第 38 卷　第 2 期(270~279)　　中 国 地 震　　Vol. 38　No. 2
2022 年 6 月　　EARTHQUAKE RESEARCH IN CHINA　　Jun. 2022

王晓湘,刘洞天,刘南江,等,2022. 基于 LSTM 的震后通信数据异常检测分析. 中国地震,**38**(2):270~279.

基于 LSTM 的震后通信数据异常检测分析

王晓湘[1)]　刘洞天[1)]　刘南江[2)]　丁一[2)]　姜立新[3)]

1)北京邮电大学,信息与通信工程学院,北京　100876
2)应急管理部国家减灾中心,北京　100124
3)中国地震台网中心,北京　100045

摘要　震后通信数据会发生异常变化,通过对通信数据异常进行分析,有助于提供有效的灾情数据以及更好地了解震后产生的影响,进而有效地为抗震救灾提供辅助支持。本文基于 LSTM 对震后的通信数据异常进行分析,研究内容主要包括通信数据流预处理、基于 LSTM 的异常检测模型以及数据分布变化检测模型。结果表明,本文模型能够对通信数据的异常变化进行识别,为后续的灾情分析提供数据。

关键词: 通信数据　数据处理　异常检测　数据分布变化

[文章编号] 1001-4683(2022)02-0270-10　[中图分类号] P315　[文献标识码] A

0　引言

我国是地震灾害最为严重的国家之一,地震会造成次生灾害,破坏当地的道路和交通,造成人员伤亡和财产损失。

震后的灾情分析对于应急救灾尤为关键,而灾情分析需要获取震后的灾情数据。目前,震后灾情数据的获取通常利用实地调查方式,但由于交通被破坏,实地调查往往耗费大量的时间和人力,延误抗震救灾的进度。

随着科学技术的发展,智能设备持有量日益增加,对通信数据的有效分析可以解决许多问题。地震发生后,受设备和电力的影响,通信数据会产生不同程度的变化。2017 年九寨沟 7.0 级地震后,就导致了 273 个基站退服,因此出现话务拥塞的情况。通过通信数据异常检测,分析震后通信数据出现的异常(白惠文,2019),可为灾情分析提供数据,进而为抗震救灾提供辅助支持,且相较传统的灾情数据获取更加快速和有效。

众多学者对震后灾情数据的获取进行了相关研究。刘在涛等(2011)以震级、震区人口密度等地震有关信息为依据,建立起地震应急响应等级的分类判别规则,从而提供了一种简便、快速的地震应急响应等级初判方法; 陈琳等(2011)通过对福建网格离散化,利用地震动峰值加速度对全省的网格进行烈度评估,提出了一种基于地震动峰值加速度的地震灾害程

[收稿日期] 2021-03-22　[修定日期] 2022-03-29

[项目类别] 国家重点研发计划(2018YFC1504500、2018YFC1504502)、国家自然科学基金青年基金项目(41807505)共同资助

[作者简介] 王晓湘,女,1969 年生,教授,主要从事数据处理、宽带移动通信等研究。E-mail:cpwang@bupt.edu.cn
姜立新,通讯作者,男,1966 年生,研究员,主要从事震害预测、地震应急技术等研究。E-mail:jlx@seis.ac.cn

度评估方法；傅征祥等(2008)通过对北方建筑物进行分类,归纳整理出基于倒塌率计算死亡率的公式,但是通过地理数据得到的震后数据只能够体现出地理层面的破坏情况,可能与实际情况不符。真实情况的灾情数据通常需要震后工作人员实地调查得出(薄涛等,2018)。

为快速获取灾情数据,近年来科研人员利用震后通信数据进行了相关研究。通信数据能够反映出人的活动,数据更为准确,且真实性高(董翔等,2007)。Bengtsson 等(2011)使用手机用户 SIM 卡的位置数据来分析震后海地人口的移动趋势,为后续研究震后人员位置和及时救援提供了帮助；Finazzi(2016)基于智能手机的网络数据来检测地震的发生,其通过在手机中加入加速度传感器,当传感器感受到振动时,将信号上传,通过分析信号便得出相关结论。

基于灾情分析的紧迫性和实地获取灾情数据的限制,本文基于长短期记忆网络 LSTM (Long Short Term Memory)对震后的通信数据进行异常检测,分析震后出现的通信数据异常,为灾情分析提供数据支持。由于地震后基站的破坏和人员的活动,震区的通信数据会发生异常变化,利用用户上传的通信数据,建立数据流异常检测模型,对数据进行异常检测并评估异常等级,得出区域通信数据的异常情况,为灾情分析提供数据支持,辅助抗震救灾。利用该模型在实验数据上进行了应用,实验结果表明,对同一区域内的数据进行分析,使用基于 LSTM 的通信数据流异常检测模型,能够分析出区域内数据的异常程度,进而对应急救灾提供辅助支持。

1　RNN、LSTM 网络概述

1.1　循环神经网络 RNN

循环神经网络 RNN(Recurrent Neural Network)由输入层、隐藏层和输出层组成,相比全连接神经网络,RNN 隐藏层的值与当前输入和之前的隐藏层值有关。RNN 模型的基本结构如图 1 所示,当模型输入 x_t 时,基本单元中隐藏层 s_t 的值取决于 x_t 和 s_{t-1},基于此可得出 RNN 的表达式,即

$$s_t = f(Ux_t + Ws_{t-1}) \tag{1}$$

其中,U、W 表示权重值,f 函数常用 tanh 函数。

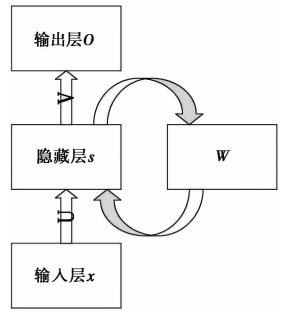

图 1　简易 RNN 模型

相比其他神经网络算法,RNN 能更好地处理序列信息,但在处理长序列信息时会出现梯度消失和梯度爆炸问题。

1.2 长短期记忆网络 LSTM

长短时记忆网络 LSTM 是针对 RNN 网络中存在的无法处理长距离依赖的问题而改进的循环神经网络。在时间序列预测分析的问题上,LSTM 可以通过过去一段时间的数据特征来预测未来的数据特征。

LSTM 网络通过“门”结构使节点可以“记忆”或者“忘记”数据,其主要包含 3 个“门”:忘记门、信息增加门和信息输出门。通过这 3 个“门”,使得每次细胞状态的输入包含上一时刻的输出,而此刻的输入也有节点本身存储的部分信息。因此,LSTM 在较长的序列上有更好的表现。

2 研究方法

考虑到本研究处理的数据为数据流,数据分布易发生变化且高速生成,故设计了基于 LSTM 的通信数据异常检测模型,并在模型中引入了数据分布检测模块。

2.1 通信数据预处理

智能设备上传的通信数据以数据流的形式进行收集,极短时间内就会产生大量数据,因此需要对数据进行快速处理,还需要对数据预处理(王晓青等,2000)。本模型使用了基于邻域粗糙集思想的特征约简方法,该方法通过计算每一个属性对于决策属性的重要性进行排序,根据设定好的重要性阈值对属性进行约简。本文利用欧几里得距离公式计算数据间的距离,当两数据距离小于设定阈值,则两数据为一类,即利用邻域粗糙集约简特征的方法进行处理。

考虑到特征间的相关性,本文计算了特征间的相关系数,根据相关系数阈值,对于高度相关的共性特征,只保留重要性更高的特征(Guo et al,2021)。

同时,为了加快模型的处理速度,对数据集进行分块处理,将每个数据块上的特征重要性进行求和,得出最终的特征重要性,使本文方法在处理高维数据时更加快速。

2.2 基于 LSTM 的通信数据异常检测模型

考虑收集到的通信数据为数据流,以及数据异常检测的基本要求,本文设计了基于 LSTM 的数据流异常检测模型。对通信数据进行预处理后,对时间序列的数据建模进行预测,且设计了差值正态建模来判断数据异常。

构建 LSTM 异常检测模型,主要环节如下:

(1)设定好时间窗口大小 K,并按照时间窗口大小对于数据集进行改造,将时间序列转化为有监督序列,即利用过去 K 个值预测下一时刻的值,原本下一时刻的值作为监督值。

(2)将使用的数据集划分为训练集和测试集,并将数据的格式转换为 LSTM 中需要的格式,即[samples,timesteps,features]。

(3)通过不断尝试,确定模型中使用的参数,包含迭代次数、每次迭代的数据量以及神经元的数量。

(4)建立 LSTM 模型。对数据流中的数据进行预测的模型构建好后,即可对数据进行预测。将数据格式进行相应的反变换,计算监督值与预测值的差,利用当前时刻前 K 时刻数据的预测差值进行正态分布建模,计算当前时刻预测差值概率密度值的倒数,作为异常分数,当异常分数大于设定的阈值时,认为数据异常。

LSTM 数据流异常检测模型算法如下：

输入：数据集 Data，时间窗口 K

输出：异常值集合

(1)将数据集转化为有监督学习问题并进行归一化；

(2)将数据集划分为训练集和测试集，并重构为规定形式；

(3)搭建 LSTM 模型并进行优化；

(4)利用模型对数据进行预测，并根据与真实值的差异来判断数据异常。

2.3　数据分布变化检测

随着时间的推移，数据分布可能会发生变化，导致异常检测模型的性能变差。本文在模型中设计了数据分布变化检测方法，通过比较前后 2 个数据块的数据分布变化进行判断，当模型性能发生显著变化，利用新数据重构异常检测模型。

SVM(Support Vector Machin)是常用的机器学习方法，其能够完成线性和非线性的分类(Minku et al,2012)。SVM 主要思想是寻找特征空间最大化间隔的线性分类超平面，将不同类的样本区分开，同时使得数据点到平面的间隔最大。

本文设计了基于 SVM 模型参数的数据分布检测法，其通过前后数据块学习得到的参数差异，判断数据分布变化的发生。通过对两数据块 M1 和 M2 进行学习，可以得出相应的分类平面向量 $W1$ 和 $W2$，通过计算向量的夹角余弦最大值来衡量是否有模型性能变化的发生，进而对模型进行相应的调整，提高模型的泛化能力。

SVM 参数法算法如下：

输入：数据块 M1，M2，阈值 θ_d

输出：是否发生模型性能变化

(1)将 M1、M2 输入 SVM 模型，计算出前后两窗口数据的分类平面向量 W_1 和 W_2；

(2)计算 $\cos\theta=\max\cos(W_1,W_2)$；

(3)如果 $\cos\theta<\theta_d$，算法收敛，返回正确；

(4)否则，返回错误。

由此，得出通信数据流异常检测模型整体的框架，如图 2 所示。利用设计的 LSTM 模型检测数据异常，得出异常分数；不断对数据流的数据分布进行检测，当模型性能产生明显变化时，利用新数据重新训练模型。

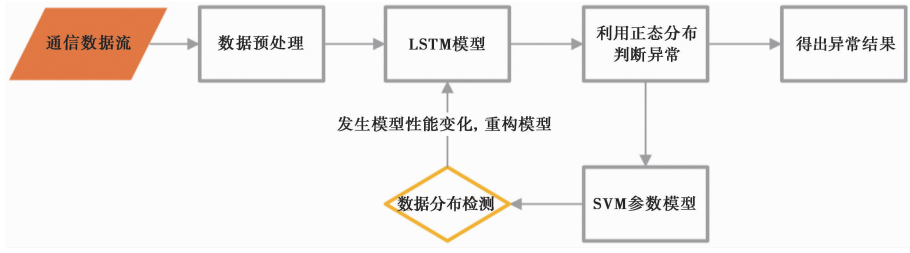

图 2　基于通信数据的异常检测模型框架

3 分析和讨论

3.1 实验相关说明

本研究实验环境为:Intel Core i5 CPU,4GB 内存,Windows10 操作系统,Python3.7,使用 PyCharm 作为开发工具。实验通信数据为基于 2017 年九寨沟 7.0 级地震发生前后数据变化规律构造的通信数据集,数据包含地震前后共 120min、10000 个样本的数据。

使用 NAB 数据集分析 LSTM 与 k-sigma、均值漂移聚类的异常检测效果,具体使用的数据流描述如下(Rampisela et al,2021):

(1)Real Tweets-GOOGLE:主要记录了 Twitter 上关于 GOOGLE 网站的提及次数,数据每 5min 进行一次统计。

(2)Real Ad Exchange:收集了在线广告点击率,包含每次点击成本 CPC(Cost per click),以及每千人印象成本 CPM(Cost per thousand impressions)。

(3)Real Known Cause:主要包含导致异常的数据,数据特征包含环境温度、CPU 使用情况、定集群 CPU 使用情况、CPU 使用率。

(4)大气污染数据集:记录了北京某空气监测站收集的天气信息和空气污染指数,数据每小时采集一次,数据特征包括日期、PM2.5 浓度、露点、温度、风向、风速、累积小时雪量和累积小时雨量。

(5)电力数据集:共有 45312 个数据,记录了澳大利亚的电价数据,数据集中有 8 个特征,两种类别表示电价的变化趋势。

3.2 结果分析

异常检测数据集通常正负样本不均衡,仅通过准确率等指标很难评估模型。因此使用 AUC(Area Under Curve)分析模型的效果,其能够避免数据样本不均衡导致的评价失误。AUC 的值为 ROC(Receiver Operating Characteristic Curve)曲线下的面积,面积越大,代表分类效果越好。将不同模型在 NAB 的 CPC 数据(Swiniarski et al,2003)、Real Tweets 中的 GOOGLE 数据、Real Known Cause 中的 request 数据上进行测试(Sachin et al,2020),得出结果如图 3 所示。

图 3中使用 LSTM+差值正态建模的 ROC 曲线基本均超过了其他曲线,曲线下的面积更大,即 AUC 值更大,分类性能更强,对数据流中的异常检测效果更好。综合模型的准确率、召回率等指标对模型进行评估,结果如表 1 所示。

在表 1 中,使用 LSTM+正态建模异常检测模型的 AUC 面积更大,且模型的 $F1$ 值更优,代表在异常数据的识别表现更好,更能捕捉数据流的变化,在异常数据检测能力上优于其余 3 种模型。

通过进一步实验测试加入数据分布变化检测模块的效果。实验基于 Python3.5 进行实现和验证。将 LSTM 整体模型和不同的检测方法结合,应用于测试数据集上,使用电力数据集和大气污染数据集,其与通信数据具有大部分时间正常、少数时间异常的相似点。

将 LSTM 模型和常用的检测方法 ADWIN、DDM 以及 SVM 参数法应用于电力数据集与大气污染数据集上,得出的结果如图 4 所示。

由图 4 可见,起初数据流分布较为稳定,模型性能没有明显差异,随着数据不断输入,不

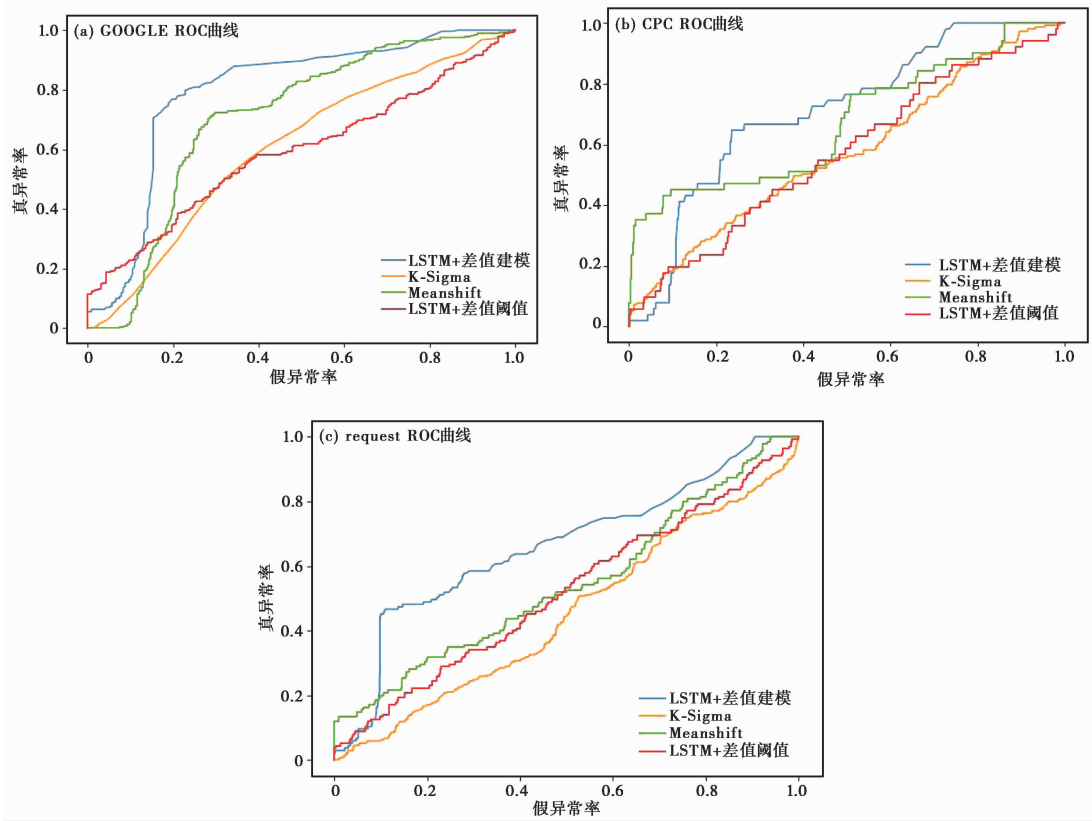

图 3　NAB 数据下不同模型的 ROC 曲线

表 1　　**不同数据下不同异常检测模型的数据**

数据集	算法名	准确率	精确率	召回率	*F*1 值	AUC 面积
Real Tweets-GOOGLE	K-sigma	0.1891	0.1319	0.9607	0.2320	0.6042
	Mean shift	0.6743	0.3860	0.4380	0.4104	0.6984
	LSTM+阈值	0.4400	0.2083	0.6333	0.3135	0.6324
	LSTM+正态	0.7956	0.4600	0.5036	0.4808	0.7939
Real Known Cause	K-sigma	0.3980	0.1119	0.9146	0.1994	0.5137
	Mean shift	0.6428	0.3150	0.4509	0.3709	0.5732
	LSTM+阈值	0.5891	0.1673	0.5850	0.2600	0.5026
	LSTM+正态	0.6900	0.3793	0.4313	0.4036	0.6546
Real Ad Exchange	K-sigma	0.6218	0.2175	0.3959	0.2798	0.5278
	Mean shift	0.7535	0.3641	0.4017	0.3819	0.6431
	LSTM+阈值	0.5475	0.1820	0.6670	0.2860	0.5482
	LSTM+正态	0.7729	0.3825	0.4671	0.4202	0.7538

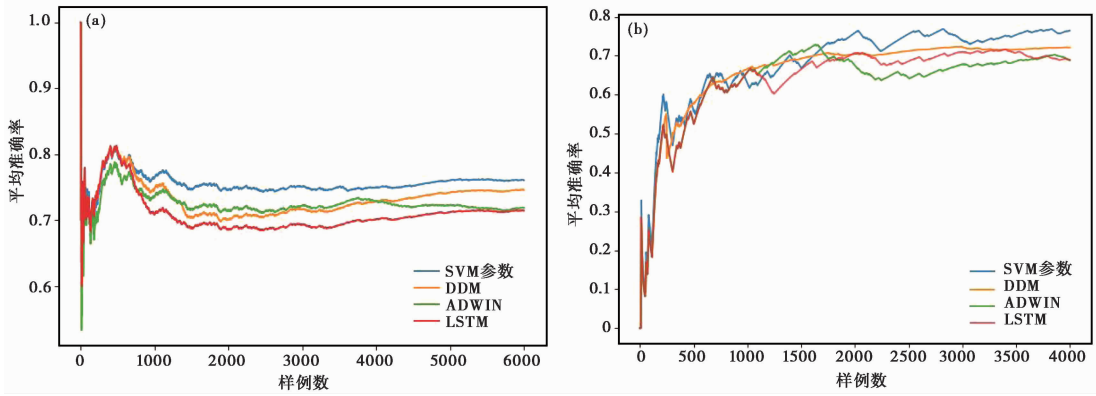

图 4 电力数据(a)和大气污染数据(b)的平均准确率曲线

同的方法根据自身对于数据分布变化的判断更新模型,模型准确率发生变化,可以看出本文方法的准确率优于其他 2 种方法。

由不同数据下的检测算法性能表现(表 2)可见,LSTM+SVM 方式的准确率较优于其他 3 种方式。在处理时间和占用内存上,由于 SVM 参数法需要计算两数据块的分类平面,因此处理时间会相对较长,占用内存也会相对较大,但考虑到通信数据流异常检测中更需要保障模型的性能,因此使用该方法分析通信数据流效果更好。

表 2 **不同数据下的检测算法性能表现**

数据集	算法名	准确率	占用内存/GB	处理时间/s
电力数据	LSTM+SVM	0.7754	8.871	131.96
	LSTM+ADWIN	0.7239	11.410	258.67
	LSTM+DDM	0.7507	6.496	111.68
	LSTM	0.7214	3.654	47.90
大气污染数据	LSTM+SVM	0.7563	10.655	135.94
	LSTM+ADWIN	0.6873	11.856	300.47
	LSTM+DDM	0.7187	9.061	30.84
	LSTM	0.6739	4.875	18.67

通过获取到的2017 年九寨沟 7.0 级地震前后部分区域的平均信号强度数据、可连接基站数和可连接 Wi-Fi 数等数据的变化,包含地震前后各 1h 共计 120min 的数据变化,通过模拟均值和方差以及震后的数据变化,对数据进行扩充,生成了实验通信数据,包含 10000 个网格的 120min 变化数据。

利用特征约简算法进行约简,得到的每个属性的重要性评分,如表 3 所示,可以看出网速和未读短信数重要性较差,因此约简后的属性保留 3 个:基站数、信号质量和 Wi-Fi 数量。

将设计的模型应用于通信数据上,其中时间窗口设置为 10,异常分数阈值设为 0.35,验证模型在通信数据异常检测上的可行性,得到的结果如表 4 所示,可以看出,LSTM+差值建模异常检测模型的性能相比其他异常检测模型更优。

表 3　　属性重要性评分

属性名	重要性评分
基站数量	0.73
信号强度	0.62
网速	0.11
未读短信数	0.17
Wi-Fi 数量	0.63

表 4　　通信数据下的模型性能表现

算法名	准确率	精确率	召回率	*F*1 值	AUC
LSTM+差值建模	0.6943	0.5281	0.6757	0.5914	0.8173
LSTM+差值阈值	0.5984	0.8132	0.3190	0.4672	0.6317
Meanshift	0.6472	0.5810	0.5478	0.5371	0.7548
K-sigma	0.5266	0.5173	0.4256	0.4538	0.6029

不同程度的地震会造成不同程度通信数据的变化，通过对异常分数进行划分，评估异常等级。将异常分数按照大小划分为 0~5 共 6 个等级，利用 2017 年九寨沟 7.0 级地震通信数据，选取第 70min 数据得出的异常分数，按照[0 ~ 0.35]、[0.35 ~ 0.48]、[0.48 ~ 0.6]、[0.6~0.7]、[0.7~0.75]、[0.75~1]的评级方式评估大致异常等级。

通过对通信数据异常等级进行评估(图 5)，能够对受影响区域进行大致划分，越靠近中心的区域，异常等级更高。同样在通信数据集上，使用数据分布检测 3 种方法进行测试，得到的结果如表 5 所示，其中 SVM 模型参数法时间窗口设置为 160，阈值设置为 0.2。

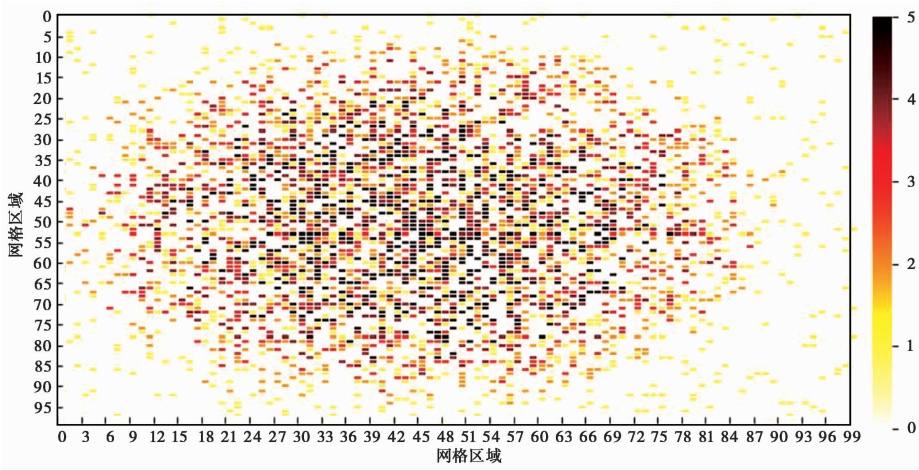

图 5　通信数据异常等级评估结果

由表 5 可见，使用 SVM 模型参数法提升了异常检测模型的准确率。本文方法在通信数据上仍有良好的表现，能够完成对通信数据流的异常检测。

将 10000 个网格按照 1km 重新划分网格，新网格异常等级按照包含网格的等级平均值

表 5 不同检测方法在通信数据上的应用

比较项	ADWIN	DDM	SVM 模型参数法
准确率	0.69	0.71	0.75
占用内存/GB	15.704	9.763	11.269
处理时间/s	346.810	104.780	198.467

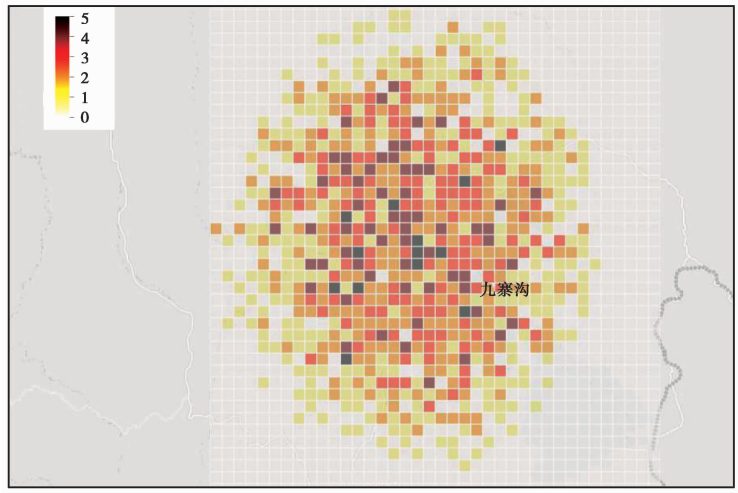

图 6 通信数据异常等级评估结果的 WebGIS 展示

计算,通过 WebGIS 可以将得到的评估结果在地图上进一步展示,如图 6 所示。

由图 6 可见,网格约简后图形更加简约直观,且排除了少数噪声点对于图的影响,能够大致判断受影响的区域,从而辅助抗震救灾。后续还可通过异常等级与灾区等级进行数据对应,利用通信数据异常分析震后的灾情(Wang et al,2016)。

因此,通过基于通信数据流的异常检测能够对通信数据的异常变化进行识别,并且能够对模型中出现的数据分布变化进行识别,并对模型进行更新,应用于震后的通信数据上,能够检测灾后出现的异常通信数据变化,从而进行一定程度上的灾情分布推测。

4 结论

针对震后灾情数据获取不及时的情况,本文设计了基于 LSTM 的数据异常检测模型,识别震后通信数据发生的异常变化;针对数据流模型中出现的数据分布变化,使用 SVM 参数法进行检测。通过在实验通信数据上测试,证明了模型对通信数据异常检测的可行性。此外,通过 WebGIS 的展示,证明通过统计所有区域内的异常等级,可为灾情分析提供有效数据,为应急救灾提供帮助。

参考文献

白惠文,2019. 基于核判别分析的数据流的在线学习算法研究. 硕士学位论文. 大连:大连理工大学.

薄涛,李小军,陈苏,等,2018. 基于社交媒体数据的地震烈度快速评估方法. 地震工程与工程振动,**38**(5):206~215.

陈琳，张明，朱耿青，等，2011. 基于地震动峰值加速度的地震灾害评估方法研究. 见：2011 中国地理信息产业大会暨中国地理信息产业协会成立大会. 北京：中国地理信息产业协会.

董翔，肖兰喜，杜宪宋，等，2007. 基于网络的山东地震灾情收集分析处理系统. 华北地震科学，**25**(3)：6～10，33.

傅征祥，丁香，王晓青. 2006. 2020 年前中国大陆发生 7 级以上大地震频次和最大震级预测的初步研究. 地震，**2006**(1)：35～39.

刘在涛，王栋梁，张维佳，等，2011. 基于贝叶斯判别分析的地震应急响应等级初判方法. 地震，**31**(2)：114～121.

王晓青，傅征祥，丁香，等，2000. 地震灾害损失预测系统计算原理与主要功能. 地震，**20**(增刊 I)：222～226.

Bengtsson L，Lu X，Thorson A，et al，2011. Improved response to disasters and outbreaks by tracking population movements with mobile phone network data：A post-earthquake geospatial study in Haiti. PLoS Med，**8**(8)：e1001083.

Finazzi F，2016. The earthquake network project：Toward a crowdsourced smartphone-based earthquake early warning system. Bull Seismol Soc Am，**106**(3)：1088～1099.

Guo Y K，Li Y Y，Xu Y，2021. Study on the application of LSTM-LightGBM Model in stock rise and fall prediction. MATEC Web Conf，**336**：05011.

Minku L L，Yao X，2012. DDD：a new ensemble approach for dealing with concept drift. IEEE Trans Knowl Data Eng，**24**(4)：619～633.

Rampisela T V，Andarlia H T，Rustam Z，2021. Classification of the likelihood of Indonesian Facebook users in spreading hoaxes using Support Vector Machine(SVM). J Phys：Conf Ser，**1725**(1)：012019.

Sachin M M，Baby M P，Ponraj A S，2020. Analysis of energy consumption using RNN-LSTM and ARIMA Model. J Phys：Conf Ser，**1716**(1)：012048.

Swiniarski R W，Skowron A，2003. Rough set methods in feature selection and recognition. Pattern Recogn Lett，**24**(6)：833～849.

Wang Y Q，Huang M L，Zhu X Y，et al，2016. Attention-based LSTM for aspect-level sentiment classification. In：Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin：Association for Computational Linguistics，606～615.

LSTM-based Anomaly Detection and Analysis of Post-earthquake Communication Data

Wang Xiaoxiang[1)]，Liu Dongtian[1)]，Liu Nanjiang[2)]，Ding Yi[2)]，Jiang Lixin[3)]

1) School of Information and Communication Engineering，BUPT，Beijing 100876，China

2) National Disaster Reduction Center of China，Ministry of Emergency Management，Beijing 100124，China

3) China Earthquake Networks Center，Beijing 100045，China

Abstract　After the earthquake，the communication data usually undergoes abnormal changes. Through the analysis of the communication data abnormality，effective disaster data can be provided，which is a basis for decision-making for earthquake relief. Analyzing abnormal communication data helps to better understand the impact of the earthquake and can provides auxiliary support for following-up command and disaster relief. This paper analyzes the communication data anomalies after the earthquake based on LSTM，which mainly includes：communication data stream preprocessing，anomaly detection model based on LSTM，and data distribution change detection model. The model in this paper is capable of identifying abnormal changes in communication data and provides disaster data for subsequent hazard analysis.

Keywords：**Communication data**；**Data processing**；**Anomaly detection**；**Data distribution changes**