Change-Point Detection in Time-Series Data by Relative Density-Ratio Estimation

Song Liu[1], Makoto Yamada[2], Nigel Collier[3], and Masashi Sugiyama[1]

[1] Tokyo Institute of Technology
2-12-1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan
{song@sg.,sugi@}cs.titech.ac.jp
[2] NTT Communication Science Laboratories
2-4, Hikaridai, Seika-cho, Kyoto, Japan 619-0237
yamada.makoto@lab.ntt.co.jp
[3] National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
collier@nii.ac.jp

Abstract. The objective of change-point detection is to discover abrupt property changes lying behind time-series data. In this paper, we present a novel statistical change-point detection algorithm that is based on non-parametric divergence estimation between two retrospective segments. Our method uses the relative Pearson divergence as a divergence measure, and it is accurately and efficiently estimated by a method of direct density-ratio estimation. Through experiments on real-world human-activity sensing, speech, and Twitter datasets, we demonstrate the usefulness of the proposed method.

Keywords: change-point detection, distribution comparison, relative density-ratio estimation, kernel methods, time-series data.

1 Introduction

Detecting abrupt changes in time-series data, called *change-point detection*, has attracted researchers in the statistics and data mining communities for decades [1–6].

Some pioneer works demonstrated good change-point detection performance by comparing the probability distributions of time-series samples over past and present intervals [1]. As both the intervals move forward, a typical strategy is to issue an alarm for a change point when the two distributions are becoming significantly different. Various change-point detection methods follow this strategy, for example, the *cumulative sum* [1], the *generalized likelihood-ratio method* [2], and the *change finder* [3].

Another group of methods that have attracted high popularity in recent years is the *subspace methods* [4, 5]. By using a pre-designed time-series model, a subspace is discovered by principle component analysis from trajectories in past

G.L. Gimel' farb et al. (Eds.): SSPR & SPR 2012, LNCS 7626, pp. 363–372, 2012.

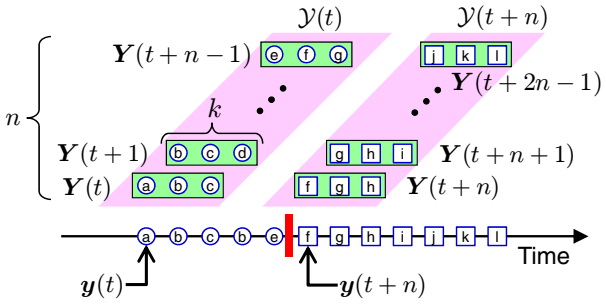


Fig. 1. Notation

and present intervals, and their dissimilarity is measured by the distance between the subspaces. One of the major approaches is called *subspace identification* [5], which compares the subspaces spanned by the columns of an *extended observability matrix* generated by a state-space model with system noise.

However, the methods explained above rely on pre-designed parametric models such as underlying probability distributions [1, 2], auto-regressive models [3], and state-space models [4, 5], for tracking some specific statistics such as the mean, the variance, and the spectrum. Thus, they are not robust against different types of changes, which significantly limits the range of applications in practice. To cope with this problem, non-parametric estimation methods such as *kernel density estimation* may be used. However, non-parametric methods tend to be less accurate in high-dimensional problems because of the so-called *curse of dimensionality*.

To overcome this difficulty, a new strategy was introduced recently which estimates the *ratio* of probability densities directly without going through density estimation [7]. In the context of change-point detection, a direct density-ratio estimation method called the *Kullback-Leibler importance estimation procedure* (KLIEP) [8] was reported to outperform competitive approaches [6] such as the *one-class support vector machine* [9] and *singular-spectrum analysis* [4].

The goal of this paper is to further advance this line of research. More specifically, our contributions in this paper are two folds.

- We apply a recently-proposed density-ratio estimation method called the *unconstrained least-squares importance fitting* (uLSIF) [10] to change-point detection. Notable advantages of uLSIF are that an analytical solution can be obtained, it achieves the optimal non-parametric convergence rate, it has optimal numerical stability, and it has higher robustness [7].
- We further improve the uLSIF-based change-point detection method by employing a state-of-the-art extension of uLSIF called *relative uLSIF* (RuLSIF) [11], which was proved to have an even better non-parametric convergence property than plain uLSIF [11], with other advantages of uLSIF maintained.

2 Problem Formulation

In this section, we formulate our change-point detection problem (see Figure 1).

Let $\boldsymbol{y}(t) \in \mathbb{R}^d$ be a d-dimensional time-series sample at time t. Let

$$\boldsymbol{Y}(t) := [\boldsymbol{y}(t)^\top, \boldsymbol{y}(t+1)^\top, \ldots, \boldsymbol{y}(t+k-1)^\top]^\top \in \mathbb{R}^{dk}$$

be a subsequence of time series at time t with length k, where $^\top$ represents the transpose. Following the previous work [6], we treat the subsequence $\boldsymbol{Y}(t)$ as a sample, instead of a single point $\boldsymbol{y}(t)$, by which time-dependent information can be incorporated naturally. Let $\mathcal{Y}(t)$ be a set of n retrospective subsequence samples starting at time t:

$$\mathcal{Y}(t) := \{\boldsymbol{Y}(t), \boldsymbol{Y}(t+1), \ldots, \boldsymbol{Y}(t+n-1)\}.$$

For change-point detection, let us consider two consecutive segments $\mathcal{Y}(t)$ and $\mathcal{Y}(t+n)$. Our strategy is to compute a certain dissimilarity measure between $\mathcal{Y}(t)$ and $\mathcal{Y}(t+n)$, and use it as the plausibility of change points. More specifically, the larger the dissimilarity is, the more likely the point is a change point.

Now the problems that need to be addressed are what kind of dissimilarity measure we should use and how we estimate it from data. We will discuss these issues in the next section.

3 Change-Point Detection via Density-Ratio Estimation

In this section, we first define our dissimilarity measure, and then show methods for estimating the dissimilarity measure.

3.1 Divergence-Based Dissimilarity Measure

We use a dissimilarity measure of the following form:

$$D(P_t \| P_{t+n}) + D(P_{t+n} \| P_t), \tag{1}$$

where P_t and P_{t+n} are probability distributions of samples in $\mathcal{Y}(t)$ and $\mathcal{Y}(t+n)$, respectively. $D(P\|P')$ denotes the *f-divergence* [12, 13]:

$$D(P\|P') := \int p'(\boldsymbol{Y}) f\left(\frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})}\right) \mathrm{d}\boldsymbol{Y},$$

where f is a convex function such that $f(1) = 0$, and $p(\boldsymbol{Y})$ and $p'(\boldsymbol{Y})$ are probability density functions of P and P', respectively. Because the f-divergence is not symmetric, we use a symmetrized divergence in Eq.(1).

The f-divergence includes various popular divergences such as the *Kullback-Leibler (KL) divergence* by $f(t) = t \log t$ and the *Pearson (PE) divergence* by $f(t) = \frac{1}{2}(t-1)^2$:

$$\mathrm{KL}(P\|P') := \int p(\boldsymbol{Y}) \log \frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})} \mathrm{d}\boldsymbol{Y} \quad \text{and} \quad \mathrm{PE}(P\|P') := \frac{1}{2} \int p'(\boldsymbol{Y}) \left(\frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})} - 1\right)^2 \mathrm{d}\boldsymbol{Y}.$$

In the rest of this section, we explain three methods of directly estimating the density ratio $\frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})}$ from samples $\{\boldsymbol{Y}_i\}_{i=1}^n$ and $\{\boldsymbol{Y}'_j\}_{j=1}^n$ drawn from $p(\boldsymbol{Y})$ and $p'(\boldsymbol{Y})$: the *KL importance estimation procedure* (KLIEP) [8] in Section 3.2, *unconstrained least-squares importance fitting* (uLSIF) [10] in Section 3.3, and *relative uLSIF (RuLSIF)* [11] in Section 3.4.

3.2 Kullback-Leibler Importance Estimation Procedure (KLIEP)

KLIEP [8] is a direct density-ratio estimation algorithm that is suitable for estimating the KL divergence.

Density-Ratio Model: Let us model the density ratio $\frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})}$ by the following kernel model:

$$g(\boldsymbol{Y};\boldsymbol{\theta}) := \sum_{\ell=1}^{n} \theta_\ell K(\boldsymbol{Y}, \boldsymbol{Y}_\ell), \tag{2}$$

where $\boldsymbol{\theta} := (\theta_1, \ldots, \theta_n)^\top$ are parameters to be learned from data samples, and $K(\boldsymbol{Y}, \boldsymbol{Y}')$ is a kernel basis function. In practice, we use the Gaussian kernel and the kernel width is chosen by cross-validation (see [8] for details).

Learning Algorithm: The parameters $\boldsymbol{\theta}$ in the model $g(\boldsymbol{Y};\boldsymbol{\theta})$ are determined so that the empirical KL divergence from $p(\boldsymbol{Y})$ to $g(\boldsymbol{Y};\boldsymbol{\theta})p'(\boldsymbol{Y})$ is minimized:

$$\max_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n} \log\left(\sum_{\ell=1}^{n} \theta_\ell K(\boldsymbol{Y}_i, \boldsymbol{Y}_\ell)\right) \text{ s.t. } \frac{1}{n}\sum_{j=1}^{n}\sum_{\ell=1}^{n} \theta_\ell K(\boldsymbol{Y}'_j, \boldsymbol{Y}_\ell) = 1, \quad \theta_1, \ldots, \theta_n \geq 0.$$

The equality constraint is for normalization purposes because $g(\boldsymbol{Y};\boldsymbol{\theta})p'(\boldsymbol{Y})$ should be a probability density function. The inequality constraint comes from the non-negativity of the density-ratio function. Since this is a convex optimization problem, the unique global optimal solution $\widehat{\boldsymbol{\theta}}$ can be simply obtained, for example, by a gradient-projection iteration. Finally, a density-ratio estimator is given as

$$\widehat{g}(\boldsymbol{Y}) = \sum_{\ell=1}^{n} \widehat{\theta}_\ell K(\boldsymbol{Y}, \boldsymbol{Y}_\ell). \tag{3}$$

KLIEP was shown to achieve the optimal non-parametric convergence rate [8].

Change-Point Detection by KLIEP: Given a density-ratio estimator $\widehat{g}(\boldsymbol{Y})$, an approximator of the KL divergence is given as

$$\widehat{\mathrm{KL}} := \frac{1}{n}\sum_{i=1}^{n} \log \widehat{g}(\boldsymbol{Y}_i).$$

In the previous work [6], this KLIEP-based KL-divergence estimator was applied to change-point detection and demonstrated to be promising in experiments.

3.3 Unconstrained Least-Squares Importance Fitting (uLSIF)

Recently, another direct density-ratio estimator called uLSIF was proposed [10], which is suitable for estimating the PE divergence.

Learning Algorithm: In uLSIF, the same density-ratio model $g(\boldsymbol{Y};\boldsymbol{\theta})$ as KLIEP (see Eq.(2)) is used. However, its training criterion is different; the density-ratio model is fitted to the true density ratio under the squared loss. More specifically, the parameter $\boldsymbol{\theta}$ in the model $g(\boldsymbol{Y};\boldsymbol{\theta})$ is determined so that the following squared loss $J(\boldsymbol{Y})$ is minimized:

$$\begin{aligned} J(\boldsymbol{Y}) &:= \frac{1}{2}\int \left(\frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})} - g(\boldsymbol{Y};\boldsymbol{\theta})\right)^2 p'(\boldsymbol{Y})\,\mathrm{d}\boldsymbol{Y} \\ &= \frac{1}{2}\int \frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})}^2 p'(\boldsymbol{Y})\,\mathrm{d}\boldsymbol{Y} - \int p(\boldsymbol{Y}) g(\boldsymbol{Y};\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{Y} + \frac{1}{2}\int g(\boldsymbol{Y};\boldsymbol{\theta})^2 p'(\boldsymbol{Y})\,\mathrm{d}\boldsymbol{Y}. \end{aligned}$$

Since the first term is a constant, we focus on the last two terms. By approximating the expectations by the empirical averages, the uLSIF optimization problem is given as follows:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^n}\left[\frac{1}{2}\boldsymbol{\theta}^\top \widehat{\boldsymbol{H}}\boldsymbol{\theta} - \widehat{\boldsymbol{h}}^\top\boldsymbol{\theta} + \frac{\lambda}{2}\boldsymbol{\theta}^\top\boldsymbol{\theta}\right], \tag{4}$$

where the penalty term $\frac{\lambda}{2}\boldsymbol{\theta}^\top\boldsymbol{\theta}$ is included for regularization purposes, and λ (≥ 0) denotes the regularization parameter, which is chosen by cross validation. (see [10] for details). $\widehat{\boldsymbol{H}}$ is the $n \times n$ matrix and $\widehat{\boldsymbol{h}}$ is the n-dimensional vector defined as

$$\widehat{H}_{\ell,\ell'} := \frac{1}{n}\sum_{j=1}^{n} K(\boldsymbol{Y}'_j, \boldsymbol{Y}_\ell) K(\boldsymbol{Y}'_j, \boldsymbol{Y}_{\ell'}) \quad \text{and} \quad \widehat{h}_\ell := \frac{1}{n}\sum_{i=1}^{n} K(\boldsymbol{Y}_i, \boldsymbol{Y}_\ell).$$

It is easy to confirm that the solution $\widehat{\boldsymbol{\theta}}$ of (4) can be analytically obtained as

$$\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}_n)^{-1}\widehat{\boldsymbol{h}}, \tag{5}$$

where $\boldsymbol{I}_n$ denotes the n-dimensional identity matrix. Finally, a density-ratio estimator is given by Eq.(3) with Eq.(5).

Change-Point Detection by uLSIF: Given a density-ratio estimator $\widehat{g}(\boldsymbol{Y})$, an approximator of the PE divergence can be constructed as

$$\widehat{\mathrm{PE}} := -\frac{1}{2n}\sum_{j=1}^{n} \widehat{g}(\boldsymbol{Y}'_j)^2 + \frac{1}{n}\sum_{i=1}^{n}\widehat{g}(\boldsymbol{Y}_i) - \frac{1}{2}.$$

This approximator is derived from the following expression of the PE divergence:

$$\mathrm{PE}(P\|P') = -\frac{1}{2}\int \left(\frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})}\right)^2 p'(\boldsymbol{Y})\mathrm{d}\boldsymbol{Y} + \int \left(\frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})}\right) p(\boldsymbol{Y})\mathrm{d}\boldsymbol{Y} - \frac{1}{2}. \tag{6}$$

Notable advantages of uLSIF are that its solution can be computed analytically, it possesses the optimal non-parametric convergence rate, it has the optimal numerical stability, and it has higher robustness [7]. As experimentally demonstrated in our supplementary technical report [14], uLSIF-based change-point detection compares favorably with the KLIEP-based method.

3.4 Relative uLSIF (RuLSIF)

Depending on the condition of the denominator density $p'(\boldsymbol{Y})$, the density-ratio value $\frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})}$ can be unbounded (i.e., they can be infinity). This is actually problematic because the non-parametric convergence rate of uLSIF is governed by the "sup"-norm of the true density-ratio function: $\max_{\boldsymbol{Y}} \frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})}$. To overcome this problem, *relative density-ratio estimation* was introduced [11].

Relative PE Divergence: Let us consider the *α-relative PE-divergence* for $0 \leq \alpha < 1$:

$$\mathrm{PE}_\alpha(P\|P') := \mathrm{PE}(P\|\alpha P + (1-\alpha)P') = \int p'_\alpha(\boldsymbol{Y}) \left(r_\alpha(\boldsymbol{Y}) - 1\right)^2 \mathrm{d}\boldsymbol{Y},$$

where $p'_\alpha(\boldsymbol{Y}) = \alpha p(\boldsymbol{Y}) + (1-\alpha)p'(\boldsymbol{Y})$ and $r_\alpha(\boldsymbol{Y}) = \frac{p(\boldsymbol{Y})}{p'_\alpha(\boldsymbol{Y})}$. We refer to $r_\alpha(\boldsymbol{Y})$ as the *α-relative density ratio*. The α-relative density ratio is reduced to the plain density ratio if $\alpha = 0$, and it tends to be "smoother" as α gets larger. Indeed, the α-relative density ratio is bounded above by $1/\alpha$ for $\alpha > 0$, even when the plain density ratio $\frac{p(\boldsymbol{Y})}{p'(\boldsymbol{Y})}$ is unbounded. This was proved to contribute to improving the estimation accuracy [11].

Learning Algorithm: In the same way as the uLSIF method, the parameter $\boldsymbol{\theta}$ of the model $g(\boldsymbol{Y};\boldsymbol{\theta})$ is learned by minimizing the squared difference between true and estimated ratios:

$$\begin{aligned} J(\boldsymbol{Y}) &= \frac{1}{2}\int p'_\alpha(\boldsymbol{Y})(r_\alpha(\boldsymbol{Y}) - g(\boldsymbol{Y};\boldsymbol{\theta}))^2 \,\mathrm{d}\boldsymbol{Y} \\ &= \frac{1}{2}\int p'_\alpha(\boldsymbol{Y}) r_\alpha^2(\boldsymbol{Y})\mathrm{d}\boldsymbol{Y} - \int p(\boldsymbol{Y}) r_\alpha(\boldsymbol{Y}) g(\boldsymbol{Y};\boldsymbol{\theta}) \,\mathrm{d}\boldsymbol{Y} \\ &\quad + \frac{\alpha}{2}\int p(\boldsymbol{Y}) g(\boldsymbol{Y};\boldsymbol{\theta})^2 \,\mathrm{d}\boldsymbol{Y} - \frac{1-\alpha}{2}\int p'(\boldsymbol{Y}) g(\boldsymbol{Y};\boldsymbol{\theta})^2 \,\mathrm{d}\boldsymbol{Y}, \end{aligned}$$

where the first term is a constant term. Note that we still use the same kernel model (2) as $g(\boldsymbol{Y};\boldsymbol{\theta})$ for approximating the α-relative density ratio.

Again, by ignoring the constant and approximating the expectations by empirical averages, the α-relative density ratio can be learned in the same way as the plain density ratio. Indeed, the optimization problem of a relative variant of uLSIF, called RuLSIF, is given as the same form as uLSIF; the only difference is the definition of the matrix $\widehat{\boldsymbol{H}}$, which is now given by

$$\widehat{H}_{\ell,\ell'} := \frac{\alpha}{n}\sum_{i=1}^{n} K(\boldsymbol{Y}_i, \boldsymbol{Y}_\ell)K(\boldsymbol{Y}_i, \boldsymbol{Y}_{\ell'}) + \frac{(1-\alpha)}{n}\sum_{j=1}^{n} K(\boldsymbol{Y}'_j, \boldsymbol{Y}_\ell)K(\boldsymbol{Y}'_j, \boldsymbol{Y}_{\ell'}).$$

RuLSIF inherits the advantages of uLSIF, i.e., its solution can be computed analytically, it has the superior numerical stability, and it has higher robustness; furthermore, RuLSIF possesses an even better non-parametric convergence property than uLSIF [11].

Change-Point Detection by RuLSIF: By using an estimator $\widehat{g}(\boldsymbol{Y})$ of the α-relative density ratio, the α-relative PE divergence can be approximated as

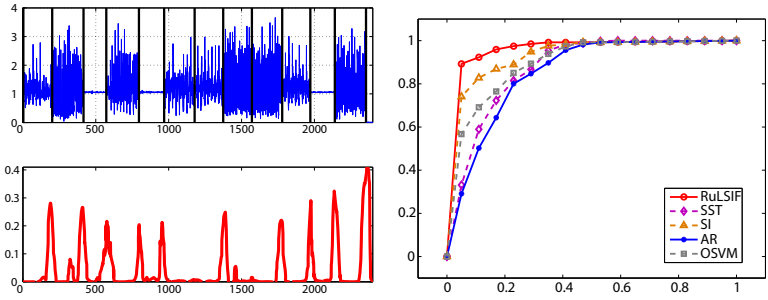
$$\widehat{\mathrm{PE}}_\alpha := -\frac{\alpha}{2n}\sum_{i=1}^{n}\widehat{g}(\boldsymbol{Y}_i)^2 - \frac{1-\alpha}{2n}\sum_{j=1}^{n}\widehat{g}(\boldsymbol{Y}'_j)^2 + \frac{1}{n}\sum_{i=1}^{n}\widehat{g}(\boldsymbol{Y}_i) - \frac{1}{2}.$$

As experimentally demonstrated in our supplementary technical report [14], the RuLSIF-based change-point detection performs even better than the plain uLSIF-based method. Thus, we focus on RuLSIF in the experiments in Section 4.

4 Experiments

In this section, we experimentally investigate the performance of the proposed and existing change-point detection methods.

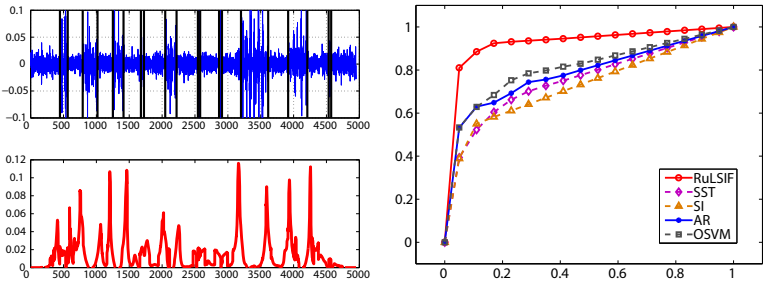
First, we use a human activity dataset and a speech dataset. The human activity dataset is a subset of the *Human Activity Sensing Consortium (HASC) challenge 2011*, which provides human activity information collected by portable three-axis accelerometers. The speech dataset is the *IPSJ SIG-SLP Corpora and Environments for Noisy Speech Recognition* (CENSREC) dataset provided by National Institute of Informatics (NII), which records human voice in a noisy



(a) One of the original time-series and RuLSIF-based change score

(b) Average ROC curves

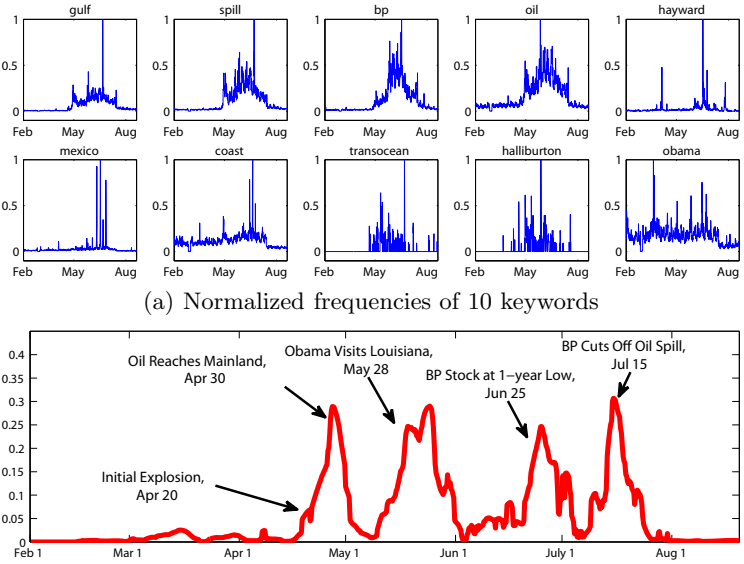
Fig. 2. HASC human-activity dataset (http://hasc.jp/hc2011/)



(a) One of the original time-series and RuLSIF-based change score

(b) Average ROC curves

Fig. 3. NII speech dataset (http://research.nii.ac.jp/src/eng/list/index.html)



(a) Normalized frequencies of 10 keywords

(b) RuLSIF-based change score and exemplary real-world events

Fig. 4. Twitter dataset (http://www.ark.cs.cmu.edu/tweets/)

environment. We compare our RuLSIF-based method with several state-of-the-art methods: *Singular spectrum transformation* (SST) [4], *subspace identification* (SI) [5], *auto regressive* (AR) [3], and *one-class support vector machine* (OSVM) [9]. Examples of RuLSIF-based change score and ROC curves over 10 datasets are plotted in Figures 2 and 3, showing that the proposed RuLSIF-based method outperforms other methods.

Finally, we apply the proposed change-point detection method to the *CMU Twitter dataset*, which is an archive of Twitter messages that have been collected from April 2010 to October 2010 via the Twitter API. Here we track the degree

of popularity of a given topic by monitoring the frequency of selected keywords. More specifically, we focus on events related to "*Deepwater Horizon oil spill in the Gulf of Mexico*" which occurred on April 20, 2010, and was widely broadcast among the Twitter community. We use the frequency of 10 keywords: "*gulf*", "*spill*", "*bp*", "*oil*", "*hayward*", "*mexico*", "*coast*", "*transocean*", "*halliburton*", and "*obama*" (see Figure 4(a)). For quantitative evaluation, we referred to the Wikipedia entry "Timeline of the Deepwater Horizon oil spill" as a real-world event source. The change-point score obtained by the proposed RuLSIF-based method is plotted in Figure 4(b), where four occurrences of important real-world events show the development of this news story.

As we can see from Figure 4(b), the change-point score increases immediately after the initial explosion of the deepwater horizon oil platform and soon reaches the first peak when oil was found on the sea shore of Louisiana on April 30. Shortly after BP announced its preliminary estimation on the amount of leaking oil, the change-point score rises quickly again and reaches its second peak at the end of May, at which time President Obama visited Louisiana to assure local residents of the federal government's support. On June 25, the BP stock was at its one year's lowest price, while the change-point score spikes at the third time. Finally, BP cuts off the spill on July 15, as the score reaches its last peak.

5 Conclusion

We extended the existing KLIEP-based change detection method and proposed to use uLSIF or RuLSIF as a building block. Through experiments, we demonstrated that the RuLSIF-based change detection method is promising.

SL was supported by NII internship fund and the JST PRESTO program. MY and MS were supported by the JST PRESTO program. NC was supported by NII Grand Challenge project fund.

References

1. Basseville, M., Nikiforov, I.V.: Detection of Abrupt Changes: Theory and Application. Prentice-Hall, Inc., Upper Saddle River (1993)
2. Gustafsson, F.: The marginalized likelihood ratio test for detecting abrupt changes. IEEE Transactions on Automatic Control 41(1), 66–78 (1996)
3. Takeuchi, Y., Yamanishi, K.: A unifying framework for detecting outliers and change points from non-stationary time series data. IEEE Transactions on Knowledge and Data Engineering 18(4), 482–489 (2006)
4. Moskvina, V., Zhigljavsky, A.: Change-point detection algorithm based on the singular-spectrum analysis. Communications in Statistics: Simulation and Computation 32, 319–352 (2003)
5. Kawahara, Y., Yairi, T., Machida, K.: Change-point detection in time-series data based on subspace identification. In: Proceedings of the 7th IEEE International Conference on Data Mining, pp. 559–564 (2007)
6. Kawahara, Y., Sugiyama, M.: Sequential change-point detection based on direct density-ratio estimation. Statistical Analysis and Data Mining 5(2), 114–127 (2012)

7. Sugiyama, M., Suzuki, T., Kanamori, T.: Density Ratio Estimation in Machine Learning. Cambridge University Press, Cambridge (2012)
8. Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Buenau, P., Kawanabe, M.: Direct importance estimation for covariate shift adaptation. Annals of the Institute of Statistical Mathematics 60(4), 699–746 (2008)
9. Desobry, F., Davy, M., Doncarli, C.: An online kernel change detection algorithm. IEEE Transactions on Signal Processing 53(8), 2961–2974 (2005)
10. Kanamori, T., Hido, S., Sugiyama, M.: A least-squares approach to direct importance estimation. Journal of Machine Learning Research 10, 1391–1445 (2009)
11. Yamada, M., Suzuki, T., Kanamori, T., Hachiya, H., Sugiyama, M.: Relative density-ratio estimation for robust distribution comparison. Advances in Neural Information Processing Systems 24, 594–602 (2011)
12. Ali, S.M., Silvey, S.D.: A general class of coefficients of divergence of one distribution from another. Journal of the Royal Statistical Society, Series B 28(1), 131–142 (1966)
13. Csiszár, I.: Information-type measures of difference of probability distributions and indirect observation. Studia Scientiarum Mathematicarum Hungarica 2, 229–318 (1967)
14. Liu, S., Yamada, M., Collier, N., Sugiyama, M.: Change-point detection in time-series data by relative density-ratio estimation. arXiv 1203.0453 (2012)