文章编号：2095-2163(2022)07-0059-10 中图分类号：TP301.6 文献标志码：A

基于 GPS 数据的震前短临异常检测

李 南，林莉莉
（福建农林大学 计算机与信息学院，福州 350000）

摘 要：随着 GPS 台站的普及，结合 GPS 数据进行时间序列数据异常检测已成为热门研究领域。针对现有方法普遍存在的主观性强、普适性差等问题，运用鞅理论，提出了一种基于 GPS 数据的震前短临异常检测算法（Anomaly Detection Algorithm based on GPS data，ADA）。实验结果表明，ADA 算法所检测到的 GPS 数据中，异常出现时间与地震发生时间存在显著相关，与时间序列异常检测中传统的 $k\sigma$ 准则和主流的异常检测模型 ARIMA、单类别支持向量机 OCSVM 以及基于两阶段聚类的异常检测算法 TSOD 相比，ADA 算法能够更直观、准确地反映震前 GPS 数据中出现的异常，不易出现误报的情况。

关键词：地震；短临异常；GPS；鞅理论

Short impending anomaly detection before earthquake based on GPS data

LI Nan，LIN Lili
（College of Computer and Information Sciences，Fujian Agriculture and Forestry University，Fuzhou 350000，China）

【Abstract】 Short impending anomaly detection before earthquake is a key part of the earthquake early warning. With the rapid popularization of GPS stations，anomaly detection before earthquake based on GPS data has become a hot research area. In order to solve the problem of strong subjectivity and poor universality in the existing methods，an algorithm based on Martingale theory is proposed to detect the short impending anomalies in GPS data. The experimental results show that the detected short impending anomaly from GPS data has a significant correlation with the corresponding earthquake. Compared with the traditional $k\sigma$ analysis method，the famous anomaly detection model ARIMA，one-class support vector machines（OCSVM）and two-stage clustering algorithm for outlier detection（TSOD），the proposed algorithm can reveal the pregnancy of the earthquake more clearly and accurately，and has less false alarms.

【Key words】 earthquake；short impending anomaly；GPS；Martingale theory

0 引 言

震前短临异常检测已成为防震减灾研究与应用领域的研究热点。当前，国内外学者大多采用射出长波辐射[1]、电离层电子总含量[2]等电、热指标，来进行短临异常检测，以探测震前一段时间及一定区域内可能发生的地球物理、化学变化。但这些方法普遍存在观测数据不足、容易受到人为活动干扰等缺陷[3]。

近年来，随着全球导航卫星系统技术（Global Navigation Satellite System，GNSS）的发展和 GPS 台站的普及，利用 GPS 数据进行短临异常检测已成为热门研究方向。与传统的电、热指标相比，利用 GPS 数据的方式，能更直接观测到大地震发生前出现的地表中、低频地形形变，具有较好的客观性和稳定性[4]。但是，现有大多数研究方法普遍依赖地理学科领域专家的知识和经验，且仅用单个典型震例的 GPS 数据来验证异常检测方法的有效性，使其存在主观性较强、普适性较差等问题[3]。

鞅理论[5]作为现代概率和随机过程的基础，适用于时间序列数据分析场合，已被广泛运用于数据挖掘中的决策优化、异常检测等领域。因此，本文结合数据挖掘的相关知识，运用鞅理论，提出一种基于 GPS 数据的震前短临异常检测算法（Anomaly Detection Algorithm based on GPS data，ADA）。

实验结果表明，ADA 算法所识别的 GPS 数据中异常出现时间与地震发生时间存在显著相关。相比于传统的 $k\sigma$ 准则分析方法、异常检测模型 ARIMA[6]、单类别支持向量机 OCSVM[7]，以及基于两阶段聚类的异常检测算法 TSOD[8]等，ADA 算法能够更直观、准确地反映震前 GPS 数据中出现的异常，可为地震预警减灾提供有效手段。

基金项目：福建省中青年教师教育科研项目（JAT190142）；福建省自然科学基金（2019J05048）。

作者简介：李 南（1987-），男，硕士，讲师，主要研究方向：数据挖掘；林莉莉（1985-），女，博士，讲师，主要研究方向：数据挖掘。
通讯作者：林莉莉 Email：binbanbinban@163.com

收稿日期：2022-01-05

1 研究方法

本文短临异常检测算法包括:数据预处理、特征提取以及异常检测 3 部分内容。

1.1 数据预处理

由于数据采集设备、传输线路故障等原因,各 GPS 台站的原始数据存在部分数据缺失的情况。另外,GPS 台站每日坐标包括东西、北南和垂直 3 个方向的数据,但垂直向数据通常误差较大[3]。因此,本文仅针对各 GPS 台站东西向和北南向的坐标数据进行处理。

首先,采用二阶多项式拟合方法,依次对东西向和北南向上的 GPS 数据进行缺失值填补。

给定某个台站特定方向上的一组 GPS 数据 (x_i,t_i) ,$i=1,2,\cdots,num$。其中,num 表示总天数;x_i 表示某天;t_i 表示第 x_i 天特定方向上的 GPS 坐标。通过拟合函数 $y(x,W)=w_0+w_1x+w_2x^2$,求解出损失函数 $E(W)=\frac{1}{2}\sum_{n=1}^{num}(y(x_n,W)-t_n)^2$ 最小化时的权重 W。当第 x' 天的数据缺失时,则使用预测值 $y(x',W)$ 进行填补。

1.2 特征提取

当同一个震例涉及多个 GPS 台站,且不同台站之间 GPS 数据出现异常的时间和强度存在较大差异时,会导致异常检测结果出现较大偏差。为了弥补以上不足,本文基于同一震例的所有相关 GPS 台站数据,采用二阶多项式拟合方法估算相应震例震中位置的每日坐标。震中坐标估算过程如算法 1 所示。若某一震例只涉及一个台站,则直接使用该台站数据即可。

算法 1 使用所有相关 GPS 台站的数据,估算震中位置的每日坐标(以东西向为例)。

输入 震中位置的经度 x',相关的 num 个 GPS 台站的经度 $x_1,x_2,\cdots,x_{num}$,相邻两日各台站东西向坐标偏移量 $\Delta x_1,\Delta x_2,\cdots,\Delta x_n$。

输出:震中东西向的每日坐标。

步骤 1 针对各台站每日坐标位移数据 $(x_i,\Delta x_i)$,$i=1,2,\cdots,num$,求解出使得二项式拟合函数的损失函数最小化时的权重 W。

步骤 2 基于拟合函数 $y(x,W)$,输入震中位置的经度 x',获得估算的偏移量 $y(x',W)$。

步骤 3 根据 x' 和 $y(x',W)$,得到预估的震中东西向的坐标 $x'+y(x',W)$。

为了降低 GPS 数据中白噪声、高斯噪声等对检测结果的影响,在运用算法 1 获得震中位置各方向的时序坐标数据后,使用滑动窗口技术对数据进行降噪处理。固定大小的滑动窗口内样本数据斜率的变化,不仅能有效刻画数据在长趋势变化下的短期特征,而且对噪声具有一定的鲁棒性[9]。因此,本文使用 GPS 台站东西向和北南向坐标数据的斜率变化范围来提取震中每日的综合特征。特征提取过程如算法 2 所示。

算法 2 根据震中东西向、北南向的每日坐标(见算法 1),提取震中的每日综合特征。

输入 震中东西向、北南向的每日坐标 $E(t)$、$N(t)$,滑动窗口大小 $window\_size$。

输出 震中第 t 天的综合特征 $V(t)$。

步骤 1 使用线性回归算法,计算第 t 天前 $window\_size$ 天内东西向、北南向坐标:

$E(i):t-window\_size\leqslant i\leqslant t$ 以及 $N(i):t-window\_size\leqslant i\leqslant t$ 的斜率,记为 $S_E(t)$、$S_N(t)$。

步骤 2 计算滑动窗口内,斜率 $S_E(t)$、$S_N(t)$ 的变化范围:

$$R_E(t)=\max\{S_E(i):t-window\_size\leqslant i\leqslant t\}-\min\{S_E(i):t-window\_size\leqslant i\leqslant t\}$$

$$R_N(t)=\max\{S_N(i):t-window\_size\leqslant i\leqslant t\}-\min\{S_N(i):t-window\_size\leqslant i\leqslant t\}$$

步骤 3 计算震中第 t 天的综合特征值:

$$V_t=\sqrt{\frac{R_E(t)^2+R_N(t)^2}{2}}$$

1.3 异常检测

地震的发生通常需要一定时间的能量累积,本文基于提取到的某震中综合特征值,评估该震中第 t 天的短临异常程度,并在此基础上利用鞅理论评估震中在某连续时间段内短临异常程度。

设:C_{t-1} 为前 $t-1$ 天综合特征 $V=\{V_1,V_2,\cdots,V_{t-1}\}$ 的均值(即中心值),即:

$$C_{t-1}=\frac{1}{t-1}\sum_{i=1}^{t-1}V_i \tag{1}$$

D_t 为 V_t 相对于 C_{t-1} 的偏移程度,即:

$$D_t=\|C_{t-1}-V_t\| \tag{2}$$

其中,$\|\cdot\|$ 表示欧式距离。

根据公式(2)得到的偏移程度,进一步计算 V_t 和 $\{V_1,V_2,\cdots,V_{t-1}\}$ 之间的相异度值 S_t,即:

$$S_t=\frac{fun(j\mid D_j>D_t)+rand\cdot fun(j\mid D_j=D_t)}{t},\quad j=1,2,\cdots,t \tag{3}$$

其中，$rand$ 是一个 (0,1] 之间的随机数，$fun()$ 是一个函数，返回满足指定条件数据的数量。

如：$fun(j \mid D_j > D_t)$ 表示在 $V = \{V_1, V_2, \cdots, V_t\}$ 中 $D_t < D_j, j = 1, 2, \cdots, t$ 的数据数量。

从公式(3)可以看出，$S_t \in (0,1]$。根据同一分布中各样本差异最小化原则[10]，S_t 越小 V_t 就越远离前 $t-1$ 天数据的中心值 C_{t-1}，则 V_t 和 $\{V_1, V_2, \cdots, V_{t-1}\}$ 之间越不相似，表明震中第 t 天的短临异常程度越高。

鞅理论适合于刻画时间序列数据的连续变化情况，使用统计量幂鞅值，可对持续一段时间内数据的异常程度进行量化。幂鞅值越高，越倾向于拒绝接受数据序列分布稳定的假设。本文采用鞅理论对数据 $\{S_1, S_2, \cdots, S_t\}$ 的分布情况进行量化分析，得到 t 天内 $\{S_1, S_2, \cdots, S_t\}$ 的幂鞅值 M_t。

$$M_t = \prod_{i=1}^{t} (\varepsilon \cdot S_i^{\varepsilon-1}) = M_{t-1}(\varepsilon \cdot S_t^{\varepsilon-1}), M_0 = 1 \quad (4)$$

其中，S_t 为 V_t 和 $\{V_1, V_2, \cdots, V_{t-1}\}$ 之间的相异度值，根据文献[1]的推论 ε 取值 0.82。

从公式(4)可见，幂鞅值 M_t 值越大，说明 t 天内频繁出现 S 值较小的情况，暗示 t 天内 GPS 数据频繁出现异常的程度越高。为了避免公式(4)中幂鞅值 M_t 值无限增大，需引入一个停止参数 h 作为 M_t 的阈值。此外，本文还引入一个稳定参数 $stable\_day$，从第 $stable\_day + 1$ 天开始计算幂鞅值，以避免过短的时间序列数据对分析结果造成误差。异常检测算法具体过程如算法 3 所示。

算法 3　使用某震中的综合特征序列，计算该震中 t 天内的幂鞅值 M_t。

输入　某震中的综合特征 $V = \{V_1, V_2, \cdots, V_t\}$、停止参数 h、稳定参数 $stable\_day$。

输出　某震中 t 天内的幂鞅值 M_t。

步骤 1　设：$t = stable\_day + 1$。

步骤 2　根据 $V = \{V_1, V_2, \cdots, V_{t-1}\}$，采用公式(1)、(2)分别计算，得到 C_{t-1} 和 D_t。

步骤 3　根据 C_{t-1} 和 D_t，采用公式(3)计算 S_t。

步骤 4　根据 S_t，采用公式(4)计算 M_t。

步骤 5　如果 $M_t \leqslant h$，则 $t = t + 1$，重新执行步骤 2 - 5，否则将第 $t + 1$ 天作为第 1 天，重新执行本算法。

1.4　ADA 算法流程

基于算法 1、算法 2 和算法 3，则 ADA 算法具体流程如图 1 所示。

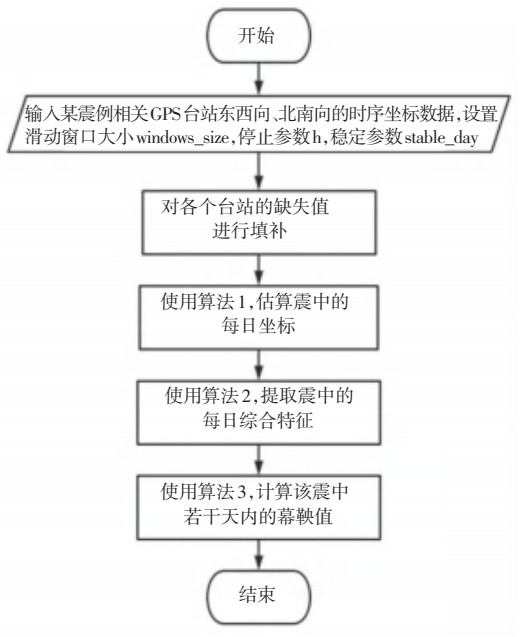


图 1　ADA 算法流程图

Fig. 1　Flow chart of ADA algorithm

2　实验设计与分析

2.1　实验数据

本文研究对象为 2001～2010 年间，北美发生的震源深度小于 60 km 且震级大于 6.0 级的地震。GPS 台站时序坐标数据来自 Nevada Geodetic Laboratory 提供的数据共享服务网站（http://geodesy.unr.edu/）。选择的 GPS 台站，需处于受相应地震孕育影响的范围[11]（即震中半径 $R = 10^{0.43Mag}$ 之内，Mag 表示地震震级，R 的单位是 km）。实验选择位于影响范围内，最靠近震中的 10 个 GPS 台站。由于地震孕育过程通常在地震前 1～30 天开始[1]，因此所使用台站的数据从地震发生前 180 天开始，到后 30 天结束。为了确保有足够的台站以供分析，单个台站在这段时间内最多允许 5% 的数据缺失。从平台获得的数据是初步处理后的 GPS 台站每日坐标分别是东西向、北南向和垂直向。文献[3]中证实，GPS 台站的时序坐标数据在垂直方向的测量误差远大于水平方向。因此，实验中只选用东西向、北南向的每日坐标作为研究数据，以保证分析结果的可靠性。

综合考虑 GPS 台站位置和数据完整性，本文最终采用的震例数据见表 1，相关信息来自美国地质调查局网站（https://earthquake.usgs.gov/）。

表 1 震例数据

Tab. 1 Earthquake data

Index	Time	Lat.(°N)	Long.(°E)	Depth(km)	Magnitude
1	2003-12-22	35.7	-121.1	8.4	6.5
2	2009-07-03	25.1	-109.8	10	6
3	2009-08-03	29.0	-112.9	10	6.9
4	2010-04-04	32.3	-115.3	9.9	7.2
5	2012-04-12	28.7	-113.1	13	7
6	2012-12-14	31.1	-119.7	13	6.3
7	2018-01-19	26.7	-111.1	10	6.3
8	2019-07-06	35.8	-117.6	8	7.1

2.2 对比算法与参数设置

为了验证基于 GPS 数据的短临异常检测算法的有效性,将本文的 ADA 算法与传统的 $k\sigma$ 准则分析方法、异常检测模型 ARIMA、单类别支持向量机 OCSVM 以及基于两阶段聚类的异常检测算法 TSOD 进行性能对比。

(1) $k\sigma$ 准则分析方法:在 $k\sigma$ 准则中,以 $\bar{x}$ 表示观测结果的平均值, σ 表示观测结果的标准差,如果数据超出 $[\bar{x}-k\sigma,\bar{x}+k\sigma]$ 的范围,则认为出现异常。依据文献[16],本文将准则中的 k 值设置为 2。

(2)ARIMA 模型:利用差分整合移动平均自回归模型,得到一个预测值,通过预测值与实际值的误差大小来判断异常位置。本文 ARIMA 模型中的信息准则函数选用贝叶斯信息准则。

(3)OCSVM:单类别支持向量机,将异常检测视为特殊的分类问题。在训练过程中,只有一类数据,首先得到可以代表这部分数据的模型。在检测过程中,判断给定样本是否属于此类别。本文 OCSVM 算法中的核函数选用高斯核函数。

(4)TSOD 算法:是基于两阶段聚类的多变量时间序列异常检测算法。第一次聚类在各个变量上筛选初始异常时间,第二次聚类结合所有变量进行异常定位,以降低误检率。TSOD 算法中第一次聚类使用基于混合高斯模型的 EM 算法,第二次聚类使用以全连接方式度量的层次聚类方法。

对比实验和参数优化实验中,使用的数据来自表 1 中震级最大的 2010-04-04 地震, $k\sigma$ 准则、ARIMA 模型、OCSVM 以及 TSOD 算法仅使用距离地震震中最近的编号为 P500 的单个 GPS 台站(Latitude:32.69°N, Longitude: - 115.30°E)数据。本文 ADA 算法则涉及多个台站的 GPS 数据,停止参数 h 设置为 2 000,窗口大小 *window_size* 设置为 7,稳定参数 *stable_day* 设置为 5。

2.3 性能对比分析

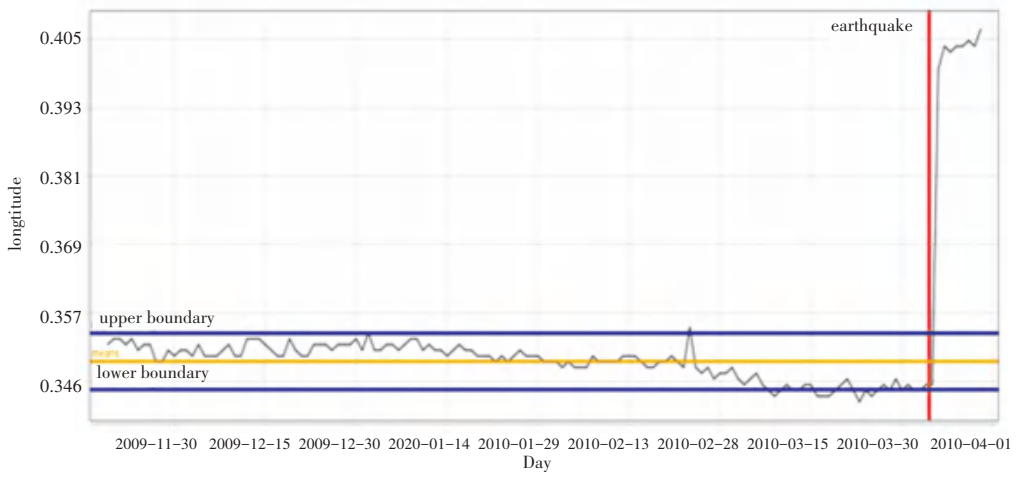
图 2~图 4、表 2 和图 5 分别给出了 $k\sigma$ 准则、ARIMA 模型、OCSVM、TSOD 算法以及 ADA 算法的运行结果。

为了更直观地观察 GPS 数据的变化,图 2 中的 GPS 数据中只使用数值的小数部分而忽略相同的整数部分。图 2(a)~2(c)中中间的水平线表示 $\bar{x}$, 上下水平线分别表示 $\bar{x}-k\sigma$ 与 $\bar{x}+k\sigma$, 垂直线表示地震发生时间。

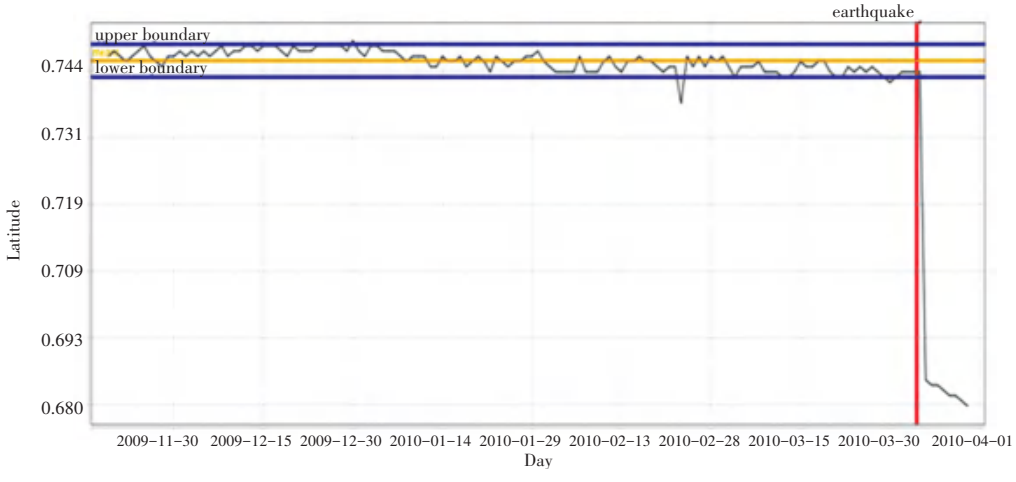
图 2(a)~2(c)中震前 GPS 数据在 3 个方向上均数次超过 $[\bar{x}-k\sigma,\bar{x}+k\sigma]$ 的警戒范围,最早的一次出现在地震发生的前 3 个月;震后东西向与北南向的坐标均发生了突变。但据文献记载,孕震活动通常发生在地震发生前的 30 天左右[1],这说明使用 $k\sigma$ 准则不能完全准确地检测到 GPS 数据中与地震相关的短临异常的存在,容易出现误报情况。此外,在不同方向上超过警戒范围的时间也不完全一致,这进一步增加了结果分析的难度。图 2(c)中 P500 台站垂直向的 GPS 数据波动较大,数值频繁超过警戒范围且无明显规律,这源于 GPS 数据在垂直向的测量误差较大。

从图 3(a)~图 3(b)可见,震前 45 天左右,P500 台站东西向和北南向的坐标出现了 ARIMA 模型的预测值与真实值误差较大(即异常)的情况,但随着地震的临近,误差并没有继续保持在较高的水平。因此,不能完全确定此次异常是否与地震相关,也可能与噪声有关。在垂直向上,ARIMA 模型的预测值和真实值之间的误差并没有显现出任何规律,这同 $k\sigma$ 准则的分析结果相一致。

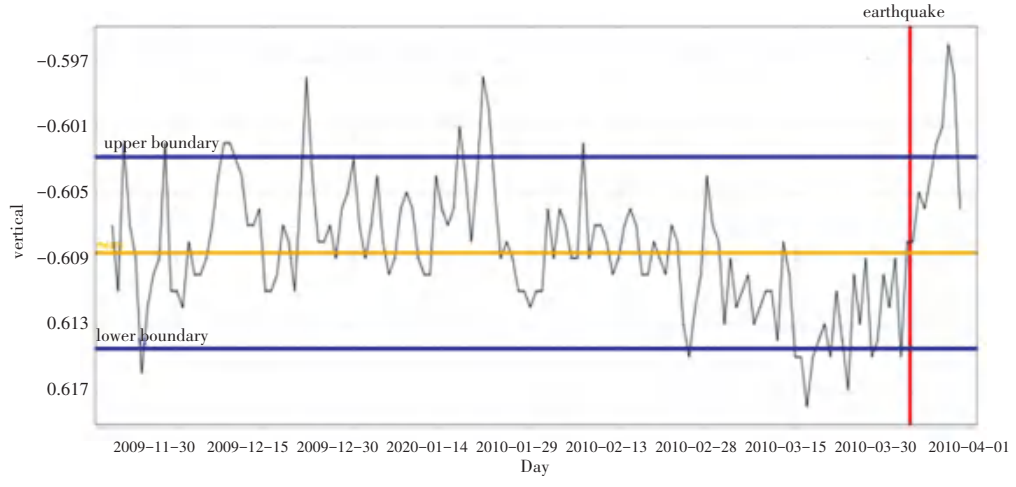
图 4(a)~图 4(c)中,横坐标表示时间,纵坐标表示当天给定方向的坐标值,在 OCSVM 算法下的类别。1 表示正常,-1 表示异常。从图 4(a)可看出,P500 台站东西向异常最早出现在震前 45 天左右,在一周后断断续续出现并延续到震前。从图 4(b)~图 4(c)可看出,OCSVM 算法在北南向和垂直向上,震前没有发现明显的异常显现规律,并出现多次误报。异常的不持续以及 3 个方向上异常出现时间的不统一都增加了结果分析的难度。



（a）东西向坐标值分析结果



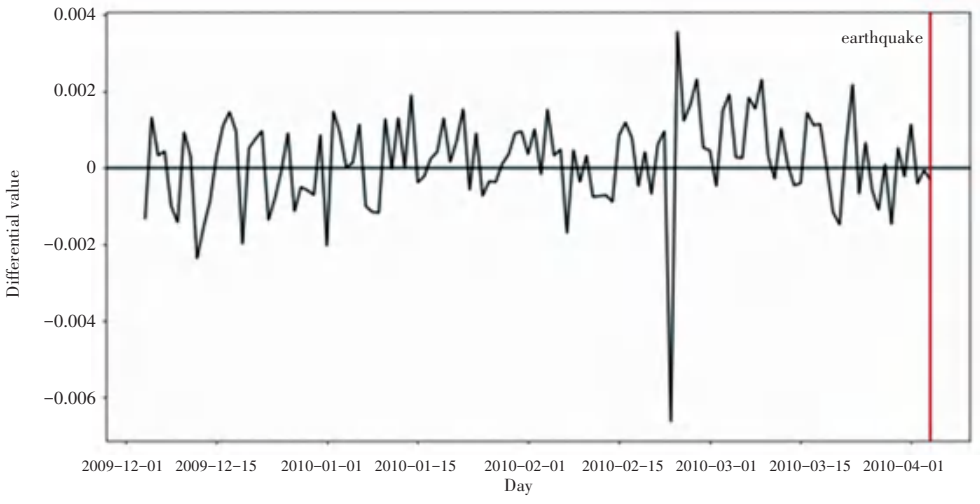
（b）北南向坐标值分析结果



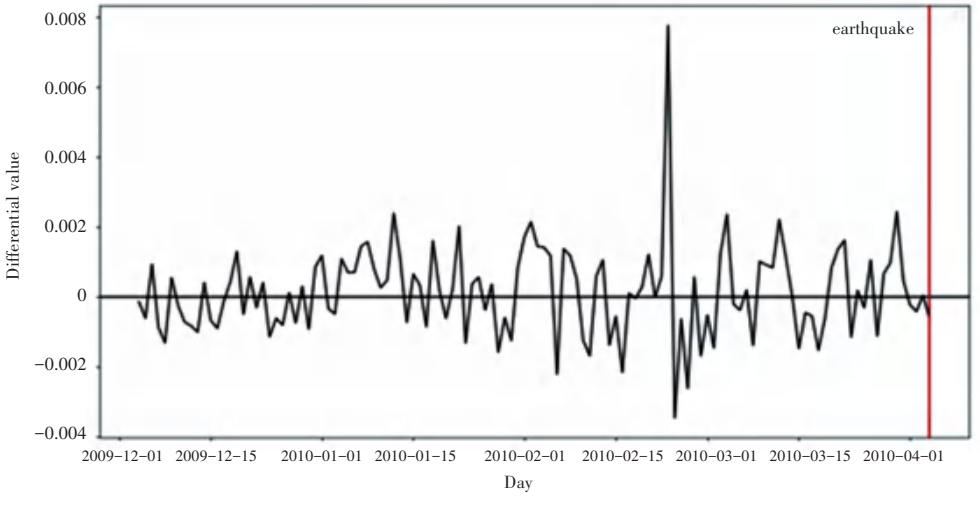
（c）垂直向坐标值分析结果

图 2　P500 台站各向 $k\sigma$ 准则分析结果

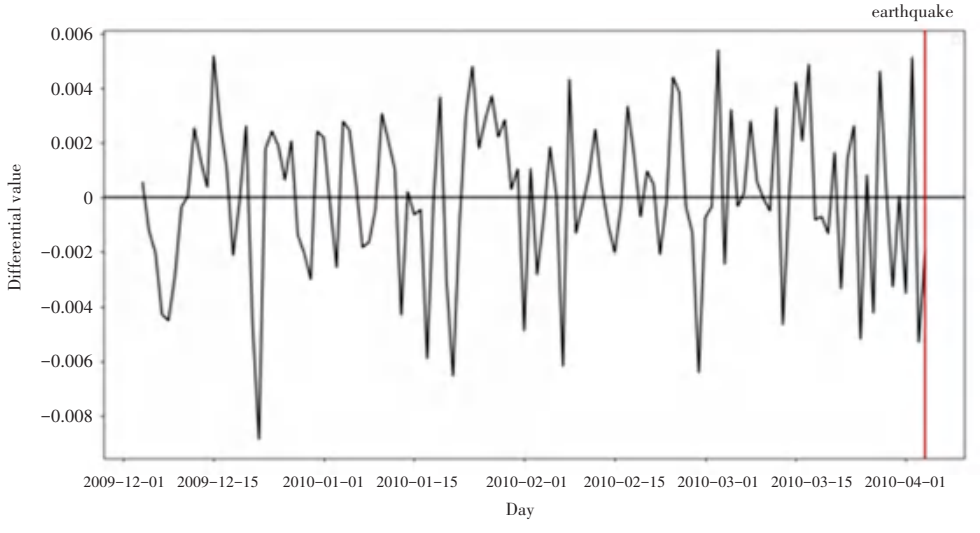
Fig. 2　Analysis result of $k\sigma$ method on P500 station



(a) 东西向预测值与真实值误差



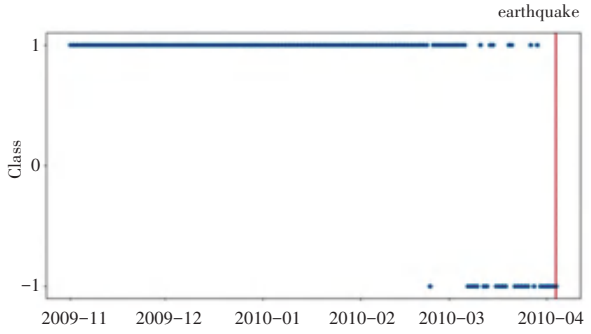
(b) 北南向预测值与真实值误差



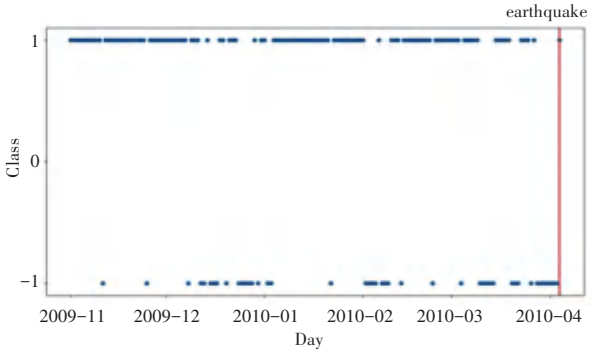
(c) 垂直向预测值与真实值误差

图 3 P500 台站的 ARIMA 模型分析结果

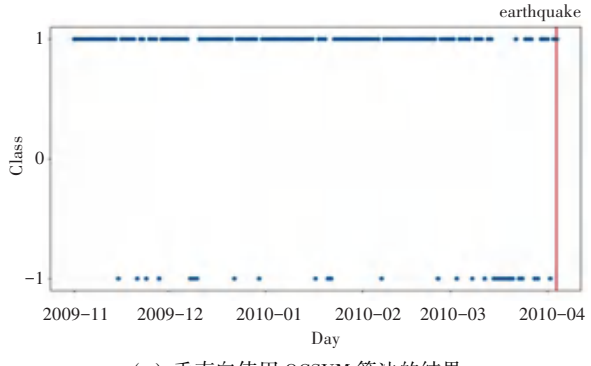
Fig. 3 Analysis result of ARIMA model on P500 station



（a）东西向使用 OCSVM 算法的结果



（b）北南向使用 OCSVM 算法的结果



（c）垂直向使用 OCSVM 算法的结果

图 4　P500 台站的 OCSVM 算法分析结果

Fig. 4　Analysis result of OCSVM algorithm on P500 station

表 2　P500 台站使用 TSOD 算法分析结果

Tab. 2　Analysis result of TSOD algorithm on P500 station

Stage	Corresponding time when anomaly is detected
东西向	2010−02−04,2010−02−22,2010−02−25, 2010−02−28,2010−03−01
北南向	2010−02−23
垂直向	2009−12−22,2010−01−21
综合 3 个方向	2009−12−22,2010−01−28,2010−02−25

从表 2 可以看出，TSOD 算法在 P500 台站 GPS 数据上，最终检测出 3 次异常。其中，距离地震发生最近的异常是在震前 40 天左右（2010−02−25），并出现了两次明显的误报（在 2009−12−22 以及 2010−01−28）。

图 5 给出了 2010−04−04 地震 ADA 算法运行结果（即幂鞅值的变化趋势）。从图 5 可明显看出，在震前绝大部分时间，幂鞅值始终保持在一个相对较小的区间内。由于大地震震前能量是一个累积的过程，幂鞅值从地震前较短的一段时间（约 1 个星期）开始缓慢增加，说明 GPS 数据开始出现异常，暗示震前局部应力场开始调整。地震后各个台站的坐标发生了较大变化，因此幂鞅值的波峰是在地震后出现，且在地震后迅速超过预设的阈值 h。这说明 ADA 算法对 2010−04−04 地震的异常检测是有效的，且比 4 种对比算法能更直观地反映出震前短临异常，不易出现误报的情况。

2.4　参数优化分析

2.4.1　稳定参数分析

为了分析稳定参数 *stable_day* 对本文方法性能的影响，在 2010−04−04 地震上分别将 *stable_day* 设置为 5、7 和 9 进行实验，幂鞅值的变化趋势如图 6 所示。

图 5　2010−04−04 地震的 ADA 算法运行结果

Fig. 5　Analysis result of ADA algorithm on 2010−04−04 earthquake

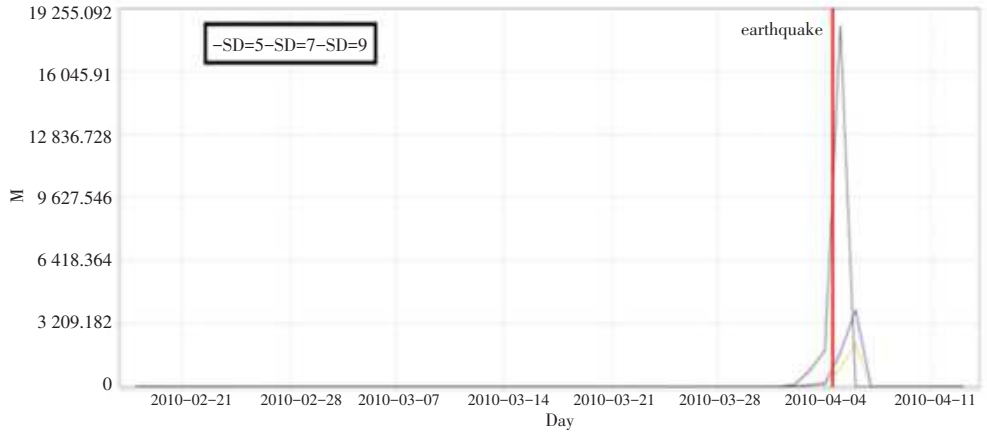


图 6 基于不同稳定参数的 ADA 算法结果

Fig. 6 Comparison of result with different stable_day

从图 6 可看出,不同的 *stable_day* 参数值,并不会对检测结果造成太大影响。不同取值下,幂鞅值的变化趋势均表现为在地震前较短的一段时间内开始增加,并在地震后的一段时间内达到波峰。这是由于当地壳运动相对稳定时,前 *stable_day* 天的 GPS 数据并不会发生太大变化,因而对结果的影响不大,不同取值下幂鞅值波峰出现的时间仅差距 1~3 天。但当参数为 5 时,幂鞅值最早出现增加的趋势。据此,实验中将稳定参数 *stable_day* 设置为 5。

2.4.2 平滑窗口分析

为了分析平滑窗口大小 *window_size* 对本文方法性能的影响,在 2010-04-04 地震上分别将 *window_size* 设置为 5、7 和 10 进行了实验,幂鞅值的变化趋势如图 7 所示。

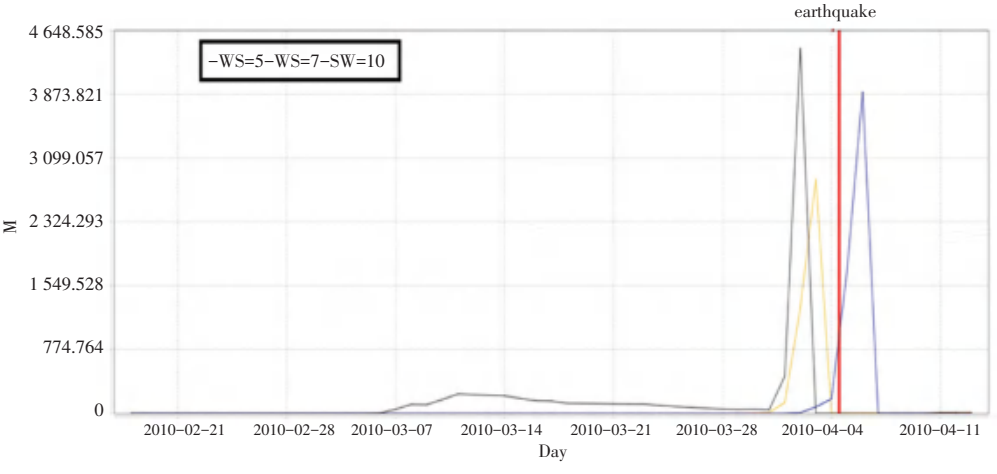


图 7 基于不同平滑窗口的 ADA 算法结果

Fig. 7 Comparison of result with different window_size

当平滑窗口 *window_size* 的取值较小时,在特征提取阶段,计算第 t 天的综合特征值所需要的样本数就越少,因此更容易受到单个样本的影响,对异常的检测也更敏感。从图 7 可看出,对比 *window_size* = 7 和 *window_size* = 10,当 *window_size* = 5 时,在地震发生前一个月就出现了幂鞅值缓慢增加的趋势,相应幂鞅值的波峰也在地震发生前最早出现。而对于 *window_size* = 7 和 *window_size* = 10,幂鞅值开始增加和波峰出现的时间并不存在显著差别。因此,为了提高算法的鲁棒性,实验中将 *window_size* 设置为 7 更为合理。

2.4.3 停止参数分析

为了分析停止参数 h 对本文方法性能的影响,在 2009-07-02 地震上分别将 h 设置为 500、1 000 和 2 000 进行了实验,幂鞅值的变化趋势如图 8 所示。

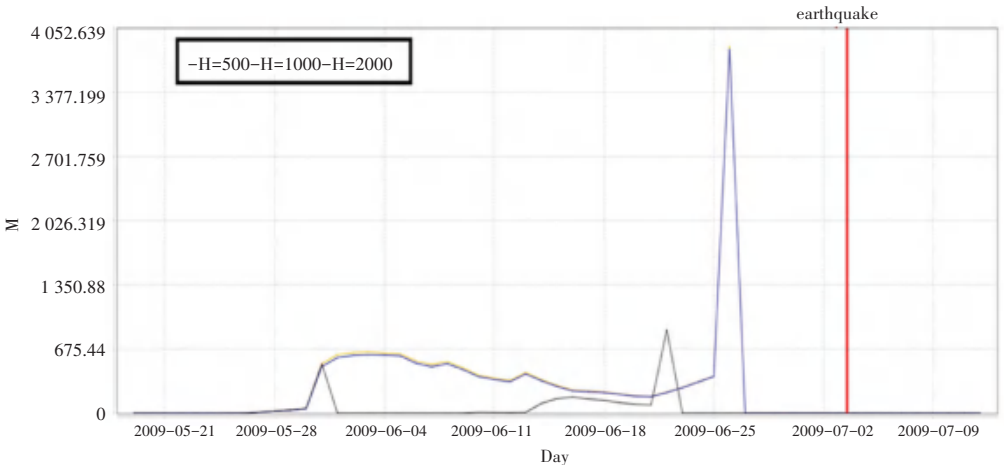


图 8　基于不同停止参数的 ADA 算法结果

Fig. 8　Comparison of result with different *h*

从图 8 可看出，对于 2009-07-02 地震，3 种停止参数设置下，幂鞅值均在地震发生前较短一段时间内显著提高，这与文献[1]中的结论相一致，即孕震活动最早在震前 30 天左右开始，并在震前几天内表现最为活跃。值得注意的是，当参数 h 设置为较小(500)时，幂鞅值的波峰多次出现，会导致异常的误报。当检测到 GPS 数据出现震前异常时，幂鞅值仅经过一天就从小于 1 000 增大到 2 000 以上。因此，参数 h = 1 000 和 h = 2 000 的幂鞅值曲线几乎重合。由于较大的停止参数会减少预警时间，降低误报可能，实验中将 h 设置为 2 000。

2.5　显著性检验

为了验证 ADA 算法所识别的 GPS 数据中存在的短临异常与对应地震之间存在关联，本文使用 Molchan 图表法[12]，在表 1 所示的 8 个震例上进行了统计显著性检验，结果如图 9 所示。

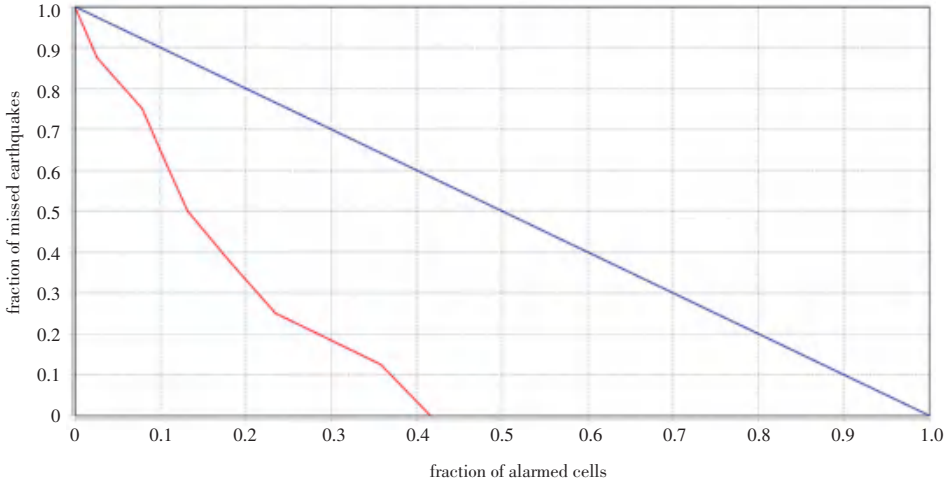


图 9　8 个震例的 Molchan 图表分析结果

Fig. 9　Analysis result of ADA algorithm oneight earthquakes by Molchan error diagram

图 9 中，横坐标表示时间占有率，纵坐标表示相应的漏报率，使用的方法相比于随机预测的优劣程度以曲线与图表边界线所包围的面积来衡量，面积越小则说明预测效果越好。若测试的结果接近于图 9 所示的对角线，则表示预测方法无统计显著性。实验中，若幂鞅值波峰出现在第 t 天，那么将前 *front* 天和后 *rear* 天作为变量，即以 $[t-front, t+rear]$ 作为预警时间范围。若地震发生在此时间段内，则表示预警成功。通过调整 *front* 和 *rear* 的取值以绘制图表。从图 9 中可以看出，ADA 算法的时间占有率-漏报率曲线远在对角线之下，说明所识别的短临异常与对应地震之间存在显著性关联。

3　结束语

震前短临异常检测是地震预警减灾的关键。本文提出的 ADA 算法能够弥补现有方法存在的主观

性较强、普适性较差的问题。8 个震例的实验结果证实了 ADA 算法检测到的短临异常与地震之间存在显著相关。另外,本文也对算法的参数进行了优化分析。

然而,地震监测预警是一项复杂的任务,会涉及与孕震相关的岩石圈-盖层-大气层-电离圈层等多个数据源。因此,如何结合这些异源数据来进行异常检测是下一步的研究方向。

参考文献

[1] KONG X, LI N, LIN L, et al. Relationship of Stress Changes and Anomalies in OLR Data of the Wenchuan and Lushan Earthquakes [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2018, 1(11): 2966-2976.

[2] ABBASI A R, SHAH M, AHMED A, et al. Possible Ionospheric Anomalies Associated with the 2009 Mw 6.4 Taiwan Earthquake from DEMETER and GNSS TEC [J]. Acta Geodaetica et Geophysica, 2021, 56(1): 77-91.

[3] WANG T, ZHANG J, KATO T, et al. Assessing the PotentiaI Improvement in Short - term Earthquakes Forcasts from Incoroiration of GPS Data [J]. Geophysical Research Letters, 2013, 40(11): 1-5.

[4] COLOMBELLI S, ALLEN RM, ZOLLO A. Application of real-time GPS to Earthquake Early Warning in Subduction and Strike-slip Environments [J]. Journal of Geophysical Research: Solid Earth, 2013, 118(7): 3448-3461.

[5] HO S S, WECHSLER H. A Martingale Framework for Detecting Changes in Data Streams by Testing Exchangeability [J]. IEEE transactions on pattern analysis and machine intelligence. 2010, 32(12), 2113-2127.

[6] 杨立宁,李艳婷. 基于 SVD 和 ARIMA 的时空序列分解与预测[J]. 计算机工程, 2021, 47(3): 53-61.

[7] HU M, JI Z, YAN K, et al. Detecting Anomalies in Time Series Data via a Meta - Feature Based Approach [J]. IEEE Access, 2018, 6: 27760-27776.

[8] 吴英友, 胡刚义, 唐静, 等. 基于两阶段聚类的设备状态异常检测方法[J]. 船舰科学技术, 2021, 43(8): 163-168.

[9] COLOMBELLI S, ALLEN R M, ZOLLO A. Application of real-time GPS to Earthquake Early Warning in Subduction and Strike-slip Environments [J]. Journal of Geophysical Research: Solid Earth, 2013, 118(7): 3448-3461.

[10] KANUNGO T, MOUNT D M, NETANYAHU N S, et al. An Efficient K - means Clustering Algorithm: Analysis and Implementation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24: 881-892.

[11] AKHOONDZADEH M, SANTIS A D, MARCHETTI D, et al. Anomalous Seismo-LAI Variations Potentially Associated with the 2017 Mw = 7.3 Sarpol-e Zahab (Iran) Earthquake from Swarm Satellites [J]. Advances in Space Research, 2019, 64(1): 143-158.

[12] PENG H, KATSUMI H, JIANCANG Z, et al. Evaluation of ULF Seimo - magnetic Phenomena in Kakioka, Japan by Using Molchan's Error Diagram [J]. Geophysical Journal International, 2017, 208(1): 482-490.

(上接第 58 页)

[2] ZHANG W, WANG C, ZHANG Y, et al. Credit risk evaluation model with textual features from loan descriptions for P2P lending [J]. Electronic Commerce Research and Applications, 2020, 42: 100989.

[3] 罗志芳.银行在信用证结算中的风险及其防范[J].中国金融, 1998(11): 26-27.

[4] KOTSIANTIS S B, ZAHARAKIS I, PINTELAS P. Supervised machine learning: A review of classification techniques [J]. Emerging artificial intelligence applications in computer engineering, 2007, 160(1): 3-24.

[5] KOKKINAKI A I. On atypical database transactions: identification of probable frauds using machine learning for user profiling[C]// Proceedings 1997 IEEE Knowledge and Data Engineering Exchange Workshop. IEEE, 1997: 107-113.

[6] 严华,胡孟梁,蔡瑞英. 防止信用卡欺诈的系统设计[J]. 微计算机信息, 2006(12): 63-65.

[7] 何杨,李洪心. 基于模糊二范数二次曲面支持向量机的信用评分研究[J]. 统计与决策, 2018, 34(5): 66-70.

[8] 刘岚,王霞,林红旭,等. 基于混合 BP 神经网络算法的信用卡消费行为风险预测[J]. 科技管理研究, 2011, 31(17): 206-210.

[9] DORNADULA V N, GEETHA S. Credit card fraud detection using machine learning algorithms[J]. Procedia computer science, 2019, 165: 631-641.

[10] 陈荣荣,詹国华,李志华. 基于 XGBoost 算法模型的信用卡交易欺诈预测研究[J]. 计算机应用研究, 2020, 37(S1): 111-112, 115.

[11] 琚春华,陈冠宇,鲍福光. 基于 KNN-Smote-LSTM 的消费金融风险检测模型——以信用卡欺诈检测为例[J]. 系统科学与数学, 2021, 41(2): 481-498.

[12] YADAV A, VISHWAKARMA D K. Sentiment analysis using deep learning architectures: a review [J]. Artificial Intelligence Review, 2020, 53(6): 4335-4385.

[13] KIM J Y, CHO S B. Towards Repayment Prediction in Peer-to-Peer Social Lending Using Deep Learning [J]. Mathematics, 2019, 7(11): 1041.

[14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[15] ZHANG W, WANG C, ZHANG Y, et al. Credit risk evaluation model with textual features from loan descriptions for P2P lending [J]. Electronic Commerce Research and Applications, 2020, 42: 100989.

[16] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks [J]. science, 2006, 313(5786): 504-507.

[17] 武建奇,何姝. 互联网贷款欺诈的形成机理与量化评估[J]. 技术经济与管理研究, 2020(11): 80-84.