Journal of Computer Applications
计算机应用,2018,38(5):1346-1352
ISSN 1001-9081
CODEN JYIIDU
2018-05-10
http://www.joca.cn

文章编号:1001-9081(2018)05-1346-07
DOI:10.11772/j.issn.1001-9081.2017102511

基于异方差高斯过程的时间序列数据离群点检测

严　宏[1,2*],杨　波[2],杨红雨[2]
(1.中国民用航空飞行学院 计算机学院,四川 广汉 618307;　2.四川大学 国家空管自动化系统技术重点实验室,成都 610064)
(*通信作者电子邮箱 yanhonggh@163.com)

摘　要:时间序列数据在测量过程中通常受到事物内在可变性以及外界干扰等因素的影响,针对各个时间点上数据受影响程度不同的情况,提出一种基于高斯过程预估模型的时间序列数据离群点检测方法。将监测数据分解为标准值和偏差项两个部分,除了对理想情况下的标准值建模,还再次使用高斯过程实现对异方差偏差项的有效描述,通过变分推断解决引入偏差项后的后验概率求解问题,将后验分布中设定的容差区间用于离群点判定。使用雅虎公司公开的网络流量时序数据进行验证,模型输出的容差区间在不同时间点上的变化趋势与标注的正常数据偏差情况相符,并在对比实验中异常检测性能指标F1-score优于自回归积分滑动平均模型、一类支持向量机以及基于密度并伴随噪声的空间聚类算法。实验结果表明,该模型能够有效描述各个时间点上正常数据的分布情况,取得误报率和召回率两方面的综合权衡,而且可以避免模型参数设置不当导致的性能问题。

关键词:离群点检测;时间序列;高斯过程;异方差;变分推断
中图分类号:TP181　　**文献标志码**:A

Outlier detection in time series data based on heteroscedastic Gaussian processes

YAN Hong[1,2*], YANG Bo[2], YANG Hongyu[2]
(1. *College of Computer Science, Civil Aviation Flight University of China, Guanghan Sichuan* 618307, *China*;
2. *National Key Laboratory of Air Traffic Control Automation System Technology, Sichuan University, Chengdu Sichuan* 610064, *China*)

Abstract: Generally, there are inevitable disturbances in time series data, such as inherent uncertainties and external interferences. To detect outlier in time series data with time-varying disturbances, an approach based on prediction model using Gaussian Processes was proposed. The monitoring data was decomposed into two components: the standard value and the deviation term. As the basis of model for the ideal standard value without any deviation, Gaussian processes were also employed to model the heteroscedastic deviations. The posterior distribution of predicted data which is analytically intractable after introducing deviation term was approximated by variational inference. The tolerance interval selected from posterior distribution was used for outlier detection. Verification experiments were conducted on the public time series datasets of network traffic from Yahoo. The calculated tolerance interval coincided with the actual range of reasonable deviation existing in labeled normal data at various time points. In the comparison experiments, the proposed model outperformed autoregressive integrated moving average model, one-class support vector machine and Density-Based Spatial Clustering of Application with Noise (DBSCAN) in terms of F1-score. The experimental results show that the proposed model can effectively describe the distribution of normal data at various time points, achieve a tradeoff between false alarm rate and recall, and avoid the performance problems caused by improper parameter settings.

Key words: outlier detection; time series; Gaussian process; heteroscedasticity; variational inference

0　引言

在数据挖掘和机器学习领域,异常检测是指监测并判定数据集中与预期行为或模式不相符的单点数据、数据集合或数据序列[1]。异常数据通常与数据集中其他大部分数据在某种相似性度量上,比如欧氏距离(Euclidean distance)、马氏距离(Mahalanobis distance)、皮尔逊相关系数(Pearson correlation)等,存在预定义或显著的差异,往往预示着风险、安全事件或事故的发生,有助于采取预防或修正措施避免或减少由异常导致的损失,因此被广泛应用于医疗、交通、金融、电力和气象等行业。

时序数据是按时间先后顺序测量或记录的序列数据,蕴含着事物发展和变化的运行模式和内在规律。时间序列数据的异常检测具有广泛的应用场景,如网络流量监测[2]、医疗数据分析[3]、水文时序数据分析[4]以及瓦斯浓度监测[5]等,这些场景中根据应用需求的差异所需要检测的异常种类有所不同。根据异常特征以及表现形式的不同,时间序列数据异常可以分为数据点异常、前后关联异常和子序列异常[1]。本文关注的是数据点异常,对其的检测通常称为离群点检测。

目前,针对时间序列数据的离群点检测方法多种多样,其

收稿日期:2017-10-23;**修回日期**:2017-12-15;**录用日期**:2017-12-22。　**基金项目**:国家空管科研资助项目(GKG201403004)。
作者简介:严宏(1984—),男,四川攀枝花人,讲师,博士研究生,CCF会员,主要研究方向:机器学习、空管自动化;　杨波(1973—),男,四川成都人,副教授,博士,主要研究方向:空管自动化、机器学习;　杨红雨(1967—),女,四川成都人,教授,博士,主要研究方向:空管自动化、图像处理。

中主要的检测方法有 3 种:1)基于分类的方法[6-8],这类方法使用已标注是否异常的数据训练模型,能够有效利用训练样本的特性进行离群点判定,但是受限于训练样本的数据分布,当某些时间点上的样本缺失或不足时,将影响模型的异常检测性能。另外,该类方法需要对训练数据进行异常标注,大量数据下的标注工作将会增加该类方法的投入成本,降低可行性。2)基于聚类的方法[9-11]自动将数据集分为若干簇类,不属于任何簇类或簇类中数据数目较少的情况判别为离群点,该类方法属于无监督学习,避免了模型训练时对标注数据的依赖,但是依赖于对聚类模型参数的合理设置,不适当的参数设置将会严重地影响异常检测效果。3)基于预估模型的方法[2,12-13]通过构建模型计算正常模式下检测时间点的数据预测值,然后与实测值比较完成离群点判定。该类方法具有较好的直观性和解释性,但是难点在于需要使模型的预估数据与正常情况下的实际数据相符,或者是偏差小于合理的阈值。求取检测时间点的数据预测值属于回归问题,采用线性回归、多项式回归以及支持向量回归等模型只能求取检测时间点的单一预测值。在考虑容许一定范围偏差的实际应用情境中,如采用固定大小的阈值判定偏差有异常,那么又将面临阈值设置合理性以及阈值固定不变带来的问题。针对上述问题,本文提出一种基于高斯过程对时间序列数据正常模式进行建模的方法,利用高斯过程模型的特性求取正常情况下标准值和偏差数据的概率分布,构建的预估模型能够输出具有容差区间的预测值,并使用公开的真实网络流量数据进行实际应用场景下的有效性和性能验证。

1　问题及方法描述

1.1　问题描述

为了便于描述,采用如下符号对研究问题中涉及的数据进行统一的形式化描述。使用符号 t 通用地表示时间序列中任一时间点,本文将 t 由协调世界时(Coordinated Universal Time, UTC)之类的标准时间转换为以某个时间点为基准的相对时间,所测量的数据值由 $\boldsymbol{y}$ 表示。在构建预估模型之前,需要获取用于训练模型的数据集 $D = \{(t_1,y_1),(t_2,y_2),\cdots,(t_N,y_N)\}$,该训练集共有 N 组数据,每组数据中两个数据项 y 和 t 一一对应。如果进一步对数据进行人工标注,将每组数据 $(t_i,y_i)(i=1,2,\cdots,N)$,使用对应符号 δ 表明是否异常,δ_i 为 1 表示数据异常,δ_i 为 0 则表示数据正常,那么将会得到标注训练集 $D = \{(t_1,y_1,\delta_1),(t_2,y_2,\delta_2),\cdots,(t_N,y_N,\delta_N)\}$。使用训练集对模型进行训练后,对于需要检测的时间点 t_*,模型会输出表示时间序列数据正常行为或模式下 t_* 对应的 y_*;但是不同于常见的回归问题求解,由于通常情况下存在各种不可避免的干扰因素,测量的数据容许出现一定范围的偏差,因此,模型最终需要输出的并非是单一数值,而是一个容差区间,表示正常情况下 y_* 的取值范围。

1.2　方法概述

本文借助于高斯过程对时间序列数据的正常行为或模式进行建模,高斯过程近些年取得了丰富的研究成果,不仅在机器学习领域近些年的重要著作[14-15]中得以着重论述,Williams 等[16]还为机器学习中的高斯过程撰写了专著。高斯过程作为一类随机过程,常用于处理非线性等复杂回归问题,通常按如下形式定义:

$$f(t) \sim \mathcal{GP}(m(t),k(t,t'))$$

其中 $m(t)$ 和 $k(t,t')$ 分别为高斯过程的均值函数和协方差函数,定义如下:

$$m(t) = \mathrm{E}[f(t)]$$

$$k(t,t') = \mathrm{E}[(f(t)-m(t))(f(t')-m(t'))]$$

通常情况下由于信息有限,无法预先获知相应的均值函数,常见方法是使用常值函数 0 作为均值函数,这样整个高斯过程交由协方差函数确定,这种做法同样能够较好地描述整个过程中变量分布的变化情况,并且还在一定程度上降低了计算的复杂度[16]。由于高斯过程使用贝叶斯推断理论不仅可以计算得出预测值的最大后验估计,还能够得到一个估计的预测值概率分布。本文正是利用高斯过程该项特点计算得出正常情况下需要检测的时间点上数据值的预计分布情况,然后确定容差区间进行离群点检测。

2　离群点检测模型

2.1　时间序列数据分解

由于传感器误差、事物内在可变性、外界环境影响等不可避免的干扰因素,时间序列数据中测量的数据值往往存在合理范围内的偏差,为此建模过程中将实际测量的数据 $\boldsymbol{y}$ 作如下分解:

$$y_i = s_i + r_i;\ i = 1,2,\cdots,N \tag{1}$$

其中:s_i 表示的是时间点 t_i 上测量值在无偏差时应当得到的理想标准值,r_i 即为 y_i 中的偏差项。通过此方法完成分解后,时间序列中测量数据的形成过程更加明确,具有可解释性,然后分别对两个构成部分建模求取正常情况下的数据分布情况。

2.2　标准值建模

模型构建过程中使用高斯过程从部分已知标准值的基础上求解需要检测的时间点 t_* 上标准值 s_* 的后验估计,具体方法如下:

假定时间点集 $\boldsymbol{t} = \{t_i \mid i = 1,2,\cdots,N\}$ 对应的数据标准值 $\boldsymbol{s} = \{s_i \mid i = 1,2,\cdots,N\}$ 已知并构建对应向量,如上所述使用均值函数为常值 0 的高斯过程对标准值建模,协方差函数选用常见的 squared exponential 核函数[16],那么由高斯过程的概率分布定义可知,$\boldsymbol{s}$ 和 s_* 满足如下的多元高斯分布:

$$p(\boldsymbol{s},s_*) \sim \mathrm{N}(0,\boldsymbol{K}_{N+1})$$

其中 $\boldsymbol{K}_{N+1}$ 为此多元高斯分布的协方差矩阵,具有如下的形式:

$$\boldsymbol{K}_{N+1} = \begin{bmatrix} \boldsymbol{K}_N & \boldsymbol{k}_* \\ \boldsymbol{k}_*^{\mathrm{T}} & k_{**} \end{bmatrix}$$

其中

$$k_{**} = k(t_*,t_*)$$

$$\boldsymbol{k}_* = [k(t_1,t_*)\quad k(t_2,t_*)\quad \cdots\quad k(t_N,t_*)]^{\mathrm{T}}$$

$$\boldsymbol{K}_N = \begin{bmatrix} k(t_1,t_1) & k(t_1,t_2) & \cdots & k(t_1,t_N) \\ k(t_2,t_1) & k(t_2,t_2) & \cdots & k(t_2,t_N) \\ \vdots & \vdots & & \vdots \\ k(t_N,t_1) & k(t_N,t_2) & \cdots & k(t_N,t_N) \end{bmatrix}$$

由此可以应用高斯过程的推导结果[14-16]计算出需要检测的时间点 t_* 上标准值 s_* 的后验分布如下:

$$p(s_* \mid t_*,\boldsymbol{s},\boldsymbol{t}) \sim \mathrm{N}(\boldsymbol{k}_*^{\mathrm{T}}\boldsymbol{K}_N\boldsymbol{s},k_{**} - \boldsymbol{k}_*^{\mathrm{T}}\boldsymbol{K}_N^{-1}\boldsymbol{k}_*)$$

2.3 偏差项建模

从建模过程中时间序列数据的分解可知，式(1)中的偏差项 $\boldsymbol{r}$ 类似于大多数高斯过程模型中引入的噪声项[16]，但是通常情况下都假设各个时间点的噪声项符合高斯分布，均值为 0 并且其方差恒定不变。这种假设主要是为了简化后续推导，使得后验估计能够直接求得解析解，然而在现实中存在不同时间点上受干扰程度不一样导致合理的数据偏差程度发生变化的场景，为此本文假设偏差项 $\boldsymbol{r}$ 满足均值为 0 的高斯分布，但是其方差不再为固定数值。为了求得偏差项 $\boldsymbol{r}$ 与时间点 $\boldsymbol{t}$ 的函数关系，本文使用另外一个高斯过程对偏差项建模，更确切地讲是对偏差项方差的对数建模，主要是为了保证偏差项方差的非负特性，如下所示：

$$p(r_i) \sim \mathrm{N}(0,\sigma_r^2(t_i)); i = 1,2,\cdots,N \tag{2}$$

$$\ln(\sigma_r^2(t)) \sim \mathcal{GP}(\mu_r, k_r(t,t')) \tag{3}$$

其中 $k_r(t,t')$ 是偏差项方差 $\sigma_r^2(t)$ 对应高斯过程定义所使用的协方差函数，同样选用 squared exponential 核函数，而常值 μ_r 用于表示其偏差项方差的平均水平。

2.4 后验分布求解

通过上述对标准值和偏差项分别建模后，根据线性高斯模型相关理论[15]可知，式(1)表明向量 $\boldsymbol{y} = (y_1,y_2,\cdots,y_N)^{\mathrm{T}}$ 和向量 $\boldsymbol{s} = (s_1,s_2,\cdots,s_N)^{\mathrm{T}}$ 之间满足线性关系，已知的测量数据 $y_1,y_2,\cdots,y_N$ 和检测时间点 t_* 对应的标准值 s_* 仍然满足多元高斯分布，其协方差矩阵为：

$$\boldsymbol{\Sigma}_{N+1} = \begin{bmatrix} \boldsymbol{K}_N + \boldsymbol{K}_D & \boldsymbol{k}_* \\ \boldsymbol{k}_*^{\mathrm{T}} & k_{**} \end{bmatrix}$$

其中 $\boldsymbol{K}_D$ 是关于时间点 $t_1,t_2,\cdots,t_N$ 上偏差项的对角矩阵 $\mathrm{diag}(\boldsymbol{r})$，对角元素 $\boldsymbol{r} = (r_1,r_2,\cdots,r_N)^{\mathrm{T}}$。至此，可以再次利用高斯过程的推导结论[14-16]，可以计算出关于 s_* 的后验分布 $p(s_* \mid t_*,\boldsymbol{y},\boldsymbol{t},\boldsymbol{r},r_*)$，附加上偏差项 r_* 就可以得到时间点 t_* 上正常情况下预估数据 y_* 的后验分布：

$$p(y_* \mid t_*,\boldsymbol{y},\boldsymbol{t},\boldsymbol{r},r_*) \sim \mathrm{N}(\mu_*,\sigma_*^2)$$

其中

$$\mu_* = \boldsymbol{k}_*^{\mathrm{T}}(\boldsymbol{K}_N + \boldsymbol{K}_D)\boldsymbol{y} \tag{4}$$

$$\sigma_*^2 = k_{**} - \boldsymbol{k}_*^{\mathrm{T}}(\boldsymbol{K}_N + \boldsymbol{K}_D)^{-1}\boldsymbol{k}_* + r_* \tag{5}$$

由于式(4)和(5)中 $\boldsymbol{K}_D$ 涉及的 $\boldsymbol{r}$ 以及式(5)中的 r_* 未完成求解，因此为了完成对 y_* 的后验分布求解，需要进一步对其中的 $\boldsymbol{r}$ 和 r_* 进行积分，然后得到：

$$p(y_* \mid t_*,\boldsymbol{y},\boldsymbol{t}) = \iint p(y_* \mid t_*,\boldsymbol{y},\boldsymbol{t},\boldsymbol{r},r_*)p(\boldsymbol{r},r_* \mid t_*,\boldsymbol{y},\boldsymbol{t})\,\mathrm{d}\boldsymbol{r}\mathrm{d}r_*$$

仅从目前的模型假设以及已知时间序列数据中无法求解 $p(\boldsymbol{r},r_* \mid t_*,\boldsymbol{y},\boldsymbol{t})$，尽管可以采用文献[17]中提到的采样方法求取近似解，但是存在计算量大、耗时长的缺点。本文采用文献[18]中的思路利用变分推断求取近似解，使用变分推断中常用的平均场(mean field)方法[15]，对边缘概率 $p(\boldsymbol{y})$ 的对数作如下分解：

$$\ln(p(\boldsymbol{y})) = L(q(\boldsymbol{s}),q(\boldsymbol{r})) + \mathrm{KL}(q(\boldsymbol{s})q(\boldsymbol{r}) \parallel p(\boldsymbol{s},\boldsymbol{r} \mid \boldsymbol{y})) \tag{6}$$

其中 $\mathrm{KL}(\cdot \parallel \cdot)$ 表示 KL 散度(Kullback-Leibler divergence)，而 $L(q(\boldsymbol{s}),q(\boldsymbol{r}))$ 是 $\ln(p(\boldsymbol{y}))$ 的下界，当寻求分布 $q(\boldsymbol{s})$ 和 $q(\boldsymbol{r})$ 最大化该下界时，也就会使式(6)中的 KL 散度最小化，从而优化 $p(\boldsymbol{s},\boldsymbol{r} \mid \boldsymbol{y})$ 在分解形式 $q(\boldsymbol{s})q(\boldsymbol{r})$ 下的近似。采用文献[18]中下界 $L(q(\boldsymbol{s}),q(\boldsymbol{r}))$ 关于 $q(\boldsymbol{s})$ 最大化的推导结论，可以进一步得到一个仅依赖于 $q(\boldsymbol{r})$ 的近似下界：

$$L(q(\boldsymbol{r})) = \ln Z(q(\boldsymbol{r})) - \mathrm{KL}(q(\boldsymbol{r}) \parallel p(\boldsymbol{r})) \tag{7}$$

其中

$$Z(q(\boldsymbol{r})) = \int \exp\left(\int q(\boldsymbol{r}) \ln p(\boldsymbol{y} \mid \boldsymbol{s},\boldsymbol{r})\,\mathrm{d}\boldsymbol{r}\right) p(\boldsymbol{s})\,\mathrm{d}\boldsymbol{s}$$

对于 $q(\boldsymbol{r})$ 的分布采用变分推断中常用的多元高斯分布，其均值向量和协方差矩阵分别用 $\boldsymbol{\mu}_q$ 和 $\boldsymbol{\Sigma}_q$ 表示，结合之前对 $\boldsymbol{s}$ 多元高斯分布的设定以及式(2)和(3)，可以进一步推导式(7)中的下界为如下形式：

$$L(q(\boldsymbol{r})) = \ln \mathrm{N}(0,\boldsymbol{K}_N + \boldsymbol{R}) - \frac{1}{4}\mathrm{tr}(\boldsymbol{\Sigma}_q) - \mathrm{KL}(\mathrm{N}(\boldsymbol{r} \mid \boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q) \parallel \mathrm{N}(r \mid \mu_r 1,\boldsymbol{K}_r)) \tag{8}$$

其中：$\mathrm{tr}(\cdot)$ 表示矩阵的迹，$\boldsymbol{R}$ 为一对角矩阵，其对角元素为 $[\boldsymbol{R}]_{ii} = \exp([\boldsymbol{\mu}_q]_i - [\boldsymbol{\Sigma}_q]_{ii})$，而 $\boldsymbol{K}_r$ 是使用协方差函数 $k_r(\boldsymbol{t},\boldsymbol{t}')$ 计算的协方差矩阵。根据式(8)的极值点与偏导数的关系，可以求得：

$$\boldsymbol{\mu}_q = \boldsymbol{K}_r\left(\boldsymbol{\Lambda} - \frac{1}{2}\boldsymbol{I}\right) + \mu_r 1$$

$$\boldsymbol{\Sigma}_q = (\boldsymbol{K}_r^{-1} + \boldsymbol{\Lambda})^{-1}$$

其中 $\boldsymbol{\Lambda}$ 表示半正定对角矩阵。至此，模型的训练可以归结为对最大化式(8)表示的下界，其中需要优化的参数包括标准值 $\boldsymbol{s}$ 和偏差项 $\boldsymbol{r}$ 对应的两个高斯过程协方差函数中的参数、矩阵 $\boldsymbol{\Lambda}$ 中的对角线元素以及用于控制偏差项方差平均水平的 μ_r，本文选用模型训练中常见的共轭梯度法[19]进行参数优化。使用训练集完成参数优化后，利用变分推断结果可以得到 s_* 的后验分布：

$$p(s_* \mid t_*,\boldsymbol{y},\boldsymbol{t}) \sim \mathrm{N}(s_* \mid a_*,c_*^2)$$

其中

$$a_* = \boldsymbol{k}_*^{\mathrm{T}}(\boldsymbol{K}_N + \boldsymbol{R})\boldsymbol{y} \tag{9}$$

$$c_*^2 = k_{**} - \boldsymbol{k}_*^{\mathrm{T}}(\boldsymbol{K}_N + \boldsymbol{R})^{-1}\boldsymbol{k}_* \tag{10}$$

根据式(2)和(3)中对偏差项方差的假设，代入其在时间点 t_* 上的期望值也是最大概率值作为近似求解 y_* 的偏差项估计，得到如下结果：

$$p(y_* \mid t_*,\boldsymbol{y},\boldsymbol{t}) \sim \mathrm{N}(y_* \mid a_*,{d_*}^2)$$

其中 y_* 的均值与式(9)等同，而方差为式(10)中的标准值方差加上偏差项引入的方差：

$${d_*}^2 = k_{**} - \boldsymbol{k}_*^{\mathrm{T}}(\boldsymbol{K}_N + \boldsymbol{R})^{-1}\boldsymbol{k}_* + \exp\left(\boldsymbol{k}_{r*}^{\mathrm{T}}\left(\boldsymbol{\Lambda} - \frac{1}{2}\boldsymbol{I}\right)1 + \mu_r\right) \tag{11}$$

其中 $\boldsymbol{k}_{r*} = [k_r(t_1,t_*) \quad k_r(t_2,t_*) \quad \cdots \quad k_r(t_N,t_*)]^{\mathrm{T}}$。

2.5 离群点判定

通过上述推导过程得到检测时间点 t_* 上 y_* 的后验分布后，可以选取其中特定范围的数据分布区间作为正常情况时间点 t_* 上测量数据的容差区间，如果需要检测的实际数据 $\boldsymbol{y}$ 不位于该区间，那么就将其判定为离群点数据，相应的判别函数可用如下形式描述：

$$f(\boldsymbol{y}) = \begin{cases} \mathrm{false}, & \boldsymbol{y} \in [a_* - z_\alpha d_*, a_* + z_\alpha d_*] \\ \mathrm{true}, & \boldsymbol{y} \notin [a_* - z_\alpha d_*, a_* + z_\alpha d_*] \end{cases}$$

其中关于 a_* 和 d_* 来自于式(9)和式(11)的推导结果，而 z_α

由模型中设置容差区间占比的超参数 α 确定,超参数 α 并不从训练数据中学习,而是根据实际应用需求进行调整设置。例如,使用高斯分布中常用的 95% 置信区间作为容差区间时,即将超参数 α 设置为 95%,那么 z_α 可近似取值为 1.96。

2.6　异常训练数据过滤

在训练集 D 中数据没有异常标注项 δ 的情况下,包含在其中的异常数据会引起模型训练中常见的噪声数据干扰问题。由于高斯过程本身对噪声数据具有一定的抗干扰能力[16],本文利用该项特性设计了一种迭代训练模型的方法,基本思路是通过当前模型输出反复过滤当前训练集然后重新训练直至过滤的异常数据小于既定的容差比例或者达到最大训练次数为止,具体操作如下描述:

输入　训练集 $D = \{(t_i, y_i) \mid i = 1, 2, \cdots, N\}$,容差区间占比 α,最大迭代训练次数 m。

输出　离群点检测模型参数 $\boldsymbol{\theta}$。

过程　离群点检测模型训练。

1)　设置集合 $D_s = D, D_a = \varnothing$

2)　随机初始化模型参数 $\boldsymbol{\theta}$

3)　$t = 0$

4)　repeat

5)　　设置当前训练集 $D_s = D_s \backslash D_a$

6)　　使用 D_s 训练模型,更新参数 $\boldsymbol{\theta}$

7)　　当前异常数据集 $D_a = \varnothing$

8)　　for $(t_j, y_j) \in D_s$ do

9)　　　计算时间 t_j 的容差区间 $[a_j - z_\alpha d_j, a_j + z_\alpha d_j]$

10)　　　if $y_j \notin [a_j - z_\alpha d_j, a_j + z_\alpha d_j]$ then

11)　　　　添加异常数据 $D_a = D_a \cup (t_j, y_j)$

12)　　　end if

13)　　end for

14)　　$t = t + 1$

15)　until $(|D_a| / |D_s|) \leqslant (1 - \alpha) \vee (t > m)$

上述训练方法能够使模型输出的容差区间反映出无标注数据集中大部分数据的分布情况,这将在下一章中得以验证。

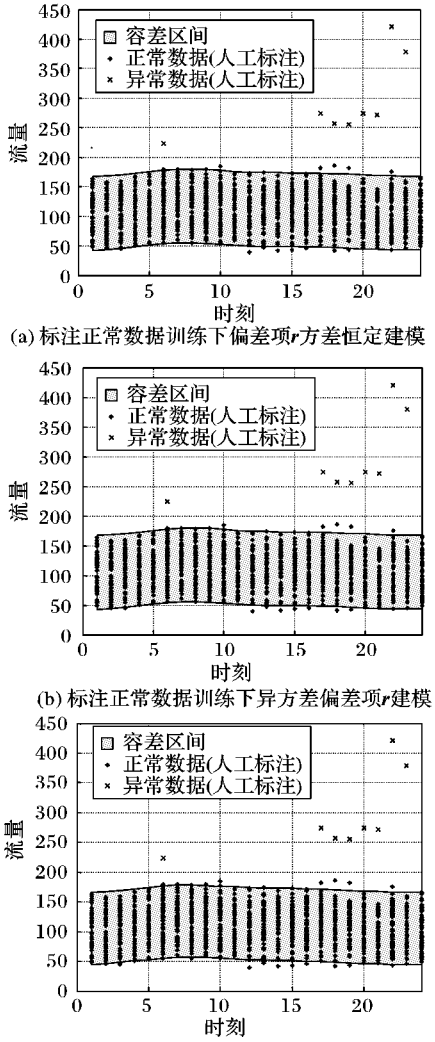
3　实验结果及分析

3.1　效果验证

为了验证对时间序列数据中离群点数据的检测效果,使用雅虎公司 Webscope 项目[20]提供的公开时序数据进行测试,该数据来源于雅虎公司真实场景中网络节点的流量统计,用于研究网络流量时序数据中的异常检测,内容包括全天 24 小时以整点开始每个小时内的流量统计,并且包含每个流量数据是否异常的人工标注。实验中使用 Matlab 编写模型代码,版本为 Matlab R2013a,操作系统使用 Windows 7 Professional,硬件环境为 Intel Core i7-6700K 4.0 GHz 以及 16 GB DDR4-2133。由于篇幅有限,选取 Webscope 项目中三个场景作为案例演示,首先基于人工标注的正常流量数据,分别采用高斯过程建模中常见的偏差项方差恒定不变和异方差偏差项两种方法对正常流量数据的容差区间进行建模,然后为了验证针对无人工标注是否异常的模型训练方法效果,将人工标注的异常数据和正常数据全部用于容差区间的模型训练,分析对比了三种方法在各个场景下容差区间建模和离群点检测效果。

1)场景一。

图 1 分别显示场景一中的网络流量数据分布和使用上述三种方法完成的容差区间建模结果,由于 Webscope 项目原始数据中没有流量单位说明故而纵坐标流量单位未标注,数据对应横坐标为流量统计的整点时刻。



(a) 标注正常数据训练下偏差项r方差恒定建模

(b) 标注正常数据训练下异方差偏差项r建模

(c) 无标注数据训练下异方差偏差项r建模

图 1　场景一的实验效果

Fig. 1　Experimental results of scenario 1

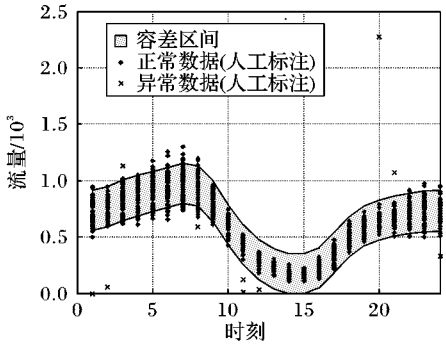
从图中数据分布可以看出该场景下正常流量数据在各个时间点上边界值以及分布区间长度方面较为相似,使用三种方法计算得出的容差区间在直观效果上几乎相同,对于远离正常流量数据分布区间的异常数据都能实现成功检测,对于孤立的边界正常数据也都出现误报的现象,但这也从一定程度上体现了这些孤立边界数据和集中分布正常数据之间的差异性。实验结果表明在各个时间点上正常数据分布相似的情况下,采用异方差高斯过程建模能够取得与常见的偏差项方差恒定假设类似的建模效果,计算得出的容差区间长度在各个时间点上近似相同,与正常数据在各个时间点的分布情况相符。

2)场景二。

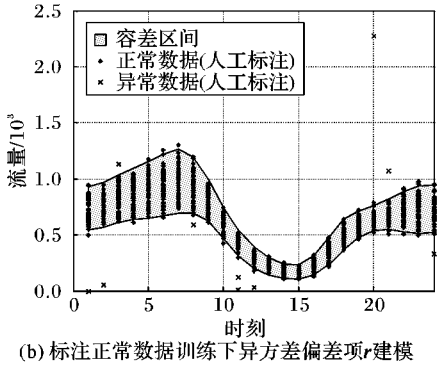
图 2 将分别显示了场景二中的网络流量数据分布和使用上述三种方法完成的容差区间建模结果。

不同于场景一,该场景中正常流量数据在各个时间点上差异较大,在使用标注正常数据训练的情况下,若采用高斯过程建模中常见的偏差项方差恒定的假设,必然会使计算得出

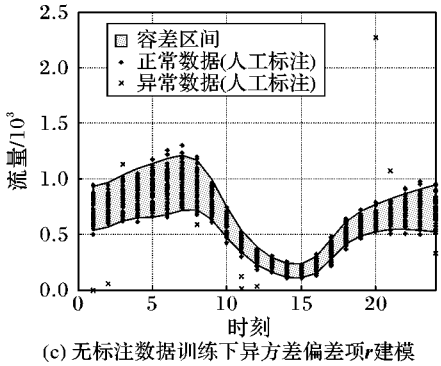
的容差区间长度在各个时间点上保持不变,与实际正常数据分布情况不符,这在图 2(a)中得到较为直观的验证,导致时间点 5、6、7 和 8 时,容差区间过小而出现较多正常数据误报。使用异方差高斯过程建模后,尽管还是出现了部分孤立边界数据误报的情况,但是在时间点 5、6、7 和 8 时,图 2(a)中部分误报的数据在图 2(b)中处于容差区间内,误报率得以下降,并且从图 2(b)中可以看出,计算得到的容差区间在各个时间点上的变化与该场景下正常流量数据在各个时间点上的合理偏差趋势更为相符。此外,对于全部数据无标注是否异常的情况,如图 2(c)所示,本文训练方法计算的容差区间相对于图 2(b)在时间点 7、8、16、17 和 18 上的容差区间更小,出现了更多的误报数据,但是计算得出的容差区间其长度在各个时间点上的变化还是与正常数据的合理偏差趋势基本相符,并且完成了所有标注异常数据的检测。



(a) 标注正常数据训练下偏差项r方差恒定建模



(b) 标注正常数据训练下异方差偏差项r建模



(c) 无标注数据训练下异方差偏差项r建模

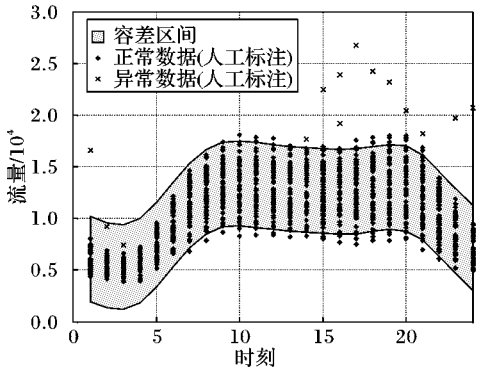
图 2 场景二的实验效果

Fig. 2 Experimental results of scenario 2

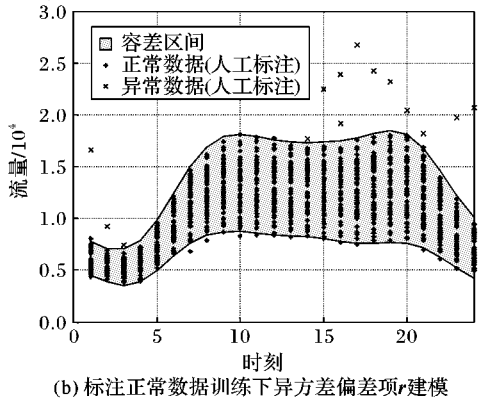
3)场景三。

图 3 将分别显示了场景三中的网络流量数据分布和使用上述三种方法完成的容差区间建模结果。

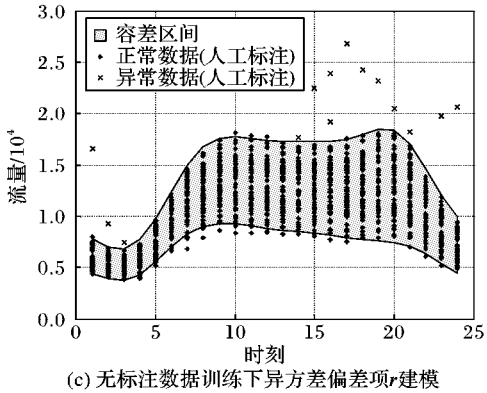
与场景二类似,该场景中实验结果再次表明:采用异方差高斯过程构建预估模型的离群点检测方法,可以如实地描述正常流量数据在各个时间点上的分布变化情况;而对于无标注的流量数据,本文模型训练方法存在对训练数据的异常判定和迭代过滤,相对于直接使用标注正常数据训练模型,会使得较为稀疏的正常边界数据中更多数据点被判定为异常,这也正是图 3(c)所示在时间点 4 ~ 17 上的容差区间相对于图 3(b) 更小,出现较多误报数据的原因,但还是与正常数据偏差范围的变化趋势基本相符。



(a) 标注正常数据训练下偏差项r方差恒定建模



(b) 标注正常数据训练下异方差偏差项r建模



(c) 无标注数据训练下异方差偏差项r建模

图 3 场景三的实验效果

Fig. 3 Experimental results of scenario 3

与场景二不同的是,该场景下高斯过程建模中常见的偏差项方差恒定假设不仅导致时间点 10 ~ 20 上误报率过高,而且在时间点 2 和 3 上的标注异常数据还被判定为正常,出现了异常漏报的情况,进一步体现了假设偏差项方差恒定的方法无法描述正常数据偏差范围变化趋势的缺陷。

3.2 性能对比

为了进一步与目前时序数据离群点检测的常用方法进行性能比较,选取相关研究中基于一类支持向量机(One-class SVM)[10-11]、自回归积分滑动平均模型(Autoregressive Integrated Moving Average Model, ARIMA)[2,12]以及基于密度

并伴随噪声的空间聚类算法(Density-based Spatial Clustering of Application with Noise, DBSCAN)[9]三种方法作为对比,选用异常检测中常用的误报率、召回率和 F1-score 作为性能对比的度量指标。实验环境中使用的硬件和软件与 3.1 节中的描述相同。实验中本文模型中容差区间占比 α 设置为 95%,鉴于此基于 One-class SVM 的模型其训练样本异常检出比率上界设置为 5%,核函数采用常用的径向基函数(Radial Basis Function, RBF),使用通常的网格搜索进行参数调优,经过预先的参数设置对误报率、召回率和 F1-score 的影响测试后,将 RBF 核函数参数 γ 搜索区间设定为[0.000 1,0.01],搜索步长为 0.000 1,最终选取三次参数设置作为示例说明,包括使得 F1-score 在其中取值最高的参数设置。基于 ARIMA 的模型根据 ADF 检验(Augmented Dickey-Fuller test)确定差分阶数 d,通过贝叶斯信息准则(Bayesian Information Criterion, BIC)确定自回归和移动平均阶数 p 和 q,同样鉴于本文模型的容差区间占比设置将预测值的 95% 置信区间作为正常数据区间,不位于该置信区间的数据判定为离群点。DBSCAN 算法虽然不像 k-means 等聚类算法需要预先确定聚类簇数,但其效果还是与邻域半径 ε 和核心对象邻域内最小对象个数 *MinPts* 两个参数密切相关,为了寻求能使反映误报率和召回率两方面的综合性能指标 F1-score 取值最大的参数设置,同样借鉴网格搜索的思路进行多组参数的比较,邻域半径 ε 搜索区间设置[0.01,0.1],搜索步长为 0.01,邻域内最小对象个数 *MinPts* 搜索区间设置[5,15],搜索步长为 1,最后除了选取 F1-score 取值最大的参数设置进行模型对比,还选取了其他 4 组能使误报率或召回率取得较好结果的参数设置作为示例说明。实验中对雅虎公司 Webscope 项目中选用的 50 个时序数据集进行数据归一化预处理,避免各个时序数据集数值范围不同造成的影响,各个对比模型、参数说明和性能指标数据如表 1 所示。

表 1　性能指标对比

Tab. 1　Comparison of performance indicators

模型	参数说明	误报率/%	召回率/%	F1-score
基于 One-class SVM 的模型	RBF 核函数参数 $\gamma=0.01$	21.67	99.04	0.874 8
	RBF 核函数参数 $\gamma=0.001$	1.91	98.49	0.982 9
	RBF 核函数参数 $\gamma=0.000\,5$	3.96	96.38	0.962 1
基于 ARIMA 的模型	预测值 95% 置信区间判定正常	2.43	98.73	0.981 5
基于 DBSCAN 的模型	$\varepsilon=0.04$, *MinPts* = 12	32.51	99.67	0.804 8
	$\varepsilon=0.06$, *MinPts* = 12	1.27	93.06	0.958 1
	$\varepsilon=0.05$, *MinPts* = 12	1.58	97.61	0.980 1
	$\varepsilon=0.05$, *MinPts* = 10	1.02	92.79	0.957 9
	$\varepsilon=0.05$, *MinPts* = 15	10.29	99.46	0.943 3
偏差项方差恒定高斯过程模型(使用标注正常数据训练)	容差区间占比 95%	13.96	89.18	0.875 8
偏差项异方差高斯过程模型(使用标注正常数据训练)	容差区间占比 95%	1.29	98.56	0.986 3
偏差项异方差高斯过程模型(使用全部无标注数据训练)	容差区间占比 95%	3.06	99.70	0.983 0

基于 One-class SVM 的模型在网格搜索参数过程中,当 RBF 核函数参数 γ 设置为 0.001 时 F1-score 性能指标数值最高,在所有模型对比中异常检测效果也相对较好,但该模型会受到核函数参数设置的影响。从表 1 中实验数据可以看出当参数 γ 增加到 0.01 时会提高召回率,但也导致误报率过高,另外当参数 γ 降低到 0.000 5 时出现了参数设置不当造成召回率降低并且误报率反而增加的情况。基于 DBSCAN 的模型的性能指标同样受到参数设置的影响,使用(ε = 0.05, *MinPts* = 12)一组参数设置时才取得误报率和召回率之间较好的权衡考量,F1-score 性能指标在该模型所有参数设置中最好,使用(ε = 0.06, *MinPts* = 12)和(ε = 0.05, *MinPts* = 10)两组参数设置时能进一步降低误报率,但也导致召回率下降,使得 F1-score 性能指标降低,使用(ε = 0.04, *MinPts* = 12)和(ε = 0.05, *MinPts* = 15)两组参数设置时召回率取得较高数值,但却出现了过高的误报率,难以投入实际应用。基于异方差高斯过程模型的离群点检测方法在使用标注正常数据训练模型的情况下,与高斯过程建模中常见的偏差项方差恒定不变的方法相比,能够计算得出更加合理的容差区间,取得显著的性能提升。尽管该模型在误报率和召回率两个单项指标上没有取得所有实验结果中的最高数值,但在综合指标 F1-score 上相对于其他模型都取得了一定程度的提升。在使用全部无标注数据训练的情况下,该模型在召回率和 F1-score 性能指标上也取得较为满意的结果。此外,该模型另一优势在于高斯过程模型相关参数通过训练集优化确定,避免了其他模型中出现的因参数设置不当造成误报率过高或召回率过低的情况。

4　结语

本文基于预估模型检测时序数据离群点检测方法并没有直接针对监测数据进行数学建模,而是首先将监测数据分解为标准值和偏差项两个部分,这种做法与常见的高斯过程建模加入噪声项的方法类似,但是区别在于并没有假定偏差项独立同分布以致于方差恒定不变,而是再次使用高斯过程对各个时间点的偏差项建模,从而能够基于异方差高斯过程对不同时间点上正常数据合理偏差范围变化实现有效的数学描述。通过实验数据表明,本文的离群点检测方法能够取得误报率和召回率两个方面较好的权衡,并且无需考虑关键模型参数的人工设置,避免参数设置不当对性能指标的严重影响。在之后的研究工作中,还需要进一步考虑不同应用场景下高斯过程协方差函数的选取以及容差区间占比设置对于离群点

检测性能的影响以及改进。

参考文献(References)

[1] CHANDOLA V, BANERJEE A, KUMAR V. Anomaly detection: a survey[J]. ACM Computing Surveys, 2009, 41(3): 1 - 58.

[2] YAACOB A H, TAN I K T, SU F C, et al. ARIMA based network anomaly detection[C]// Proceedings of the 2nd International Conference on Communication Software and Networks. Piscataway, NJ: IEEE, 2010: 205 - 209.

[3] LIN J, KEOGH E, FU A, et al. Approximations to magic: finding unusual medical time series[C]// Proceedings of the 2005 IEEE Symposium on Computer-Based Medical Systems. Piscataway, NJ: IEEE, 2005: 329 - 334.

[4] 余宇峰, 朱跃龙, 万定生, 等. 基于滑动窗口预测的水文时间序列异常检测[J]. 计算机应用, 2014, 34(8): 2217 - 2220. (YU Y F, ZHU Y L, WAN D S, et al. Time series outlier detection based on sliding window prediction[J]. Journal of Computer Applications, 2014, 34(8): 2217 - 2220.)

[5] 张宝燕, 李茹, 穆文瑜. 基于混沌时间序列的瓦斯浓度预测研究[J]. 计算机工程与应用, 2011, 47(10): 244 - 248. (ZHANG B Y, LI R, MU W Y. Study on gas concentration prediction based on chaotic time series[J]. Computer Engineering and Applications, 2011, 47(10): 244 - 248.)

[6] SEVAKULA R K, VERMA N K. Clustering based outlier detection in fuzzy SVM[C]// Proceedings of the 2014 IEEE International Conference on Fuzzy Systems. Piscataway, NJ: IEEE, 2014: 1172 - 1177.

[7] MARTINS H, PALMA L, CARDOSO A, et al. A support vector machine based technique for online detection of outliers in transient time series[C]// Proceedings of the 2015 10th Asian Control Conference. Piscataway, NJ: IEEE, 2015: 1 - 6.

[8] DANG T T, NGAN H Y T, LIU W. Distance-based k-nearest neighbors outlier detection method in large-scale traffic data[C]// Proceedings of the 2015 IEEE International Conference on Digital Signal Processing. Piscataway, NJ: IEEE, 2015: 507 - 510.

[9] ABID A, KACHOURI A, MAHFOUDHI A. Outlier detection for wireless sensor networks using density-based clustering approach [J]. IET Wireless Sensor Systems, 2017, 7(4): 83 - 90.

[10] JIANG J, YASAKETHU L. Anomaly detection via one class SVM for protection of SCADA systems[C]// Proceedings of the 2013 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery. Washington, DC: IEEE Computer Society, 2013: 82 - 88.

[11] NGAN H Y T, YUNG N H C, YEH A G O. A comparative study of outlier detection for large-scale traffic data by one-class SVM and kernel density estimation[J]. Proceedings of SPIE - the International Society for Optical Engineering, 2015, 9405: 94050I-1 - 94050I-10.

[12] PENA E H M, BARBON S, RODRIGUES J J P C, et al. Anomaly detection using digital signature of network segment with adaptive ARIMA model and paraconsistent logic[C]// Proceedings of the 2014 IEEE Symposium on Computers and Communication. Piscataway, NJ: IEEE, 2014: 1 - 6.

[13] FERNANDES G, PENA E H M, CARVALHO L F, et al. Statistical, forecasting and metaheuristic techniques for network anomaly detection[C]// Proceedings of the 30th Annual ACM Symposium on Applied Computing. New York: ACM, 2015: 701 - 707.

[14] BISHOP C M. Pattern Recognition and Machine Learning (Information Science and Statistics)[M]. New York: Springer, 2006: 303 - 319.

[15] MURPHY K P. Machine Learning: a Probabilistic Perspective [M]. Cambridge, MA: MIT Press, 2012: 79 - 91, 515 - 542.

[16] WILLIAMS C K I, RASMUSSEN C E. Gaussian Processes for Machine Learning[M]. Cambridge, MA: MIT Press, 2006: 7 - 30, 79 - 102.

[17] GOLDBERG P W, WILLIAMS C K I, BISHOP C M. Regression with input-dependent noise: a Gaussian process treatment[C]// NIPS 1998: Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 1998: 493 - 499.

[18] LÁZARO-GREDILLA M, TITSIAS M K. Variational heteroscedastic Gaussian process regression[C]// ICML 2011: Proceedings of the 2011 International Conference on Machine Learning. New York, NY: ACM, 2011: 841 - 848.

[19] NOCEDAL J, WRIGHT S. Numerical Optimization[M]. New York: Springer, 2006: 101 - 134.

[20] Yahoo! Inc. Webscope dataset ydata labeled time series anomalies v1.0 [EB/OL]. [2015-03-24]. https://webscope.sandbox.yahoo.com/catalog.php?datatype=s&did=70.

This work is partially supported by the National Air Traffic Control Research Project (GKG201403004).

YAN Hong, born in 1984, Ph. D. candidate, lecturer. His research interests include machine learning, air traffic control automation.

YANG Bo, born in 1973, Ph. D., assistant professor. His research interests include air traffic control automation, machine learning.

YANG Hongyu, born in 1967, Ph. D., professor. Her research interests include air traffic control automation, image processing.

(上接第 1333 页)

[23] SUN Y, HAN J, GAO J, et al. iTopicModel: Information network-integrated topic modeling[C]// Proceedings of the 9th IEEE International Conference on Data Mining. Washington, DC: IEEE Computer Society, 2009: 493 - 502.

This work is partially supported by the Ministry of Education - China Mobile Research Fund (MCM20160307), the Sichuan Province Science and Technology Innovation Talent Project and the International Cooperation Project of Chengdu Municipal Science and Technology Bureau (2016-GH02-00048-HZ, 2015-GH02-00041-HZ).

QIU Qingyu, born in 1994, M. S. candidate. His research interests include machine learning, data mining.

LI Jing, born in 1988, M. S., engineer. Her research interests include big data, artificial intelligence.

QUAN Bing, born in 1988, M. S., engineer. His research interests include big data, artificial intelligence.

TONG Chao, born in 1988, M. S. His research interests include big data, machine learning.

ZHANG Lijun, born in 1978, M. S., engineer. Her research interests include machine learning, data mining.

ZHANG Haixian, born in 1980, Ph. D., associate professor. Her research interests include deep neural network.