Adaptive Streaming Anomaly Analysis

Zhao Xu
NEC Laboratories Europe
Heidelberg, Germany

Lorenzo von Ritter
Technical University of Munich
Garching, Germany

1 Introduction

Anomaly detection in streaming time series is of increasing importance in many real applications for enhancing the availability and security of systems. Streaming data usually contains complex dynamic patterns, which makes automatic model learning challenging. In this paper, we present a robust nonparametric Bayesian method AOTS for adaptive online anomaly detection in streaming time series. In particular, we develop an online prediction method for Student-t process (TP) to analyze sequentially arrived data. The heavy-tailed distributions also provide robustness against unknown anomalies in time series to better capture the normal patterns. Additionally, inspired by the idea of automatic Bayesian covariance discovery [1], we propose a kernel selection method based on submodular optimization to reduce manual efforts in model construction. Experiments on real data demonstrate the effectiveness of our method.

2 Automated Online Anomaly Detection

Assume that there is a time series stream $\mathbf{y} = \{y_1, y_2, \ldots\}$ of an infinite number of observations. We model the time series as a function $y_t = f(t)$ (shortened as f_t) with time t as predictor. The modeling method can avoid introducing anomalies into predictors like the autoregression based methods do, thus potentially reduces the complexity of modeling anomalies in time series.

Algorithm 1: Model construction for the AOTS

Input : Ω (candidate kernels), $\mathbf{y}$ (observed time series), τ (stop condition, default 0.01)
Initialization: $A = \emptyset$, K = None, k=None, r=1.0;
while $r > \tau$ **do**
  $A \leftarrow A \cup k$, $\Omega \leftarrow \Omega \setminus k$, $K \leftarrow K + k$;
  Find a kernel $k \in \Omega$ that offers minimal $\mathcal{NLL}$ together with the selected kernels A. Its hyperparameters are computed using gradient descent method with the $\mathcal{NLL}$ and the derivatives computed as follows:;
  $\mathcal{NLL} = \frac{n}{2}\log(\nu-2) + \log\left(B\left(\frac{\nu}{2}, \frac{n}{2}\right)\right) + \frac{1}{2}\log\left(|K+k|\right) + \frac{\nu+n}{2}\log\left(1 + \frac{\mathbf{y}^T(K+k)^{-1}\mathbf{y}}{\nu-2}\right)$;
  $\frac{\partial}{\partial\theta_i}\mathcal{NLL} = -\frac{1}{2}Tr\left(\left(\frac{\nu+n}{\nu+\beta-2}\alpha\alpha^T - (K+k)^{-1}\right)\frac{\partial k}{\partial\theta_i}\right)$;
  $\frac{\partial}{\partial\tilde{\nu}}\mathcal{NLL} = (\nu-2)\left[\frac{n}{2(\nu-2)} - \psi\left(\frac{\nu+n}{2}\right) + \psi\left(\frac{\nu}{2}\right)\frac{1}{2}\log\left(1+\frac{\beta}{\nu-2}\right) - \frac{(\nu+n)\beta}{2(\nu-2)(\nu+\beta-2)}\right]$;
  $r = (\mathcal{NLL}^{(A)} - \mathcal{NLL}^{(A\cup k)})/\mathcal{NLL}^{(A)}$;
Output : Selected kernels A and their hyperparameters

$K = LL^T$ (Cholesky decomposition), $\alpha = L^T \setminus (L \setminus \mathbf{y})$, and $\hat{\nu} = \log(\nu - 2)$. ψ is the digamma function.

We assume that the function f itself is unknown and random, and drawn from a Student-t process (TP) [2], which defines a robust nonparametric distribution over functions. Formally the generative process is: $\zeta|\nu, k \sim \mathcal{IWP}(\nu, k)$, $f|\mu, \zeta \sim \mathcal{GP}(\mu\ , (\nu-2)\zeta)$. That is, we first draw a kernel function ζ from an inverse Wishart process $\mathcal{IWP}(\nu, k)$, which is then used to draw functions (i.e. the time

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Algorithm 2: Online one-step ahead prediction for streaming time series

Input : y_{new} (new observation), $\mathbf{y}$ (previous observations), $\tilde{\mathbf{k}}_{new}$ (covariance of y_{new} and $\mathbf{y}$), $\tilde{\mathbf{k}}_*$ (covariance of f_{n+2} and $\mathbf{y}, y_{new}$), $k_{new,new}$ and $k_{*,*}$ (variances of f_{new} and f_{n+2})

$\tilde{\ell}_{new} = L \setminus \tilde{\mathbf{k}}_{new}$, $\tilde{\ell}_{new,new} = \left(k_{new,new} + \sigma^2 - \tilde{\ell}_{new}^T \tilde{\ell}_{new}\right)^{1/2}$;

$\tilde{a}_{new} = (y_{new} - \tilde{\ell}_{new}^T \mathbf{a})/\tilde{\ell}_{new,new}$, $L \leftarrow \begin{bmatrix} L & 0 \\ \tilde{\ell}_{new}^T & \tilde{\ell}_{new,new} \end{bmatrix}$, $\mathbf{a} \leftarrow \begin{bmatrix} \mathbf{a} \\ \tilde{a}_{new} \end{bmatrix}$;

$\mathbf{b} = L \setminus \tilde{\mathbf{k}}_*$, $\beta = \mathbf{a}^T \mathbf{a}$, $m_* = \mathbf{b}^T \mathbf{a}$, $var_* = \frac{\nu+\beta-2}{\nu+n-1}(k_{*,*} - \mathbf{b}^T \mathbf{b})$;

Output : m_* (mean) and var_* (variance) of the one-step ahead prediction.

series) from a Gaussian process $\mathcal{GP}(\mu, (\nu-2)\zeta)$. The TP has one more level than the GP, thus the sampled time series can be more flexible.

Kernels k play an important role, especially when dynamics of time series are often complicated [1]. To automate model construction of Student-t process with less manual intervention, we develop a greedy search method inspired by submodular optimization. The model construction problem is now formulated as: from a finite set Ω of base kernels (e.g. linear, squared exponential, periodic kernels) and their multiplicative compositions (e.g. linear $\times$ periodic), select a subset A minimizing the negative log likelihood $\mathcal{NLL}$ of the observed time series. Here we only consider multiplication of two kernels due to the possible overfitting problem. Not like Gaussian process, Student-t process is not closed under addition. Thus we incorporate all the selected kernels into the base kernel of the inverse Wishart process. The selection method is shown as Alg. 1. As the time series arrive as streams, batch mode prediction is prohibitively expensive. We develop an online learning method for efficient prediction, shown as Alg. 2. Then a predictive interval can be computed to describe the normality of the time series in the near future for anomaly detection.

3 Experiments and Conclusion

We evaluate the AOTS method using the airline passenger data with randomly added outliers. The first 120 time steps are given to learn the dynamics, and the rest is used as sequentially arrived streaming time series to detect anomalous events. The experimental results are reasonable and illustrated as Fig. 1. The left panel shows the predicted time series and the detected anomalies. The middle panel is the iteratively selected kernels with the gradually decreasing negative log likelihood. To give intuitions of the selected kernels, the right panel visualizes sample time series drawn with the kernels. The results demonstrate the superiority of the proposed online nonparametric Bayesian method with automatic kernel selection in streaming anomaly detection.

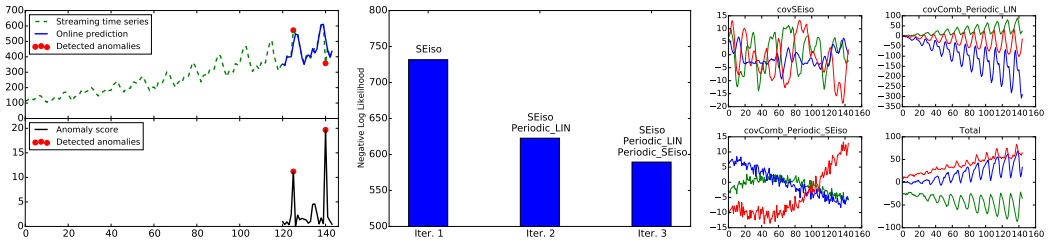

Figure 1: Detected anomalies from streaming time series.

References

[1] J. R. Lloyd, D. Duvenaud, R. Grosse, J. B. Tenenbaum, and Z. Ghahramani. Automatic construction and natural-language description of nonparametric regression models. In *AAAI*, 2014.

[2] A. Shah, A. Wilson, and Z. Ghahramani. Student-t processes as alternatives to gaussian processes. In *Proceedings of AISTATS*, 2014.