Description of Kawahara and Sugiyama's online changepoint detection algorithm based on direct density ratio estimation

Christopher Natoli

1 Framework of the solution

The following online changepoint detection algorithm was designed by Kawahara and Sugiyama [2].

Suppose $\{\boldsymbol{y}(t)\}_{t=1}^{T} \subset \mathbb{R}^d$ is a d-dimensional time series and let $\boldsymbol{Y}(t) \in \mathbb{R}^{dk}$ be a k-long subsequence of the data beginning at time t:

$$\boldsymbol{Y}(t) := \left[\boldsymbol{y}(t)^\top, \boldsymbol{y}(t+1)^\top, \ldots, \boldsymbol{y}(t+k-1)^\top\right]^\top,$$

i.e., $\boldsymbol{Y}(t)$ is the concatenation of the k vectors $\boldsymbol{y}(t), \ldots, \boldsymbol{y}(t+k-1)$. Rather than consider the "feature space" or "state space" $\mathbb{R}^d$, i.e., the space in which the datapoints live, we will consider the space $\mathbb{R}^{dk}$ in which the subequences $\boldsymbol{Y}(t)$ live, since these subsequences can describe some of the temporal structure of the time series that points in $\mathbb{R}^d$ cannot.

This online algorithm makes use of a sliding window. Divide the sliding window into two intervals, one before the possible changepoint called the reference interval and one after the possible changepoint called the test interval. Denote the beginnings of the reference and test intervals by t_{rf} and t_{te}, respectively, and denote their lengths by n_{rf} and n_{te}. If t is the current time, i.e., the front of the sliding window, then $t = t_{\text{te}} + n_{\text{te}}$ and $t_{\text{te}} = t_{\text{rf}} + n_{\text{rf}}$. Define $\boldsymbol{Y}_{\text{rf}}(i) = \boldsymbol{Y}(t_{\text{rf}}+i-1)$ and $\boldsymbol{Y}_{\text{te}}(i) = \boldsymbol{Y}(t_{\text{te}}+i-1)$. Let p_{rf} and p_{te} be the probability densities of the reference and test intervals, respectively.

We want to test

$$\begin{aligned} & H_0 \colon p(\boldsymbol{Y}(i)) = p_{\text{rf}}(Y(i)) && \text{for } i = t_{\text{rf}}, \ldots, t-1 \\ \text{vs} \quad & H_1 \colon p(\boldsymbol{Y}(i)) = p_{\text{rf}}(Y(i)) && \text{for } i = t_{\text{rf}}, \ldots, t_{\text{te}}-1 \\ & \qquad\; p(\boldsymbol{Y}(i)) = p_{\text{te}}(Y(i)) && \text{for } i = t_{\text{te}}, \ldots, t-1. \end{aligned}$$

The likelihood ratio and log-likelihood ratio for these hypotheses are therefore

$$\begin{aligned} \Lambda &= \frac{\prod_{i=1}^{n_{\text{rf}}} p_{\text{rf}}(\boldsymbol{Y}_{\text{rf}}(i)) \prod_{i=1}^{n_{\text{te}}} p_{\text{te}}(\boldsymbol{Y}_{\text{te}}(i))}{\prod_{i=1}^{n_{\text{rf}}} p_{\text{rf}}(\boldsymbol{Y}_{\text{rf}}(i)) \prod_{i=1}^{n_{\text{te}}} p_{\text{rf}}(\boldsymbol{Y}_{\text{te}}(i))} \\ &= \frac{\prod_{i=1}^{n_{\text{te}}} p_{\text{te}}(\boldsymbol{Y}_{\text{te}}(i))}{\prod_{i=1}^{n_{\text{te}}} p_{\text{rf}}(\boldsymbol{Y}_{\text{te}}(i))} \\ S := \log \Lambda &= \sum_{i=1}^{n_{\text{te}}} \log \frac{p_{\text{te}}(\boldsymbol{Y}_{\text{te}}(i))}{p_{\text{rf}}(\boldsymbol{Y}_{\text{te}}(i))}. \end{aligned}$$

We choose the log-likelihood ratio S as our score. If the score is greater than a certain threshold, then we say that tme t_{te} is a changepoint.

2 Directly estimating the density ratio

Rather than estimate the density ratio $w := \frac{p_{\text{te}}}{p_{\text{rf}}}$ by first individually estimating the densities p_{te} and p_{rf} and then dividing, the ratio itself will be directly estimated by employing the Kullback-Leibler Importance Estimation Procedure (KLIEP) developed by Sugiymama et al. [4].

Suppose the estimate $\hat{w}$ lives in a finite-dimensional function space spanned by the basis functions $\{\phi_\ell\}_{\ell=1}^b$, where $\phi_\ell \colon \mathbb{R}^{dk} \to \mathbb{R}^+$ for all ℓ. Then $\hat{w}(\cdot) = \sum_{\ell=1}^b \alpha_\ell \phi_\ell(\cdot)$, and so the parameters we want to learn are the coordinates $\{\alpha_\ell\}_{\ell=1}^b$. We do so by minimizing the Kullback-Leibler divergence of $\hat{p}_{\text{te}} = p_{\text{rf}}\hat{w}$ from the true distribution p_{te}, i.e.,

$$D_{KL}(p_{\text{te}} \| \hat{p}_{\text{te}}) = \int p_{\text{te}}(\boldsymbol{Y}) \log \frac{p_{\text{te}}(\boldsymbol{Y})}{p_{\text{rf}}(\boldsymbol{Y})\hat{w}(\boldsymbol{Y})} \, d\boldsymbol{Y}.$$

The minimization problem is subject to two constraints. First, w is nonnegative: requiring $\alpha_\ell \geq 0$ for all ℓ implies $\hat{w} \geq 0$. Second, p_{te} is a probability density and therefore integrates to 1: the same should hold for $p_{\text{rf}}\hat{w}$. Approximation reduces this constrained minimization problem, which is convex, to the following maximization problem:

$$\begin{aligned} &\max_{\{\alpha_\ell\}_{\ell=1}^b} \left\{ \sum_{i=1}^{n_{\text{te}}} \log \sum_{\ell=1}^b \alpha_\ell \phi_\ell(\boldsymbol{Y}_{\text{te}}(i)) \right\} \\ \text{subject to } &\frac{1}{n_{\text{rf}}} \sum_{i=1}^{n_{\text{rf}}} \sum_{\ell=1}^b \alpha_\ell \phi_\ell(\boldsymbol{Y}_{\text{rf}}(i)) = 1 \\ &\alpha_\ell \geq 0, \quad \ell = 1, \ldots, b. \end{aligned}$$

For this changepoint algorithm, we choose that the basis functions $\phi_\ell(\cdot)$ are Gaussian kernels $K_\sigma(\cdot, \boldsymbol{Y}_{\text{te}}(\ell))$ centered at the test subsequences, where

$$K_\sigma(\boldsymbol{Y}, \boldsymbol{Y}') = \exp\left(-\frac{\|\boldsymbol{Y} - \boldsymbol{Y}'\|_2^2}{2\sigma^2} \right).$$

It follows that $b = n_{\text{te}}$. The rationale for centering kernels on the test points is that it is effective to place kernels where the target function w is large. The ratio w is large when p_{te} is large and/or when p_{rf} is small. The density p_{te} of the test distribution is probably large when its argument is pulled from the test distribution, i.e., a test point $\boldsymbol{Y}_{\text{te}}(i)$. Kernels are therefore centered at the test points $\{\boldsymbol{Y}_{\text{te}}(i)\}_{i=1}^{n_{\text{te}}}$.

The kernel parameter σ is selected by cross-validation. The maximization problem is solved by gradient ascent.

3 Updating the parameters $\{\alpha_\ell\}$ online

Kawahara and Sugiyama propose a way to update $\{\alpha_\ell\}$ in an online manner rather than repeating the above procedure with every new datapoint.

Let $\mathcal{Y}_{\text{te}} = \{\boldsymbol{Y}_{\text{te}}(i)\}_{i=1}^{n_{\text{te}}}$ be the set of test subsequences. Since the Gaussian kernel $K_\sigma \colon \mathcal{Y}_{\text{te}} \times \mathcal{Y}_{\text{te}} \to \mathbb{R}^+$ is a positive definite kernel, we have by the Moore-Aronszajn theorem that there

exists a unique Hilbert space $\mathcal{H}$ of functions on $\mathcal{Y}_{\text{te}}$ with K_σ as the reproducing kernel [1]. Moreover, the span of $\{K_\sigma(\cdot, \boldsymbol{Y}))\}_{\boldsymbol{Y} \in \mathcal{Y}_{\text{te}}}$ is dense in $\mathcal{H}$, so if $w \in \mathcal{H}$, then $\hat{w}$ can approximate w arbitrarily well. This partially justifies our previous assumption that $\hat{w}$ lives in the span of $\{K_\sigma(\cdot, \boldsymbol{Y}))\}_{\boldsymbol{Y} \in \mathcal{Y}_{\text{te}}}$.

We choose to minimize the regularized risk functional

$$E_i(\hat{w}) = -\log \hat{w}(\boldsymbol{Y}_{\text{te}}(i)) + \frac{\lambda}{2}\|\hat{w}\|_{\mathcal{H}}^2$$

over the span of $\{K_\sigma(\cdot, \boldsymbol{Y})\}_{\boldsymbol{Y} \in \mathcal{Y}_{\text{te}}}$, where $-\log \hat{w}(\boldsymbol{Y}_{\text{te}}(i))$ is the loss function, $\frac{\lambda}{2}\|\hat{w}\|_{\mathcal{H}}^2$ is the regularization term, and $\|\cdot\|_{\mathcal{H}}$ is the norm in $\mathcal{H}$. Rather than minimize the average regularized risk $\frac{1}{n_{\text{te}}}\sum_{i=1}^{n_{\text{te}}} E_i(\hat{w})$, we follow the method of stochastic gradient descent in a function space [3] and minimize the regularized risk $E_{n_{\text{te}}+1}$ at only the new training example.

Before differentiating the regularized risk with respect to the function $\hat{w}$, we note a few useful derivatives. Let $e_{\boldsymbol{Y}}(\hat{w}) := \hat{w}(\boldsymbol{Y})$ be the evaluation functional at $\boldsymbol{Y}$. Since K_σ is a reproducing kernel, we have by definition that

$$\langle \hat{w}(\cdot), K_\sigma(\cdot, \boldsymbol{Y})\rangle = \hat{w}(\boldsymbol{Y}) = e_{\boldsymbol{Y}}(\hat{w}).$$

Therefore,

$$\partial_{\hat{w}} e_{\boldsymbol{Y}}(\hat{w}) = \partial_{\hat{w}}\langle(\hat{w}(\cdot), K_\sigma(\cdot, \boldsymbol{Y})\rangle = K_\sigma(\cdot, \boldsymbol{Y})$$

and so

$$\partial_{\hat{w}}\left(-\log \hat{w}(\boldsymbol{Y})\right) = \partial_{\hat{w}}\left(-\log e_{\boldsymbol{Y}}(\hat{w})\right) = -\frac{1}{\log e_{\boldsymbol{Y}}(\hat{w})}\partial_{\hat{w}} e_{\boldsymbol{Y}}(\hat{w}) = -\frac{K_\sigma(\cdot, \boldsymbol{Y})}{\log \hat{w}(\boldsymbol{Y})}.$$

We also note that $\partial_{\hat{w}}\left(\frac{1}{2}\|\hat{w}\|^2\right) = \hat{w}$. Putting these together, we find that

$$\partial_{\hat{w}} E_i(\hat{w}) = -\frac{K_\sigma(\cdot, \boldsymbol{Y}_{\text{te}}(i))}{\hat{w}(\boldsymbol{Y}_{\text{te}}(i))} + \lambda\hat{w}.$$

Denote the estimated density ratio after updating by $\hat{w}'(\cdot) = \sum_{i=1}^{n_{\text{te}}} \alpha'_\ell K_\sigma(\cdot, \boldsymbol{Y}_{\text{te}}(\ell+1))$, where kernels are placed on the $(\ell+1)$th test subsequence because the test interval moves one step forward after updaing. Then the update equation with learning rate η is

$$\begin{aligned}
\hat{w}' &= \hat{w} - \eta\partial_{\hat{w}} E_{n_{\text{te}}+1}(\hat{w}) \\
&= \hat{w} - \eta\left(-\frac{K_\sigma(\cdot, \boldsymbol{Y}_{\text{te}}(n_{\text{te}}+1))}{\hat{w}(\boldsymbol{Y}_{\text{te}}(n_{\text{te}}+1))} + \lambda\hat{w}\right) \\
&= (1-\eta\lambda)\hat{w} + \eta\frac{K_\sigma(\cdot, \boldsymbol{Y}_{\text{te}}(n_{\text{te}}+1))}{\hat{w}(\boldsymbol{Y}_{\text{te}}(n_{\text{te}}+1))} \\
\sum_{i=1}^{n_{\text{te}}} \alpha'_\ell K_\sigma(\cdot, \boldsymbol{Y}_{\text{te}}(\ell+1)) &= (1-\eta\lambda)\sum_{\ell=1}^{n_{\text{te}}} \alpha_\ell K(\cdot, \boldsymbol{Y}_{\text{te}}(\ell)) + \eta\frac{K_\sigma(\cdot, \boldsymbol{Y}_{\text{te}}(n_{\text{te}}+1))}{\hat{w}(\boldsymbol{Y}_{\text{te}}(n_{\text{te}}+1))}.
\end{aligned}$$

Matching the kernels on both sides of the equation gives

$$\begin{aligned}
\alpha'_\ell K_\sigma(\cdot, \boldsymbol{Y}_{\text{te}}(\ell+1)) &= (1-\eta\lambda)\alpha_{\ell+1} K(\cdot, \boldsymbol{Y}_{\text{te}}(\ell_1)) \\
\alpha'_\ell &= (1-\eta\lambda)\alpha_{\ell+1} \qquad (1)
\end{aligned}$$

for $\ell = 1, \ldots, n_{\mathrm{te}} - 1$, and for $\ell = n_{\mathrm{te}}$, we have

$$\alpha'_{n_{\mathrm{te}}} K_\sigma(\cdot, \boldsymbol{Y}_{\mathrm{te}}(n_{\mathrm{te}} + 1)) = \eta \frac{K_\sigma(\cdot, \boldsymbol{Y}_{\mathrm{te}}(n_{\mathrm{te}} + 1))}{\hat{w}(\boldsymbol{Y}_{\mathrm{te}}(n_{\mathrm{te}} + 1))}$$
$$\alpha'_{n_{\mathrm{te}}} = \frac{\eta}{\hat{w}(\boldsymbol{Y}_{\mathrm{te}}(n_{\mathrm{te}} + 1))}. \tag{2}$$

Thus, at every time step, the parameters $\{\alpha_\ell\}$ are updated according to equations (1) and (2). Normalization is carried out with respect to the following contraint:

$$\frac{1}{n_{\mathrm{rf}}} \sum_{t=1}^{n_{\mathrm{rf}}} \sum_{\ell=1}^{n_{\mathrm{te}}} \alpha_\ell K_\sigma(\boldsymbol{Y}_{\mathrm{rf}}(t + 1), \boldsymbol{Y}_{\mathrm{te}}(\ell + 1)) = 1.$$

4 Algorithm

The algorithm first uses cross-validation to select the optimal σ of the Gaussian kernel. The offline KLIEP is then performed on the first frame of the sliding window to estimate the density ratio w. At each new time step, the parameters $\{\alpha_\ell\}$ of the estimate $\hat{w}$ are updated online. If at any point the score S exceeds some threshold, then a changepoint is detected at time t_{te}. The sliding window then advances so that back of the window t_{rf} becomes the front of the old window $t_{\mathrm{te}} + n_{\mathrm{te}}$. This algorithm proceeds until we run out of data. The details of this algorithm are discussed in four separate algorithms in Kawahara and Sugiyama's paper.

References

[1] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*, volume 3. Springer, 2004.

[2] Y. Kawahara and M. Sugiyama. Sequential change-point detection based on direct density-ratio estimation. *Statistical Analysis and Data Mining*, 5(2):114–127, 2012.

[3] J. Kivinen, A. J. Smola, and R. C. Williamson. Online learning with kernels. In *Advances in neural information processing systems*, pages 785–792, 2001.

[4] M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1433–1440. Curran Associates, Inc., 2008.