldots, d_T^M$ obtained by applying the detection algorithm to M random permutations of $\mathbf{x}_1, \ldots, \mathbf{x}_T$. If d_T^{orig} is the k-th largest among $\{d_T^{orig}, d_T^1, \ldots, d_T^M\}$, then we can reject the IID hypothesis at level $k/(M+1)$. If the level is less than the desired false alarm rate, we signal a change and "reset the clock" by discarding $\mathbf{x}_1, \ldots, \mathbf{x}_T$. (Note that in this case the detection threshold will vary with time.)

The problem with the analytical and permutation-based approaches is that their validity depends critically on the piecewise IID assumption, which seems inherently implausible since we are observing a time series. If it is violated, the results can be wildly off the mark, a fact we can demonstrate using a simple detection algorithm based on a two-sample test. Detection algorithms based on such tests have been discussed previously (see, e.g., [10] and references therein). The idea is to compare the distribution P of the most recently observed data with the distribution Q of a reference set observed earlier. The value d_T of the detection stream is the test statistic of the two-sample test, and the (nominal) false alarm rate for detection threshold τ is the probability that $d_T \geqslant \tau$ under the null hypothesis $P = Q$ (the qualifier "nominal" is a reminder that the significance level is derived under the IID assumption). For our illustration we assume the data stream to be one dimensional and define the current and reference sets to be the N_C most recent observations and the N_R immediately preceding observations. The square of a two-sample t-test forms the detection stream at the current time T:

$$d_T = \frac{(\bar{x}_C - \bar{x}_R)^2}{\left(\dfrac{1}{N_C} + \dfrac{1}{N_R}\right)\hat{\sigma}^2},$$

where $\bar{x}_C$ and $\bar{x}_R$ are the sample means of the current and reference sets, and

$$\hat{\sigma}^2 = \frac{1}{N_C + N_R - 2}\left[\sum_{n=0}^{N_C-1}(x_{T-n} - \bar{x}_C)^2 + \sum_{n=0}^{N_R-1}(x_{T-N_C-n} - \bar{x}_R)^2\right]$$

is the pooled variance estimate. Although the t-test is designed to test the null hypothesis that the current and reference sets have the same mean values, we can still use it as a test of the IID hypothesis, recognizing that it might have little or no power for detecting changes other than mean shifts.

If we are willing to assume that the observations in the current and reference sets are realizations of IID Gaussian random variables, then the threshold τ for false alarm rate α is the square of the $\alpha/2$ quantile of the t distribution with $N_C + N_R - 2$ degrees of freedom. If we drop the Gaussianity assumption, we can use the permutation approach described above. The problem in either case is that the actual false alarm rate can be vastly different from the desired (nominal) rate if the independence assumption is violated.

As an example, choose $N_C = 4$, $N_R = 16$, and let $X_1, \ldots, X_{20}$ be a segment of an autoregressive process $X_t = \phi X_{t-1} + \varepsilon_t$, where $|\phi| < 1$ is the correlation between X_{t-1} and X_t, and the ε_t are IID standard Gaussian. If $\phi = 0$, the X_t are IID; if $\phi \neq 0$, they are no longer independent. The alarm threshold for false alarm rate $\alpha = 0.1$ when $\phi = 0$ is $\tau \doteq 3$ (the square of the 5th percentile for a t distribution with 18 degrees of freedom). For five different ϕ, we simulate 10,000 independent realizations of $X_1, \ldots, X_{20}$ and compute d_{20} for each realization. We then estimate α using the fraction of times when $d_{20} > 3$ in the 10,000 realizations. Table 1 shows that, as expected, the false alarm rate is close to 0.1 when $\phi = 0$, but is dramatically off the mark otherwise.

Table 1
False alarm rate α for the squared two-sample t-test using a threshold level of $\tau = 3$ and data generated from a Gaussian first-order autoregressive process with a unit-lag autocorrelation of ϕ.

ϕ	−0.9	−0.5	0	0.5	0.9
α	0.008	0.018	0.098	0.282	0.537

To illustrate the failure of the permutation approach, we generate an additional 1000 independent realizations of $X_1, \ldots, X_{20}$ for our selected values of ϕ. For each realization, we generate 1000 random permutations, compute d_{20} and keep track of the proportion of times that $d_{20} > 3$—this proportion is what a permutation test would declare the false alarm rate to be for $\tau = 3$. When averaged over all 1000 realizations of the AR process, this proportion is very close to 0.1 for all five values of ϕ: the permutation approach gives the correct false alarm rate when $\phi = 0$ (the IID case) but it underestimates (overestimates) the correct rate α when $\phi > 0$ ($\phi < 0$), with the discrepancy becoming more serious as ϕ approaches 1 (-1). We conclude that the permutation-based approach for setting the alarm threshold is not viable in the presence of correlated data (the usual case when dealing with time series).

5. Comparing change detectors

In this section we propose a method for evaluating the relative performance of change detectors that takes into account sampling variability.

Suppose we have two change detectors with ROC curves $\tau \longrightarrow (f_1(\tau), h_1(\tau))$ and $\tau \longrightarrow (f_2(\tau), h_2(\tau))$. There are two obvious ways to use these curves for assessing the relative performance of the detectors. For a given hit rate $h_1(\tau_1) = h_2(\tau_2) \equiv h$, we can compare the false alarm rates $f_1(\tau_1)$ and $f_2(\tau_2)$ and declare the first detector to be better if $f_1(\tau_1) < f_2(\tau_2)$; alternatively, for a given false alarm rate, we can compare hit rates. More elaborate comparison schemes are possible [15]. In our boat example, we define the hit rate $h(\tau)$ in terms of the onset of a small number of events, so it is easier to compare the false alarm rates for a given hit rate. This approach yields false alarm rates for the two detectors that are functions of h. We denote these functions as $f_1(h)$ and $f_2(h)$ and compare them using the ratio

$$r_{1,2}(h) = \frac{\max(f_1(h), \varepsilon)}{\max(f_2(h), \varepsilon)}, \quad (1)$$

where ε is a small number that allows the false alarm rate to be zero.

We use a modified version of the block bootstrap to assess if $r_{1,2}(h)$ is significantly different from unity. Block bootstrapping is an adaptation of the standard bootstrap designed for use with time series [4,11,19]. In the standard bootstrap, the basic unit for resampling is an individual observation; in a block bootstrap, it is a block of consecutive observations, with each block having the same size. The block size is selected such that, within a block, the dependence structure of the original time series is preserved, while values at the beginning and end of each block are approximately independent. Our input stream is naturally broken up into blocks of unequal size, namely, boat events and quiescent periods. We use these blocks to define the basic unit in two modified block bootstraps. The first is an "uncoupled" bootstrap. Given n_e boat events and $n_q = n_e + 1$ quiescent periods, we resample (with replacement) n_e boat events and n_q quiescent periods to form a bootstrap sample with the same structure as the original stream (n_q quiescent periods separated by n_e events). The second is a "coupled" bootstrap, in which the basic unit is taken to be an event and the quiescent period immediately following it. The motivation for this scheme is to preserve potential dependence between the quiescent period following an event and the event itself due to boat wakes.

The method for comparing detectors is the same for the coupled and the uncoupled bootstraps. For a given bootstrap sample, we evaluate $f_1(\tau)$, $h_1(\tau)$, $f_2(\tau)$ and $h_2(\tau)$ over a grid of thresholds τ, from which we calculate the curve $r_{1,2}(h)$. We repeat this procedure n_b times, yielding n_b bootstrap replicates of $r_{1,2}(h)$. We then construct $(1-\alpha)$ two-sided non-simultaneous confidence intervals for the ratio $r_{1,2}(h)$ based upon the empirical distribution of the bootstrap replicates. (The "matched pair" design by which we evaluate $f_1(\tau)$ and $f_2(\tau)$ for each bootstrap sample and then compute confidence intervals for $r_{1,2}(h)$ will lead to sharper comparisons than an unmatched design in which bootstrap samples are generated separately for each detector.) As we vary h, the end points of these confidence intervals trace out confidence bands. If the confidence interval for $r_{1,2}(h)$ does not include unity, we have evidence at the $(1-\alpha)$ confidence level that one change detector outperforms the other in the sense that it has a smaller false alarm rate for hit rate h.

6. An illustrative example

In this section we illustrate the methodology presented in the previous sections by considering two change detectors that are designed to detect the boat events described in Section 2. We define the two-sample tests behind the change detectors in Section 6.1, after which we demonstrate the pitfalls of using a permutation approach to determine the false alarm rate (Section 6.2). We then illustrate how we can compare the performance of the two change detectors in a manner that takes into account sampling variability (Section 6.3).

6.1. *Definition of detection streams based on two-sample test statistics*

The two detectors we use to illustrate our methodology are quite different in their intent, but both are based on two-sample tests. The first detector is designed to be sensitive to mean changes, while the second uses a nonparametric test with power against all alternatives. To simplify notation, we define the tests for samples $\mathbf{c}_1, \ldots, \mathbf{c}_n$ (the current set) and $\mathbf{r}_1, \ldots, \mathbf{r}_m$ (the reference set), with the understanding that we would obtain the values of the corresponding detection streams at the current time T by comparing the $n = N_C$ most recent observations with the $m = N_R$ observations immediately preceding them.

The first detection stream, denoted as $d_T^{(\max)}$, is based on the largest squared element of the vector $\bar{\mathbf{c}} - \bar{\mathbf{r}}$, where $\bar{\mathbf{c}}$ is the average of $\mathbf{c}_1, \ldots, \mathbf{c}_n$, and $\bar{\mathbf{r}}$ is similarly defined. The detection stream will be large if there has been a recent large change in one of the 280 variable in the input stream, i.e., a large change in mean gray level for one of the bins in the image. Boats are small and their appearance changes the mean gray level for a small number of bins; therefore we want a test that is sensitive to large changes in a few bins, rather than to small changes in a large number of bins.

The second change detector we consider is based on a so-called "energy" test statistic that has been advocated as a nonparametric test for equality of two multivariate distributions [2,20,22–24]. This statistic is given by

$$d_T^{(\mathrm{e})} = \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\|\mathbf{c}_i - \mathbf{r}_j\| - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\|\mathbf{c}_i - \mathbf{c}_j\| - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\|\mathbf{r}_i - \mathbf{r}_j\|,$$

where $\|\cdot\|$ denotes the Euclidean norm. This test is consistent against all alternatives to H_0 and hence is not focused on any particular aspect of the difference in distribution between the current and reference sets [24]. Because it is an omnibus test, it cannot be expected to have as much power for detecting a change in means as a test specifically designed for that type of change. Fig. 3 shows the detection streams $d_T^{(\max)}$ (top pane) and $d_T^{(\mathrm{e})}$ (bottom) plotted against time for the case $N_C = 4$ and $N_R = 16$.

6.2. *Pitfalls of setting the alarm threshold via permutation tests*

To complement the simulated example of Section 4, we now present an empirical demonstration of our assertion that we cannot expect to get reasonable estimates of the false alarm rate using a permutation argument.

We apply the change detector based on the energy test statistic with $N_C = 4$ and $N_R = 16$ to the longest quiescent period in our boat data (1030 images). For each of the 1011 segments of length 20 we calculate the permutation-based p-value (i.e., the observed level of significance) of the energy test statistic: we compare the original value of the test statistic for the segment with a reference set of 500 "permuted" values obtained by applying the test to a randomly shuffled version of the segment. If the original value is the k-th largest amongst these 501 values, then the p-value is $\hat{\alpha} = k/501$ [1].

Please cite this article as: A.Y. Kim, et al., Using labeled data to evaluate change detectors in a multivariate streaming environment, Signal Process. (2009), doi:10.1016/j.sigpro.2009.04.011

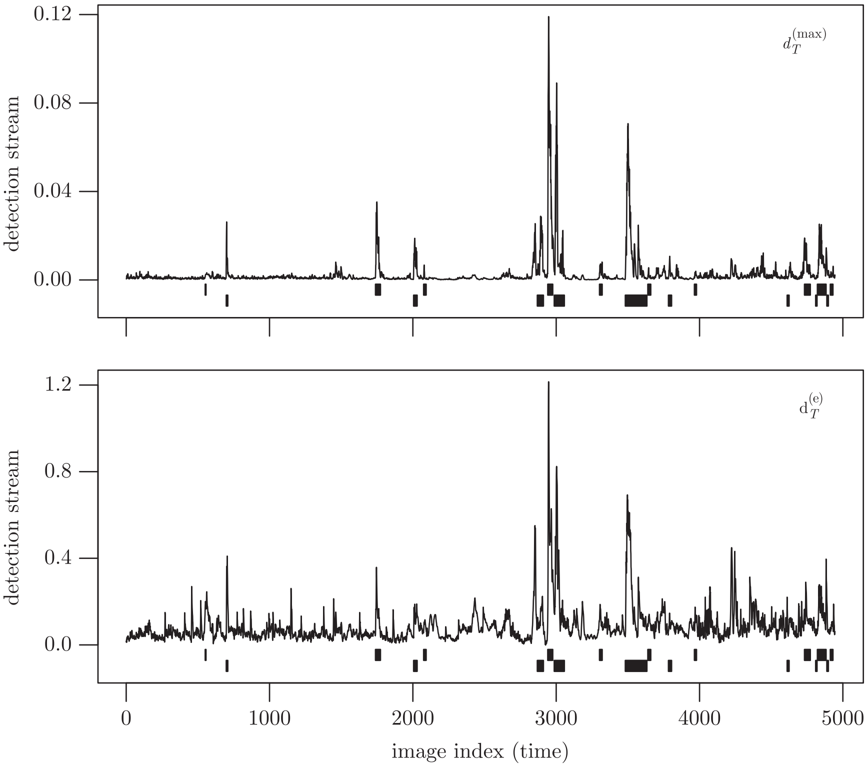

Fig. 3. Two detection streams plotted versus image index T, with boat events marked as in Fig. 2. The top plot shows $d_T^{(\max)}$, which is based upon the maximum squared difference in means; the bottom is for $d_T^{(e)}$, which is based upon the energy test statistic. The settings $N_C = 4$ and $N_R = 16$ are used for both detectors at each current time T.

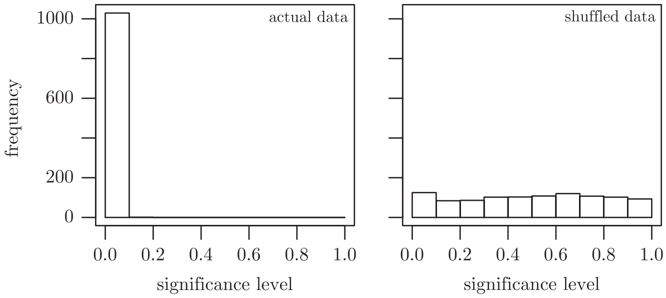

Fig. 4. Histograms of p-values (levels of significance) as empirically determined by a permutation test based upon data from a quiescent period (left-hand plot) and upon data from the same period but randomly shuffled (right-hand).

Since we are dealing with a quiescent period, the distribution of $\hat{\alpha}$ across all 1011 values in the detection stream should be uniform over the interval [0,1] (see Lemma 3.3.1 of [13]). The left-hand pane of Fig. 4 shows a histogram of the p-values, which clearly is not consistent with a uniform distribution. To demonstrate that it is indeed the correlated nature of the input stream that is causing the problem, we reran the entire procedure using the same 1030 images, but shuffling the order of the images at random. This shuffling removes the correlation between images that are close to one another. We now obtain the histogram in the right-hand pane, which is clearly much more consistent with a uniform distribution. This demonstrates that we can use a permutation argument to determine the false alarm rate if indeed the IID assumption is valid.

6.3. *Comparison of change point detectors*

Here we compare the two change detectors whose detection streams are based on the two-sample test statistics defined in Section 6.1 (again using $N_C = 4$ and $N_R = 16$). As discussed in Section 3, we declare that a change detector has successfully identified a boat event if the detection stream exceeds the alarm threshold at least once during a tolerance window of width N_W. Here we set N_W equal to $N_C = 4$, but other choices could be enter-

Please cite this article as: A.Y. Kim, et al., Using labeled data to evaluate change detectors in a multivariate streaming environment, Signal Process. (2009), doi:10.1016/j.sigpro.2009.04.011

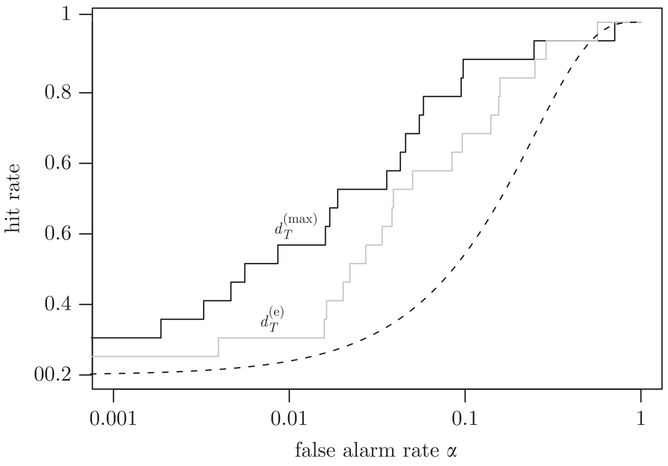

Fig. 5. ROC curves for detection streams $d_T^{(\max)}$ and $d_T^{(e)}$, along with a dashed curve appropriate for a null algorithm that ignores the data and raises an alarm at each time t with probability α.

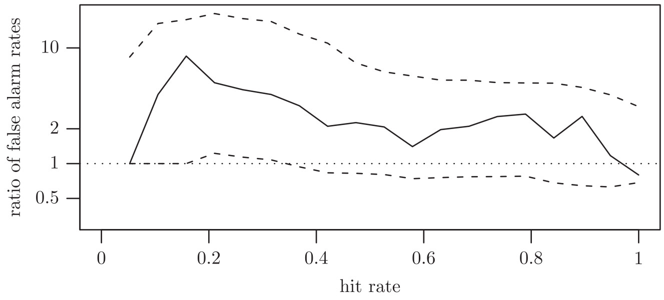

Fig. 6. Comparing ROC curves using $r_{\max,e}(h)$ versus hit rate h (solid curve). The dashed curves indicate non-simultaneous 90% empirical confidence intervals based upon 5000 bootstrap samples.

tained (i.e., there is no compelling reason to couple N_W and N_C).

Fig. 5 shows the ROC curves for the change detectors based on $d_T^{(\max)}$ and $d_T^{(e)}$ along with a curve appropriate for null detector based upon $N_W = 4$ coin tosses (dashed). Except at the very highest hit and false alarm rates (upper right-hand corner), the $d_T^{(\max)}$ detector (sensitive to mean changes) generally outperforms the $d_T^{(e)}$ detector (sensitive to arbitrary changes) in the sense of having a smaller false alarm rate for a given hit rate. To assess whether this difference between the detectors is statistically significant, we use the bootstrap procedures discussed in Section 5 to determine a 90% (non-simultaneous) confidence band for the ratio $r_{\max,e}(h)$ defined in Eq. (1) with $\varepsilon = 0.001$. The uncoupled and coupled bootstrap procedures yield basically the same results, so we only present results from the uncoupled scheme. Fig. 6 shows the confidence bands based upon 5000 uncoupled bootstrap samples. Except for a limited range of hit rates around 0.2–0.3, the intervals for $r_{\max,e}(h)$ include 1, indicating that for most hit rates the difference between the two detectors is not significant. A possible explanation for this inconclusive result is the small number of events in our training data.

7. Summary and discussion

We have proposed a method for comparing two change detectors. The method is based on labeled data, i.e., a segment of the input stream in which we have identified events and quiescent periods. The key element is an adaptation of the block bootstrap. The adaptation constructs bootstrap streams by piecing together events and quiescent periods randomly chosen (with replacement) from those making up the original stream. The bootstrap allows us to assess the effect of sampling variability on pairwise comparisons of ROC curves, and thereby determine whether a particular change detector is significantly better than another. Our example compared two change detectors whose detection streams are constructed using two-sample tests, but our method is not dependent upon this particular construction and can be applied to other kinds of change detectors (e.g., the output of cumulative sum statistics [21], which are not based explicitly on the two-sample notion).

Our proposed method can be extended to compare the performance of $K > 2$ change detectors that might arise in many different ways (e.g., different sizes for the current and references windows [10] or a time-varying geometry in which the reference window grows monotonically in time while maintaining a constant size for the current window). A problem with comparing K detectors is that none might emerge as uniformly best for all hit rates. Ignoring the question of statistical significance, Fig. 5 shows that the $d_T^{(\max)}$ detector generally outperforms the $d_T^{(e)}$ detector, but not at the highest hit/false alarm rates. We could focus on a single hit rate and then order the K detectors by their false alarm rates. The natural generalization of the matched pairs design for comparing the false alarm rates of two detectors is a blocked design where each bootstrap sample is a block, and the detectors are the treatments. Detectors can then be compared using standard multiple comparison procedures.

If we have labeled data, we can do more than just evaluate the performance of predefined change detectors—we can use the data to design new detectors. One possibility is to look for linear or nonlinear combinations of existing change detectors that outperform any single detector. For example, suppose that a change of interest is associated with a change in both the mean and the variance of a distribution, and suppose that we have two change detectors, one of which has power against changes in means, and the other, against changes in variance. Then some combination of these two detectors is likely to be superior to either individual detector in picking up on the change of interest, and the particular combination that is best can be determined using the labeled data. Using labeled data to construct new "targeted" change detectors is an interesting area for future research.

Acknowledgments

The authors thank Kevin Williams, Russ Light and Timothy Wen of the SORFED project for supporting the use of their data. This work was funded by the US Office of Naval Research under Grant number N00014–05–1–0843.

References

[1] T.W. Anderson, Sampling permutations for nonparametric methods, in: B. Ranneby (Ed.), Statistics in Theory and Practice: Essays in Honour of Bertil Matérn, Swedish University of Agricultural Sciences, Umeå, Sweden, 1982, pp. 43–52.

[2] B. Aslan, G. Zech, New test for the multivariate two-sample problem based on the concept of minimum energy, Journal of Statistical Computation and Simulation 75 (2005) 109–119.

[3] P. Bélisle, L. Joseph, B. MacGibbon, D.B. Wolfson, R. du Berger, Change-point analysis of neuron spike train data, Biometrics 54 (1998) 113–123.

[4] P. Bühlmann, Bootstraps for time series, Statistical Science 17 (2002) 52–72.

[5] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, 1993.

[6] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters 27 (2006) 861–874.

[7] M. Frisén, Statistical surveillance. Optimality and methods, International Statistical Review 71 (2003) 403–434.

[8] M. Grimm, K. Kroschel (Eds.), Robust Speech Recognition and Understanding, I-Tech Education and Publishing, Vienna, Austria, 2007.

[9] T. Hesterberg, D.S. Moore, S. Monaghan, A. Clipson, R. Epstein, Bootstrap Methods and Permutation Tests, second ed., W.H. Freeman, New York, 2005;
D.S. Moore, G.P. McCabe, Introduction to the Practice of Statistics, fifth ed., W.H. Freeman, New York, 2005 ⟨http://bcs.whfreeman.com/ips5e/content/cat_080/pdf/moore14.pdf⟩ (Chapter 14 in online version).

[10] D. Kifer, S. Ben-David, J. Gehrke, Detecting change in data streams, in: Proceedings of the 30th Very Large Data Base (VLDB) Conference, Toronto, Canada, 2004, pp. 180–191.

[11] S.N. Lahiri, Resampling Methods for Dependent Data, Springer, New York, 2003.

[12] T.L. Lai, Sequential changepoint detection in quality control and dynamical systems, Journal of the Royal Statistical Society, Series B (Methodological) 57 (1995) 613–658.

[13] E.L. Lehmann, J.P. Romano, Testing Statistical Hypotheses, third ed., Springer, New York, 2005.

[14] F. Li, G.C. Runger, E. Tuv, Supervised learning for change-point detection, International Journal of Production Research 44 (2006) 2853–2868.

[15] S.A. Macskassy, F. Provost, Confidence bands for ROC curves: methods and an empirical study, in: First Workshop on ROC Analysis in AI, ECAI-2004, Spain, 2004.

[16] M. Markou, S. Singh, Novelty detection: a review—part 1: statistical approaches, Signal Processing 83 (2003) 2481–2497.

[17] M. Markou, S. Singh, Novelty detection: a review—part 2: neural network based approaches, Signal Processing 83 (2003) 2499–2521.

[18] A. Pievatolo, R. Rotondi, Analysing the interevent time distribution to identify seismicity phases: a Bayesian nonparametric approach to the multiple-changepoint problem, Applied Statistics 49 (2000) 543–562.

[19] D.N. Politis, J.P. Romano, M. Wolf, Subsampling, Springer, New York, 1999.

[20] R. Rubinfeld, R. Servedio, Testing monotone high-dimensional distributions, in: Proceedings of the 37th Annual Symposium on Theory of Computing (STOC), 2005, pp. 147–156.

[21] G.C. Runger, M.C. Testik, Multivariate extensions to cumulative sum control charts, Quality and Reliability Engineering International 20 (2004) 587–606.

[22] G.J. Székely, Potential and kinetic energy in statistics, in: Lecture Notes, Budapest Institute of Technology, Technical University, 1989.

[23] G.J. Székely, E-statistics: energy of statistical samples, Technical Report No. 03-05, Bowling Green State University, Department of Mathematics and Statistics, 2000.

[24] G.J. Székely, M.L. Rizzo, Testing for equal distributions in high dimension, InterStat, 2004.