DISTANCE-BASED CHANGE-POINT DETECTION WITH ENTROPY ESTIMATION

Noboru Murata[1,2], *Kensuke Koshijima*[1,3] *and Hideitsu Hino*[4]

[1]Waseda University, 169-8555 Tokyo, JAPAN,
[2]noboru.murata@eb.waseda.ac.jp, [3]kkoshijima113@gmail.com
[4] University of Tsukuba, 305-8577 Ibaraki, JAPAN,
hinohide@cs.tsukuba.ac.jp

ABSTRACT

Change-point detection is a problem of finding time points where stochastic properties of time series suddenly change. This problem has been actively discussed under various contexts in the area of data mining. We introduce a non-parametric method for change-point detection by an entropy estimator. Since our entropy estimator is based on distances between data points, it is applied not only for ordinal vectorial data but also variable length data and non vectorial data. We will demonstrate its validity through numerical experiments with real-world data.

1. INTRODUCTION

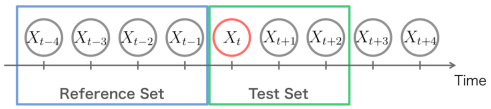
Finding change-points, at which the generating mechanism of time series changes, is of great importance in many applications for real-world data analysis. It is applied to, for example, intrusion detection in computer networks, irregular-motion detection in vision surveillance systems, signal segmentation in data stream, fraud detection in cellular systems. For further details, see [1, 2] and references therein.

The problem is generally formulated as follows. Let X_t be an observation at time t, which is a random variable of a certain stochastic process. Usually fixed-length data vectors are considered as observations. Then our objective is to examine whether the test set (future sequence) $\{X_t, X_{t+1}, \dots\}$ is different from the reference set (past sequence) $\{X_{t-1}, X_{t-2}, \dots\}$ as shown in Figure 1(a). In other words, the objective is to examine whether $X_t, X_{t+1}, \dots$ are well predicted based on $X_{t-1}, X_{t-2}, \dots$. A typical approach for solving this problem is to define a change-point score, e.g.

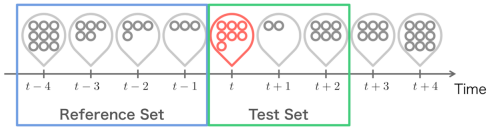
$$\begin{aligned}&\mathrm{score}(X_t)\\ &\quad = -\log \Pr(X_t, X_{t+1}, \dots | X_{t-1}, X_{t-2}, \dots), \quad (1)\end{aligned}$$

where $\Pr(\cdot|\cdot)$ is the conditional probability distribution estimated from the dataset. So far, many algorithms have been proposed from different viewpoints of generative models of time series, computational costs, scalability of data size, and sensitivity. Representative algorithms are, for example, Singular Spectrum Analysis [3], ChangeFinder [1], Kullback-Leibler Importance Estimation Procedure [2].

In this paper, we focus on a slightly different problem as shown in Figure 1(b). In our setting, an observation at time t is not a single random variable, but a collection of random variables, which we call *"a bag of data"*. We assume that observations are generated from certain probability distributions, and we want to find changes of those distributions behind the stream of observed bags. This concept is schematically shown in Figure 2 where each bag is represented by a histogram.



(a) time series



(a) stream of bags of data

Figure 1. Change-point detection problem

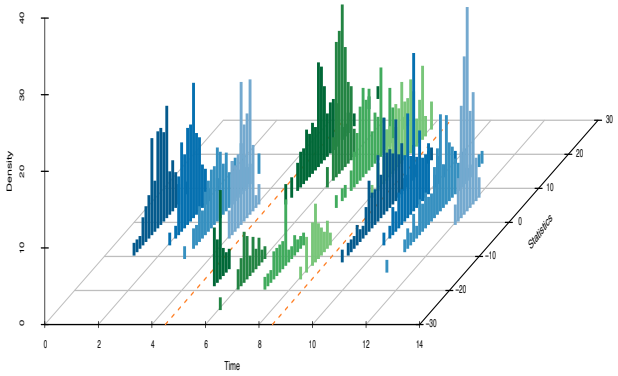


Figure 2. Sequence of histograms of bags

Interesting examples are found in graph-structured data analysis, e.g. community structure analysis, internet incident detection, market trading analysis. In these examples, graph-related statistics such as numbers of connected nodes and volumes of flow on edges are observed

at each time, and based on those observations we investigate changes of graph structures. We will analyze the ENRON email dataset [4] as an example of community structure analysis based on mail transaction.

2. METHOD

Let $B_t = \{X_i; i = 1, \ldots, n_t\}$ be an observed bag of data at time t. We assume that the size of the bag n_t can be different in time. A datum X in a bag is a random variable from a certain sample space $\mathcal{X}$, typically d-dimensional Euclidean space $\mathbb{R}^d$. In the case of detecting graph structure change under sender-receiver scenarios, X can be any value calculated from graph structure, such as the degree of a sender or a receiver, and the number of messages from a sender to a receiver.

Our objective is to examine whether the future bags $B_t, B_{t+1}, \ldots$ differ from the past bags $B_{t-1}, B_{t-2}, \ldots$. By introducing the probability distribution of the bag P_{B_t}, the bag B_t is identified by the distribution P_{B_t}, then the objective is rephrased as examining whether the sequence of distributions $P_{B_t}, P_{B_{t+1}}, \ldots$ are different from the sequence of $P_{B_{t-1}}, P_{B_{t-2}}, \ldots$.

To construct a probability distribution from a bag, there are two basic approaches: one is utilizing parametric models, and the other is utilizing non-parametric models such as a histogram and a kernel density estimate. When the probability distribution of B_t is represented by a parametric model P_{θ_t}, the problem is reduced to the ordinal change-point detection problem of the sequence of vectors $\{\theta_t\}$. However, it is sometimes difficult to choose a proper model which is valid for the target data all the time. On the other hand, non-parametric models are more flexible for modeling distributions. Here we propose the following two-step procedure for dealing with non-parametric models: first, we embed the distributions of bags $\{P_{B_t}\}$ in an appropriate metric space with a distance measure between histograms (see Figure 3), then we examine whether the fluctuation of $\{P_{B_t}\}$ is anomalous or not (see Figure 4) based on the distance-based entropy estimator, which we will explain below.

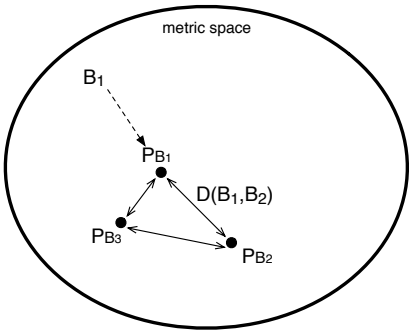
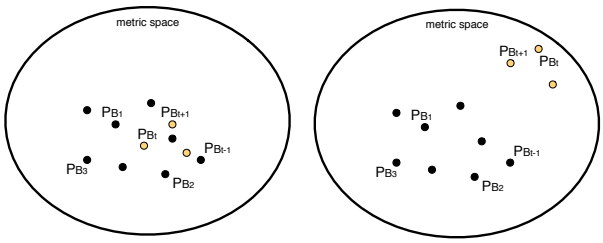


Figure 3. Embedding in a metric space



regular case anomalous case

Figure 4. Change-point detection

For a distance measure of histograms, we employ Earth Mover's Distance (EMD) [5]. It is proposed as a perceptually natural dissimilarity measure in computer vision, and is efficiently calculated by linear programming. EMD is known to be mathematically equivalent to the Wasserstein distance defined as

$$D(P, Q) = \inf_R E_R[d(X, Y)],$$
$$\text{subj. to } P(X) = \int dR(X, y),\ Q(Y) = \int dR(x, Y), \tag{2}$$

where P and Q are probability distributions on $\mathcal{X}$, and d can be any distance of $\mathcal{X}$. Intuitively speaking, it corresponds to the least amount of work needed to match two distributions, i.e. a kind of edit distance (size of shaded area $D(P, Q)$ in Figure 5).

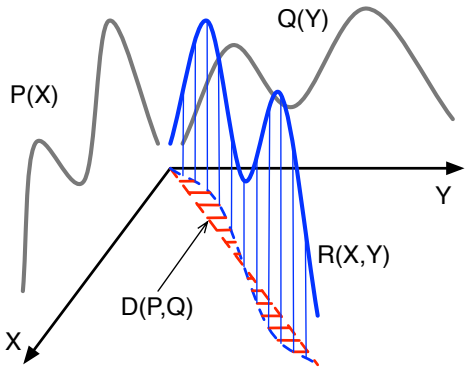


Figure 5. Earth Mover's distance

To measure the difference between the sequence of past bags $\{P_{B_{t-1}}, P_{B_{t-2}}, \ldots\}$ and the sequence of future bags $\{P_{B_t}, P_{B_{t+1}}, \ldots\}$, we adopt distance-based entropy estimators [6]. Let $\mathfrak{D} = \{(B_i, w_i); i = 1, \ldots, n\}$ be a weighted dataset, where weights satisfy $w_i > 0$ and $\sum_i w_i = 1$. For a weighted dataset $\mathfrak{D}$, the empirical average of $h(B)$ for a function h is defined as

$$\overline{h(B)} = \sum_{i=1}^{n} w_i h(B_i). \tag{3}$$

This is a simple extension of the ordinal empirical average, which is defined as

$$\overline{h(B)} = \frac{1}{n} \sum_{i=1}^{n} h(B_i), \tag{4}$$

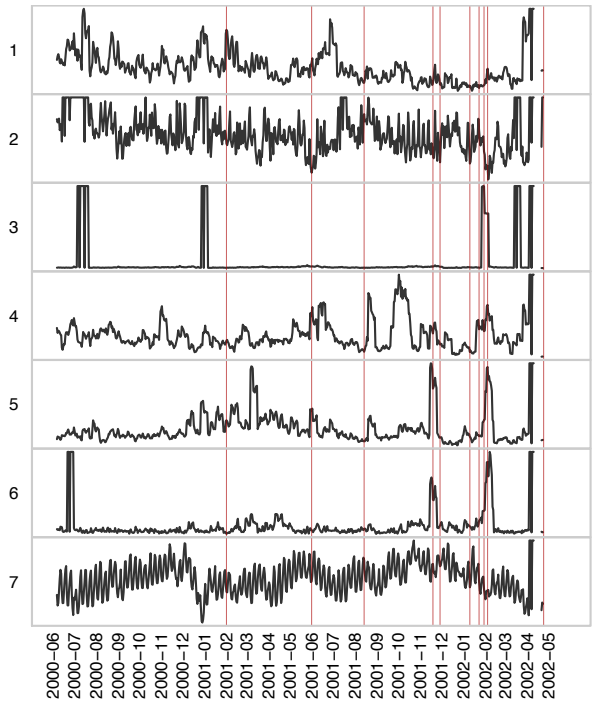


Figure 6. Change-point score (KL-based)

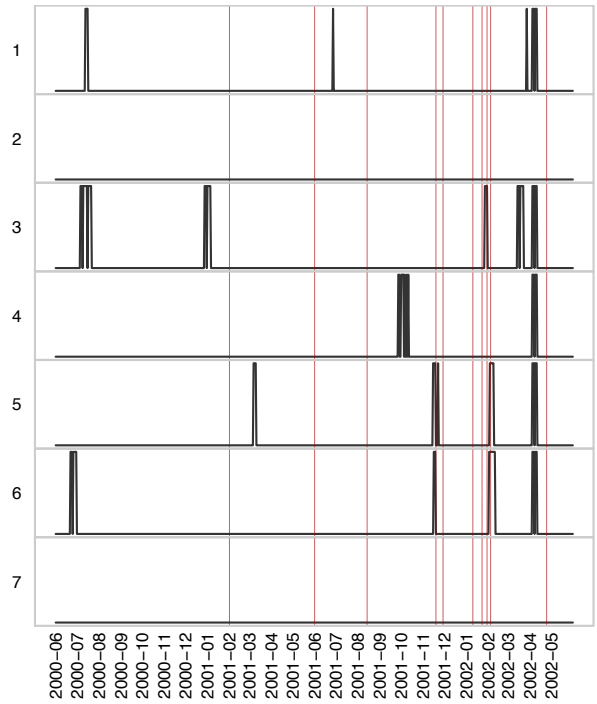


Figure 7. Alert signals

where $w_i = 1/n$. In time series analysis, weights are introduced to deal with discounting in order to reduce long-term effects of time series. In [6], three types of computationally efficient estimators (mean quantile estimators) are proposed by using a distance $D(B, B')$:

- information content of B:

$$I(B; \mathfrak{D}) = c + d \sum_{B_i \in \mathfrak{D}} w_i \log D(B_i, B) \qquad (5)$$

- cross-entropy between $\mathfrak{D}$ and $\mathfrak{D}'$:

$$H(\mathfrak{D}, \mathfrak{D}') = c + d \sum_{\substack{B_i \in \mathfrak{D} \\ B'_j \in \mathfrak{D}'}} w_i w'_j \log D(B_i, B'_j) \quad (6)$$

- auto-entropy of $\mathfrak{D}$:

$$H(\mathfrak{D}) = c + d \sum_{\substack{B_i, B_j \in \mathfrak{D} \\ B_j \neq B_i}} \frac{w_i w_j}{1 - w_i} \log D(B_i, B_j) \quad (7)$$

where c and d are constant.

Based on these estimators with Earth Mover's distance $D(B, B') = D(P_B, P_{B'})$, we propose two change-point scores as follows. First, we define reference (past) and test (future) datasets:

$$\mathfrak{D}_t^{\mathrm{ref}} = \{(B_i, w_i); i = t, t-1, \dots\} \qquad (8)$$

$$\mathfrak{D}_t^{\mathrm{test}} = \{(B_i, w_i); i = t, t+1, \dots\} \qquad (9)$$

where weights reflect discounting, e.g. $w_i \propto 1/|i - t|$. Then, we define two different change-point scores

- log density ratio:

$$\begin{aligned} \mathrm{score}(B_t) &= \log \frac{f_{\mathrm{test}}(B_t)}{f_{\mathrm{ref}}(B_t)} \\ &= I(B_t; \mathfrak{D}_t^{\mathrm{ref}}) - I(B_t; \mathfrak{D}_t^{\mathrm{test}}) \end{aligned} \qquad (10)$$

- symmetrized Kullback-Leibler divergence:

$$\begin{aligned} &\mathrm{score}(B_t) \\ &= \frac{2H(\mathfrak{D}_t^{\mathrm{ref}}, \mathfrak{D}_t^{\mathrm{test}}) - H(\mathfrak{D}_t^{\mathrm{ref}}) - H(\mathfrak{D}_t^{\mathrm{test}})}{2} \end{aligned} \qquad (11)$$

These two scores have different statistical properties. Roughly speaking, the KL-based score is usually more conservative than the density-ratio-based score, because the KL-based score uses all the pairs of $\mathfrak{D}_t^{\mathrm{ref}}$ and $\mathfrak{D}_t^{\mathrm{test}}$ but the density-ratio-based score uses a part of them.

3. ILLUSTRATIVE EXAMPLE

The ENRON email dataset [4] is a collection of email transmission data from about 150 users, mostly senior management of ENRON, around ENRON accounting scandal at 2001. We used the data between June 2000 and May 2002. The time window size of a bag is one day, and the sizes of reference and test datasets are six bags and two bags, respectively. We considered seven different statistics of graph structure:

1. the degree of a sender (how many receivers are connected with a sender),
2. the degree of a receiver,

3. the 2nd order degree of a sender (how many senders are connected with a sender through receivers),
4. the 2nd order degree of a receiver,
5. the number of mails from a sender to any receivers,
6. the number of mails from any senders to a receiver,
7. the number of mails between a pair of a sender and a receiver.

Figure 6 shows KL-based change-point scores calculated from histograms of the seven statistics. The vertical solid lines indicate important events related to the ENRON company, such as "Jeffrey Skilling takes over as CEO (Feb 2001)", "ENRON restates 3rd quarter earnings (19 Nov 2001)" and "FBI begins investigation of document shredding (23 Jan 2002)". For more detailed information, see [7].

Figure 7 shows alert signals where change-point scores exceed their means by one standard deviation. We see that most of signals correspond with the above mentioned events.

4. CONCLUSION

We have formulated a change-point detection problem of stream of bags, and proposed a method based on a statistically appropriate distance between bags of data and distance-based entropy estimators. In the proposed method, we can use various statistics extracted from bags, but the choice of statistics is crucial for the performance. We have to carefully choose the statistics based on prior knowledge of the target data or performance assessment such as cross-validation, for example. Also integrating multiple statistics by convex combination and majority rule can be possible. As a future work, we plan to extend our method to on-line detection with stable entropy estimators and on-line adaptive thresholding in order to deal with huge amount of data.

5. REFERENCES

[1] J. Takeuchi and K. Yamanishi, "A unifying framework for detecting outliers and change points from time series," *IEEE Trans. Knowledge and Data Engineering*, vol. 18, no. 4, pp. 482–492, Apr. 2006.

[2] Y. Kawahara and M. Sugiyama, "Change-point detection in time-series data by direct density-ratio estimation," in *SDM*, 2009, vol. 9, pp. 389–400.

[3] V. Moskvina and A. Zhigljavsky, "An algorithm based on singular spectrum analysis for change-point detection," *Communications in Statistics - Simulation and Computation*, vol. 32, no. 2, pp. 319–352, 2003.

[4] W. W. Cohen, *Enron Email Dataset*, Aug. 2009, `https://www.cs.cmu.edu/~enron/`.

[5] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, Nov. 2000.

[6] H. Hino and N. Murata, "Information estimators for weighted observations," *Neural Networks*, vol. 46, pp. 260–275, Oct. 2013.

[7] J. Sun, C. Faloutsos, S. Papadimitriou, and P. S. Yu, "Graphscope: parameter-free mining of large time-evolving graphs," in *Proc. KDD '07*, 2007, pp. 687–696.