文章编号：2095-2163(2020)02-0228-04 中图分类号：TP29 文献标志码：A

基于 LSTM 的地震前兆数据分析算法设计与实现

王圆圆，孙可可
（防灾科技学院 应急管理学院，河北 三河 065201）

摘 要：地震前兆数据具有短期、中期、长期的变化，研究地震前兆数据的变化规律对地震预测具有重要意义。论文采用目前最流行的深度学习技术之 LSTM 模型，对海量前兆数据进行学习，学习出其短期、中期、长期的变化规律，建立前兆数据深度模型，并根据模型预测出前兆波形。论文将预测的前兆波形与实际的观测的波形数据进行对比，实验表明，LSTM 算法能很好地拟合观测数据。

关键词：地震前兆数据；LSTM；数据拟合；时间序列分析

Design and implementation of seismic precursor data analysis algorithm based on LSTM

WANG Yuanyuan, SUN Keke
(School of Emergency Management, Institute of Disaster Prevention, Sanhe Hebei 065201, China)

【Abstract】 Earthquake precursor data have short-term, medium-term and long-term changes. Studying the change rule of earthquake precursor data is of great significance for earthquake prediction. This paper uses the most popular LSTM model of deep learning technology to learn the massive precursory data, learn its short-term, medium-term and long-term change rules, establish the depth model of precursory data, and predict the precursory waveforms according to the model. In this paper, the predicted precursory waveform is compared with the actual observed waveform data. Experiments show that the LSTM algorithm can well fit the observation data.

【Key words】 seismic precursor data; LSTM; data fitting; time series analysis

0 引 言

地震前兆现象主要分为宏观现象和微观现象。本文主要分析地震前兆现象中的微观现象，例如逸出气氡、气压等。地震台站检测到的地震前兆数据在不间断且不规律的波动中会蕴藏着动态演化和信号变化[1]。地震前兆数据具有在结构上的复杂性、前兆观测方法的不固定性、数据位精度的可变性、数据采样率的不一致性、数据源的多样性等特点。地震前兆数据变化规律有长期、中期、短期变化[2]。通常用逐级降采样率取年、季度、月、周、日、小时、分钟、秒的平均值进行数据分析。正是由于这些大量高采样率的观测数值和与其协作的分析人员逐天逐台的采集和处理模式，传统的处理模式和计算方法已经很难在海量的观测数据中迅速自动定位精确位置，这也制约了人类研究地震前兆数据的进展[3]。未来，在保证数据完整性的前提下，面对海量的地震前兆数据，利用机器学习进行地震前兆数据分析是一个至关重要的研究方向[4]。如果人类通过数据分析掌握了地震前兆数据变化规律，会对我们的研究带来莫大的帮助[5]。

1 LSTM 原理

长短期记忆(long short-term memory，LSTM)模型由不同的记忆单元组成，例如单元状态(cell state)和通过“门”(gate)，其中通过“门”又分为 3 类[6]，分别是：遗忘门(forget gate)、输入门(input gate)、输出门(output gate)[7]。LSTM 的通过“门”(gate)发挥增加或删除信息的功能，对应着模型中的记忆或遗忘的功能。“门”是一种将抽象具体化的结构，进行信息过滤，且由一个点乘和一个 sigmoid 函数构成。sigmoid 函数的输出值域区间为[0,1]，1 代表全部通过，0 表示直接全部丢掉。3 个这样的门组成一个 LSTM 单元。LSTM 记忆单元总图如图 1 所示。对此拟做研究分述如下。

(1)遗忘门。遗忘门的 sigmoid 函数的输入值是上一单元的输出 h_{t-1} 和本单元的输入 x_t 数据，再为 c_{t-1} 中的每一项产生一个在[0,1]内的值。通过这种方式来控制上一个单元状态被遗忘的程度[8]。主要函数如下：

作者简介：王圆圆(1998-)，女，本科生，主要研究方向：机器学习；孙可可(1994-)，男，硕士研究生，主要研究方向：机器学习。
通讯作者：孙可可 Email：405070470@qq.com

收稿日期：2019-11-17

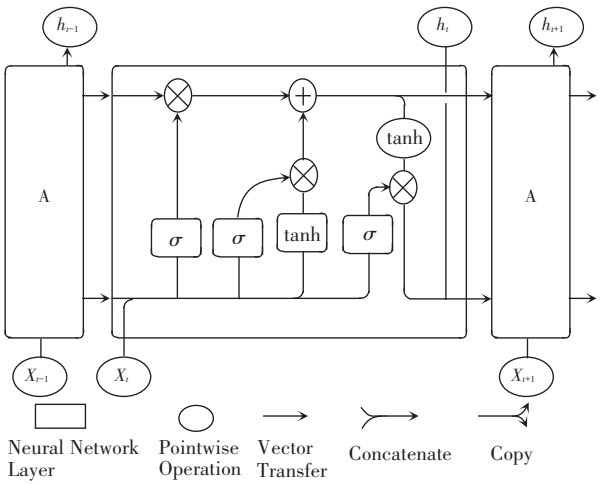


图 1　LSTM 记忆单元总图

Fig. 1　LSTM memory unit general diagram

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + b_f), \tag{1}$$

(2)输入门。输入门和一个 tanh 函数匹配用来限制加入哪些数据。tanh 函数可以计算出下一轮的候选向量 $\vec{C}_t$，输入门为 $\vec{C}_t$ 中的每一项生成一个值，并将其控制在[0,1]内，限制增加新信息的数量。此时，可以计算出遗忘门的输出 f_t，用来控制上一单元被遗忘的程度，同时加上输入门的输出 f_t，用来限制增加新信息的数量，在此基础上，更新本记忆单元的单元状态[9]。主要公式为：

$$\boldsymbol{C}_t = f_t * \boldsymbol{C}_{t-1} + i_t * \overrightarrow{\boldsymbol{C}_t}, \tag{2}$$

$$i_t = \sigma(\boldsymbol{W}_i \cdot [h_{t-1}, x_t + b_i]), \tag{3}$$

$$\overrightarrow{\boldsymbol{C}_t} = \tanh(\boldsymbol{W}_C \cdot [h_{t-1}, x_t] + b_C), \tag{4}$$

(3)输出门。输出门用来控制当前的单元状态有多少被过滤掉。先将单元状态激活，输出门为其中每一项产生一个在[0,1]内的值，控制单元状态被过滤的程度[10]。主要公式如下：

$$o_t = \sigma(\boldsymbol{W}_o [h_{t-1}, x_t] + b_o), \tag{5}$$

$$h_t = o_t * \tanh(\boldsymbol{C}_t). \tag{6}$$

(4)单元状态(cell state)。这是 LSTM 的关键，即用图 1 上半部分的水平直线来表示，可以将数据从上一个单元传输到下一个单元，就象一条数据传送带一样贯穿在整个结构中，在传输数据的同时只会有很少的线性相互作用[11]。单元状态局部图如图 2 所示。

2　实验数据预处理

由于数据是精确到秒的检测值，据统计分析可知，一个月的分钟数据会达到三十万。而在庞大的数据量中，却会因为检测仪器故障、自然环境、人为因素等作用导致监测结果中存在缺失值。为了保证提取数据的完整性和预测结果的准确性，就要对缺失值进行处理。在本次研究中，则将缺失值补齐，再进行数据分析。对此可做分析论述如下。

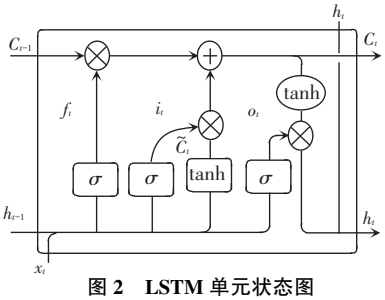


图 2　LSTM 单元状态图

Fig. 2　LSTM unit status diagram

2.1　缺失值处理

缺失值处理方法有 3 种，分别是：数据补差、删除记录和不处理。本次研究中，采用补差记录的方法。原始数据中的缺失数据，采用补差法，用其周围的数据进行补差。

2.2　数据规范化

由于个别数据会影响正常数据，不进行数据规范化会影响数据分析结果的准确性。本文采用 Z-score 方法进行数据规范化，因为 Z-score 的数据分布情况是正态分布 ($N(0,1)$)，并且正态分布又被称为零-均值规范化。Z-score 公式可表示为：

$$z = \frac{x - \mu}{\sigma}. \tag{7}$$

其中，x 是原始数据，z 是规范后的数据[12]。

研究可知，mu 是均值，signma 是标准差，Z-score 的分布如图 3 所示。

3　仿真实验

3.1　不同降采样方法对数据分析的影响

本实验中取 2005～2009 年山西省临汾地震观测站第三个测项的气压值，全球精确坐标度为 $(36.073 * N, 111.505 * E)$、海拔为 443.31 m 的数据。和 2008 ～ 2013 年地震研究所测点为白浮的逸出气氡值，精确位置度为 $(40.184 * N, 116.234 * E)$、海拔为 45 m 的数据。

3.2　气压值的不同降采样方法

分析不同的降采样方法对气压值数据拟合结果的影响。根据 4 种最大值、最小值、均值、中位数不同的降采样方法得出的采样率为 3 天时的气压值的数据拟合结果图和误差结果图，详见图 4～图 7。

分析图 4～图 7 可知，当降采样方法为最大值时，$RMSE$ = 160.956 3、最小值时，$RMSE$ = 224.664 1、平均值时，$RMSE$ = 9.522、中值时，$RMSE$ = 12.390 9。通过比较 4 种降采样的数据拟合结果图和误差值 $RMSE$，选出误差最小的情况为平均值法。

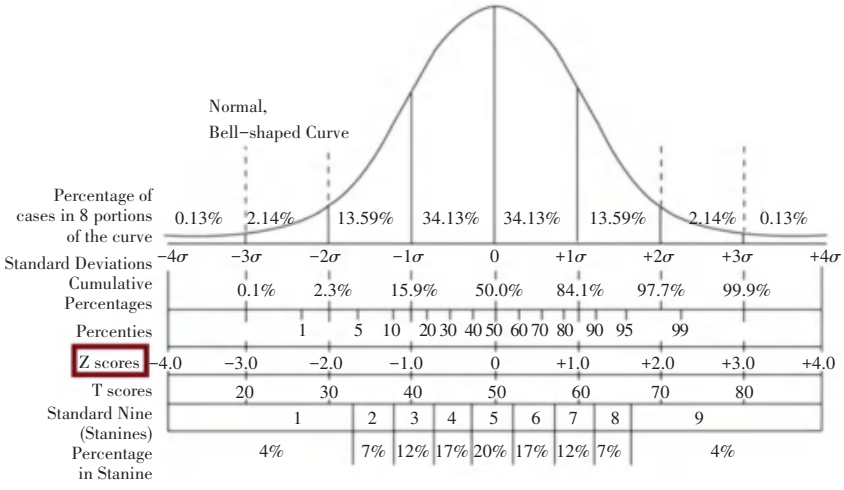
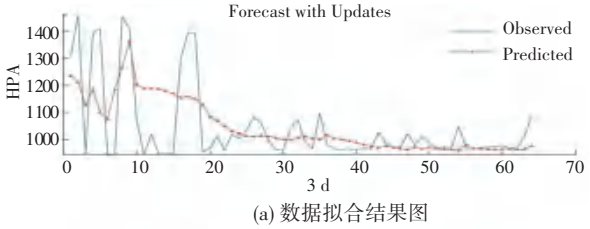


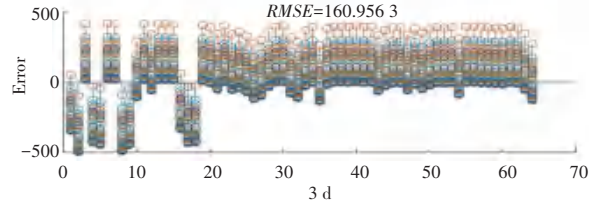
图 3 Z-scores 分布图

Fig. 3 Z-scores distribution diagram



(a) 数据拟合结果图

(a) Data fitting results graph

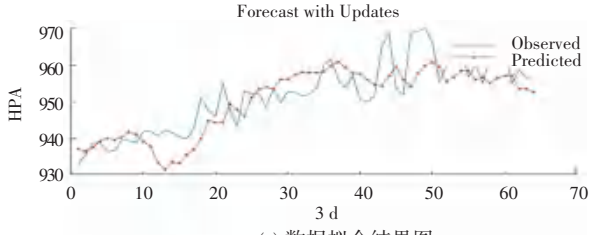


(b) 误差结果图

(b) Error result graph

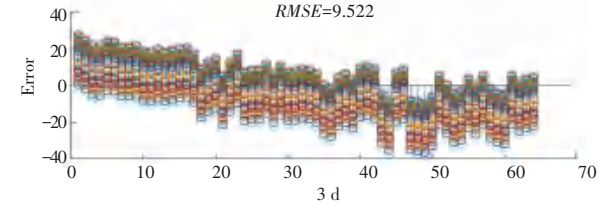
图 4 MAX 气压数据拟合图

Fig. 4 MAX air pressure data fitting diagram



(a) 数据拟合结果图

(a) Data fitting results graph

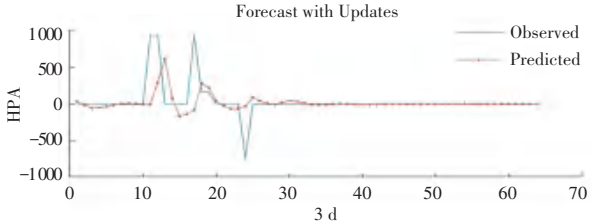


(b) 误差结果图

(b) Error result graph

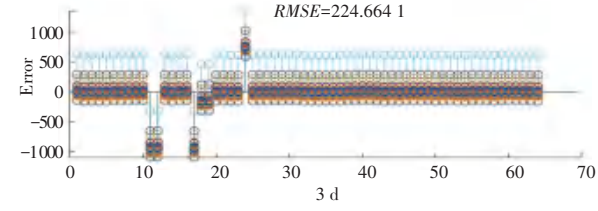
图 6 MEAN 气压数据拟合图

Fig. 6 MEAN air pressure data fitting diagram



(a) 数据拟合结果图

(a) Data fitting results graph

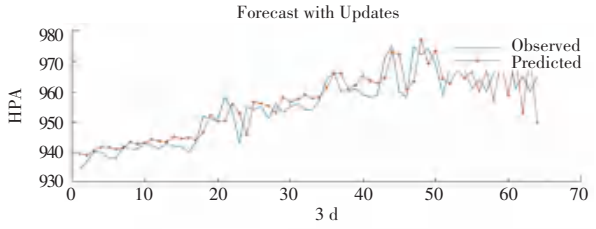


(b) 误差结果图

(b) Error result graph

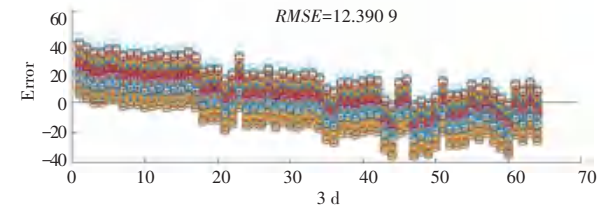
图 5 MIN 气压数据拟合图

Fig. 5 MIN air pressure data fitting diagram



(a) 数据拟合结果图

(a) Data fitting results graph



(b) 误差结果图

(b) Error result graph

图 7 MEDIAN 气压数据拟合图

Fig. 7 MEDIAN air pressure data fitting diagram

4　结束语

综前所述可知，在数据分析过程中，首先对数据进行预处理。预处理分为两步，分别是：缺失值处理，采用补差法；降采样处理，有最大值、最小值、平均值、中位数四种方法。然后，给出了具体的实验步骤，即：选出误差值最小的降采样方法，并用不同的采样率运行，再选出误差最小和数据拟合最优的情况。最后，得出数据预测结果。经过上述的实验步骤得出如下结论：2005～2009 年山西省临汾地震观测站，全球精确坐标度为(36.073 * N,111.505 * E)、海拔为 443.31 m 的最优情况是采用平均值降采样方法，采样率为 3 天的情况下得到的误差值最小，数据拟合结果最优。

参考文献

[1] 刘子维. 地震前兆数据异常识别关键技术研究[D]. 武汉：武汉大学，2016.

[2] 杨满栋，李闽峰，郝平，等. 非结构化时间序列地震数据信息网络服务系统[C]//2001 年中国地球物理学会年刊——中国地球物理学会第十七届年会论文集. 昆明：中国地球物理学会，2001：1.

[3] 徐光宇. 日本国立防灾科学技术中心的地震前兆数据分析处理系统概述[J]. 国际地震动态，1988(6)：5.

[4] 武安绪，张永仙，张小涛. 融合地震前兆背景趋势变化与短临异常提取的一种定量时序数据处理新算法[C]//中国地震学会 2009 年汶川地震学术研讨会. 重庆：中国地震学会，2009：35.

[5] DOETSCH P，KOZIELSKI M，NEY H. Fast and robust training of recurrent neural networks for offline handwriting recognition [R]. Aachen，Germany：RWTH Aachen University，2014.

[6] KIM H Y，WON C H. Forecasting the volatility of stock price index：A hybrid model integrating LSTM with multiple GARCH-type models[J]. Expert Systems With Applications，2018，103.

[7] 池国民，赵银刚，杜桂林，等. 安丘地震台地震前兆数据跟踪分析[J]. 地震地磁观测与研究，2017，38(3)：203.

[8] 陈卓，孙龙祥. 基于深度学习 LSTM 网络的短期电力负荷预测方法[J]. 电子技术，2018，47(1)：39.

[9] 杨煜，张炜. TensorFlow 平台上基于 LSTM 神经网络的人体动作分类[J]. 智能计算机与应用，2017，7(5)：41.

[10] 蒲春，孙政顺，赵世敏. Matlab 神经网络工具箱 BP 算法比较[J]. 计算机仿真，2006，5(5)：142.

[11] ZHANG Duo，LINDHOLM G，RATNAWEERA H. Use long short-term memory to enhance Internet of Things for combined sewer overflow monitoring[J]. Journal of Hydrology，2018，556：409.

[12] 米硕，孙瑞彬，李欣，等. 基于 LSTM 的循环神经网络模型确立睡眠与病例诊断结果的关系[J]. 科技与创新，2018(7)：99.

(上接第 227 页)

[3] SERRANO M A，BOGUNA M. Topology of the World Trade Web[J]. Phys Rev E Stat Nonlin Soft Matter Phys，2003，68(2)：015101.

[4] 宋周莺，车姝韵，杨宇. “一带一路”贸易网络与全球贸易网络的拓扑关系[J]. 地理科学进展，2017，36(11)：1340.

[5] 杨丽梅，翟婧帆. 中国与“一带一路”沿线国家贸易网络分析[J]. 商业经济研究，2019(2)：119.

[6] HOEKMAN B，NICITA A. Trade policy，trade costs，and developing country trade[J]. World Development，2011，39(12)：2069.

[7] 胡晓丹. “一带一路”交通基建项目对提升沿线地区贸易效率的作用[J]. 湖南科技大学学报(社会科学版)，2019，22(2)：60.

[8] 杜方叶，王姣娥，谢家昊，等. “一带一路”背景下中国国际航空网络的空间格局及演变[J]. 地理科学进展，2019，38(7)：963.

[9] 卓志强，姚红光. “一带一路”沿线航空网络结构及其鲁棒性研究[J]. 物流科技，2018，41(5)：78.

[10] 种照辉，覃成林. “一带一路”贸易网络结构及其影响因素—基于网络分析方法的研究[J]. 国际经贸探索，2017，33(5)：16.

[11] BARRAT A，BARTHÉLEMY M，PASTORSATORRAS R，et al. The architecture of complex weighted networks.[J]. Proc Natl Acad Sci U S A，2004，101(11)：3747.

[12] 刘军. 社会网络分析导论[M]. 北京：社会科学文献出版社，2004.

[13] ALAVI M. Computer-mediated collaborative learning：An empirical evaluation[J]. MIS Quarterly，1994，18(2)：159.

[14] 陈银飞. 2000-2009 年世界贸易格局的社会网络分析[J]. 国际贸易问题，2011(11)：31.

[15] 吴祥平. 基于复杂网络的昌九一体化物流网络研究[D]. 上海：上海大学，2015.

[16] 赵晓媚. 我国互联网金融发展的空间效应分析—基于社会网络分析的方法[D]. 上海：上海师范大学，2019.

[17] 杨文龙，杜德斌，马亚华，等. “一带一路”沿线国家贸易网络空间结构与邻近性[J]. 地理研究，2018，37(11)：2218.