· 调查报告与分析 ·

长短时记忆神经网络模型在河北省麻疹疫情发病趋势预测中应用

许晓萌，崔世恒，王亚菲，孙丽，丛艳丽，王晶辉，李静，张振国

河北省疾病预防控制中心免疫规划管理所，石家庄 050021
通信作者：孙丽，E-mail：1126sl@163.com

【摘　要】目的 探讨长短时记忆（LSTM）神经网络模型在麻疹疫情发病趋势预测上的可行性，为科学防控麻疹提供参考依据。**方法** 收集中国疾病预防控制信息系统传染病监测系统中河北省发病日期为 2004 年 1 月 — 2020 年 12 月的 51 012 例麻疹病例发病数据构建 LSTM 神经网络模型，选择最优模型对河北省麻疹疫情发病趋势进行预测，并采用均方误差平方根（RMSE）和平均绝对误差（MAE）评价模型预测效果。**结果** 河北省 2004、2005、2006、2007、2008、2009、2010、2011、2012、2013、2014、2015、2016、2017、2018、2019 和 2020 年分别报告麻疹病例 950、4 837、7 953、4 973、2 273、3 359、14 457、79、38、353、5 365、3 825、1 825、287、241、130 和 67 例，从 2015 年开始河北省麻疹发病数逐年下降，且发病具有明显的季节性；视窗长度分析结果显示，当视窗长度取 3 时，模型预测效果最好，RMSE 和 MAE 值分别为 17.288 和 12.334；本研究构建 LSTM 神经网络模型对河北省 2017 — 2020 年麻疹发病情况进行预测，模型预测的发病趋势与实际趋势基本一致，RMSE 和 MAE 值在 2017、2019 和 2020 年均 < 10，但 2018 年误差略大。**结论** LSTM 神经网络模型在河北省麻疹疫情发病趋势预测中效果较好，可用于麻疹发病趋势的研判和风险评估。

【关键词】 麻疹；发病趋势；预测；长短时记忆 (LSTM) 神经网络模型；应用

Predicting trend of measles epidemic in Hebei province: an empirical study with long short-term memory neural network model

XU Xiaomeng, CUI Shiheng, WANG Yafei, SUN Li, CONG Yanli, WANG Jinghui, LI Jing, ZHANG Zhenguo (*Department of Immunization Program Administration, Hebei Provincial Center for Disease Control and Prevention, Shijiazhuang* 050021, *China*)
Corresponding author: SUN Li, E-mail: 1126sl@163.com

【Abstract】 Objective To explore the feasibility of predicting the trend of measles epidemic using long short-term memory (LSTM) neural network model for conducting prevention and control of measles. **Methods** The data on 51 012 measles cases reported in Hebei province form 2004 through 2020 were collected from China Information System for Disease Control and Prevention. The LSTM neural network model was constructed and the optimal model was selected to predict the trend of measles epidemic in the province. Rooted mean squared error (RMSE) and mean absolute error (MAE) were used to evaluate the prediction of model established. **Results** The annual number of measles cases reported in the province during the 17-year period were 950, 4 837, 7 953, 4 973, 2 273, 3 359, 14 457, 79, 38, 353, 5 365, 3 825, 1 825, 287, 241, 130, and 67, respectively, with a persistent decline since 2015. In addition, an obvious seasonality was observed in the incidence of measles. Using the collected data of 2017, the window length of 3 was determined for the constructed LSTM neural network model, with the RMSE of 17.288 and the MAE of 12.334, and the model was adopted to predict monthly number of measles cases from 2017 through 2020. The predicted monthly numbers of measles incidence were basically consistent with the number observed and the values of RMSE and MAE for years of 2017, 2019 and 2020 were all below 10, but the values for 2018 were slightly higher. **Conclusion** The constructed LSTM neural network model in this study showed a good efficiency in predicting monthly measles incidence in Hebei province and the model could be used in the analysis on measles incidence trend and epidemic risk assessment.

【Keywords】 measles; incidence trend; prediction; long short-term memory neural network model; application

 DOI：10.11847/zgggws1141808
基金项目：河北省 2021 年度医学科学研究课题计划（20210680）
第一作者：许晓萌（1992 – ），主管医师，硕士，研究方向：传染病预防与控制。
收稿日期：2023-03-21　修回日期：2023-06-16　录用日期：2023-09-04　责任编辑：郭薇
利益冲突：不存在　伦理审查：已获取　出版授权：全体作者已与编辑部签署作者贡献声明及版权转让协议

麻疹是由麻疹病毒引起的一种急性呼吸道传染病，其传染性强，属于中国法定乙类传染病[1]。随着含麻疹成分疫苗的广泛应用以及消除麻疹工作目标的制订，我国麻疹大规模流行得到了有效控制，但由于我国地域广阔、人口众多、人口流动性强以及各地卫生工作发展水平不一致，麻疹小规模暴发疫情仍时有发生，麻疹的防控仍是公共卫生工作的重要部分[2]。因此，若能合理预测麻疹发病趋势，及早采取有针对性的预防措施，将有助于做好麻疹疫情的防控工作。人工神经网络（artificial neural network, ANN）作为一种非线性的建模和预测方法，由于具有自组织、自学习和自适应等优点而受到很多学者的广泛关注，为传染病疫情预测提供了一种新的分析方法[3－4]。为探讨长短时记忆（long short term memory, LSTM）神经网络模型在麻疹疫情发病趋势预测上的可行性，为科学防控麻疹疫情提供参考依据，本研究收集中国疾病预防控制信息系统传染病监测系统中河北省发病日期为 2004 年 1 月 — 2020 年 12 月的 51 012 例麻疹病例发病数据构建 LSTM 神经网络模型，选择最优模型对河北省麻疹疫情发病趋势进行预测，并采用均方误差平方根（rooted mean squared error, RMSE）和平均绝对误差（mean absolute error, MAE）评价模型预测效果。结果报告如下。

1　资料与方法

1.1　资料来源　资料来源于中国疾病预防控制信息系统传染病监测系统，收集其中发病时间为 2004 年 1 月 — 2020 年 12 月的河北省 51 012 例麻疹病例发病数据。主要研究者已完成传染病监测系统的业务管理备案，负责河北省麻疹监测信息的收集、录入、分析、评价、报告和反馈等工作。本研究已通过河北省疾病预防控制中心伦理委员会批准（批号：IRBS2021－009）。

1.2　统计分析　将河北省 2004 — 2020 年麻疹月发病例数进行数据预处理，选取某一年数据进行视窗长度的确定，选择最佳视窗长度构建 LSTM 神经网络模型，对河北省 2017 — 2020 年麻疹月发病情况进行预测，评价模型预测效果。基于 Python 3.6（程序设计语言）和 Keras 2.1.6（深度学习框架）构建 LSTM 神经网络模型进行麻疹发病趋势预测分析。

1.2.1　模型原理　LSTM 神经网络模型是一种特殊的循环神经网络（recurrent neural network, RNN），解决了 RNN 中梯度消亡和对长期依赖学习较慢的问题，常应用于分析长距离的时间序列数据[5]。LSTM 在记忆块中引入了 1 个记忆单元和控制记忆单元的 3 个门结构，分别为遗忘门、输入门和输出门，其中遗忘门的作用为选择上一时刻中有用的信息保留到当前时刻；输入门的作用为选择当前时刻有用的信息存储在记忆单元中；输出门的作用为选择记忆单元中有用的信息作为当前时刻网络的输出值。

1.2.2　数据预处理　使用最大最小归一化方法[6]对数据进行预处理，即对每一年的发病数采用下述公式进行归一化处理：

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

其中，x^*为数据归一化后的值；x 为原始数据，$x_{\max}$和$x_{\min}$分别为其最大值和最小值。

1.2.3　模型构建（图 1）　本研究采用全局趋势预测模型和局部数量预测模型相结合的方式进行预测。在基于全局趋势预测的模型中，从 2004 年开始以 12 个月的数据作为训练数据，以下一个月的数据作为训练标签，模型拓扑结构如图 1 左侧部分所示。例如，在训练阶段，将 2004 年 1 — 12 月数据作为训练数据，以 2005 年 1 月数据作为训练标签；将 2004 年 2 月 — 2005 年 1 月数据作为训练数据，以 2005 年 2 月数据作为训练标签；以此类推，得到训练模型。在测试阶段，通过上述模型迭代得到 2017 — 2020 年的预测趋势。例如，预测 2017 年 1 月发病趋势时，网络输入为 2016 年 1 — 12 月发病趋势；预测 2017 年 2 月份发病趋势时，网络输入为 2016 年 2 月 — 2017 年 1 月发病趋势。基于上述计算，本文得到 2017 — 2020 年所有月份的全局发病趋势。全局趋势预测模型的输入维度为（N × 12, 12, 1），输出维度为（N × 12, 1），其中 N 为年数、N × 12 为总月数。在基于局部数量预测的模型中，需设置视窗长度及步长，模型拓扑结构如图 1 右侧部分所示。假如视窗长度设置为 3、步长为 1，在训练阶段以 2016 年 1 — 3 月的月发病数作为训练数据，4 月的发病数作为训练标签；以 2016 年 2 — 4 月的月发病数作为训练数据，5 月的发病数作为训练标签；以此类推，得到训练模型。在测试阶段，该模型以 2016 年 10 — 12 月的月发病数作为输入，预测 2017 年 1 月的发病数，以 2016 年 11 月 — 2017 年 1 月的月发病数作为输入，预测 2017 年 2 月的发病数；以此类推，得到 2017 — 2020 年局部发病预测数。该局部数量预测模型的输入维度为（3 × 3, 3, 1），输出维度为（12, 1）。模型最终预测结果将整体趋势与局部数目相结合，其中整体趋势预测在得到趋势后将映射至实际发病数，局部数目预测将直接得到预测月发病数，最终结果通过权重拟合计算得出。

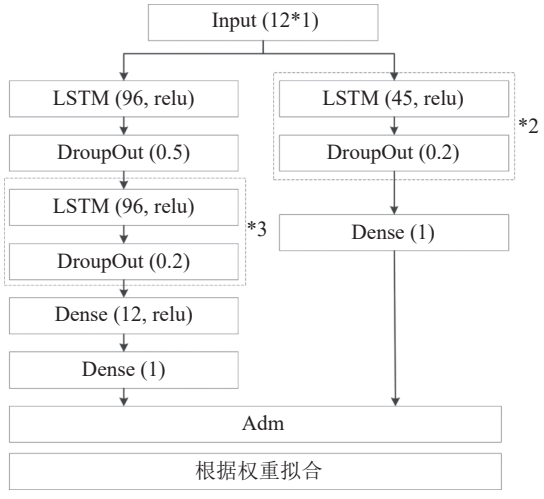

图 1 模型拓扑结构图

Fig. 1 Topology structure for predicting monthly number of measles cases with long short term memory neural network model

1.2.4 模型预测效果评价 模型预测效果采用 RMSE 和 MAE 值进行评价，RMSE 和 MAE 的值越小代表模型的预测效果越好[7]。RMSE 和 MAE 值的具体计算公式如下：

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{x}_i)^2} \quad (2)$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|x_i-\hat{x}_i| \quad (3)$$

其中，x_i表示时刻 i 模型的真实值，$\hat{x}_i$表示时刻 i 模型的预测值，n 表示测试时的样本量。

2 结 果

2.1 河北省 2004—2020 年麻疹疫情流行趋势（表 1）河北省 2004 年 1 月 — 2020 年 12 月共报告 51 012 例麻疹发病病例，其中 2010 年的发病例数最多（14 457 例），从 2015 年开始发病数逐年下降；除 2020 年外，河北省每年 3 — 5 月的麻疹发病例数均较多，麻疹发病具有明显的季节性，且各年份发病趋势趋于一致。

表 1 河北省 2004 — 2020 年麻疹月发病数

Tbl. 1 Reported monthly number of measles cases by year in Hebei province: 2004 – 2020

年份	月份												合计
	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月	
2004	10	32	126	115	99	63	37	40	19	57	166	186	950
2005	179	282	816	1 206	979	567	271	124	78	79	103	153	4 837
2006	314	509	1 119	1 659	2 016	1 192	381	153	81	70	199	260	7 953
2007	429	652	937	1 347	1 040	334	85	27	27	17	25	53	4 973
2008	66	138	387	431	564	323	161	75	36	24	30	38	2 273
2009	32	190	583	1 089	689	246	128	65	44	23	52	218	3 359
2010	655	1 239	2 712	3 974	4 569	1 016	175	65	26	8	10	8	14 457
2011	4	7	12	33	14	4	0	0	1	1	0	3	79
2012	0	1	4	6	10	6	5	0	2	1	0	3	38
2013	6	18	57	58	63	13	17	13	19	21	15	53	353
2014	344	912	1 466	1 205	801	288	110	54	19	30	18	118	5 365
2015	339	512	798	766	847	247	97	48	24	17	23	107	3 825
2016	144	216	412	491	307	126	66	20	6	9	12	16	1 825
2017	17	20	59	40	49	30	12	11	19	9	7	14	287
2018	48	25	26	36	22	25	19	6	5	13	7	9	241
2019	6	6	17	36	28	14	8	7	2	1	1	4	130
2020	42	5	4	2	1	4	1	1	2	1	0	4	67

2.2 河北省麻疹疫情发病趋势 LSTM 神经网络模型预测

2.2.1 视窗长度分析 本研究选取河北省 2017 年 1 — 12 月的麻疹月发病数用于确定视窗长度，以不同的视窗长度对数据进行处理，将处理过的数据作为 LSTM 模型的输入，采用 RMSE 和 MAE 值评价模型预测效果，选取最佳视窗长度，从而进行下一步的预测分析。结果显示，视窗长度为 3、4、5、6、7、8、9 时的 RMSE 和 MAE 值分别为 17.288 和 12.334、17.729 和 12.335、20.772 和 14.315、24.021 和 16.974、23.191 和 16.328、23.987 和 16.640、26.535 和 20.315，其中视窗长度取 3 时模型的预测效果最好，为此本研究将利用前 3 个月的发病数据预测后 1 个月的数据。

2.2.2 模型预测结果（表 2，图 2） 对河北省 2017—2020 年的月发病情况进行预测，模型预测效果评

价结果显示，训练数据年份和测试数据年份分别为 2004 — 2016 年和 2017 年、2004 — 2017 年和 2018 年、2004 — 2018 年和 2019 年、2004 — 2019 年和 2020 年，模型的 RMSE 和 MAE 值分别为 9.731 和 6.224、14.877 和 10.168、8.341 和 4.940、8.467 和 4.899，RMSE 和 MAE 值在 2017、2019 和 2020 年均＜10，2018 年的误差略大。河北省 2017 — 2020 年麻疹月发病预测值结果见表 2，实际发病趋势与预测趋势对比情况见图 2，结果显示，河北省 2017 — 2020 年模型预测的麻疹发病趋势与实际发病趋势基本保持一致。

表 2　河北省 2017 — 2020 年麻疹月发病实际值与预测值

Tbl. 2　Reported and predicted monthly number of measles cases by year in Hebei province: 2017 – 2020

年份	指标	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
2017	实际值	17	20	59	40	49	30	12	11	19	9	7	14
	预测值	16	25	47	86	77	54	17	10	10	10	10	12
2018	实际值	48	25	26	36	22	25	19	6	5	13	7	9
	预测值	11	19	44	43	48	31	6	6	6	6	7	8
2019	实际值	6	6	17	36	28	14	8	7	2	1	1	4
	预测值	9	10	15	21	19	19	12	3	3	3	3	3
2020	实际值	42	5	4	2	1	4	1	1	2	1	0	4
	预测值	16	15	4	3	4	3	7	5	3	4	4	3

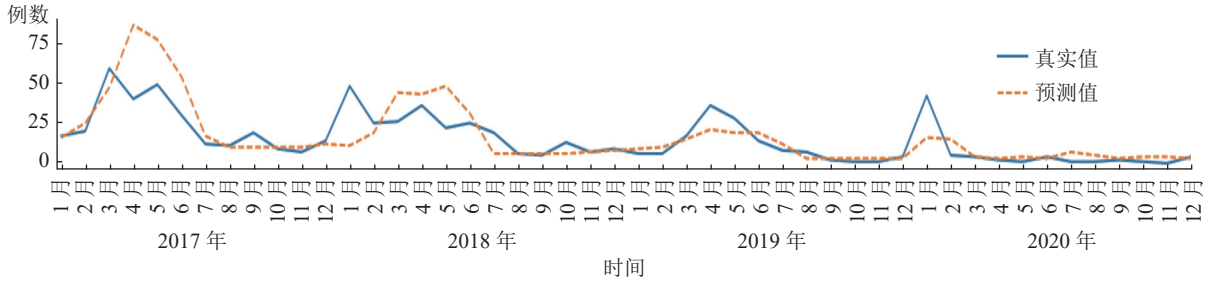

图 2　基于 LSTM 神经网络模型的河北省 2017 — 2020 年麻疹发病真实情况与预测情况比较图

Fig. 2　Chang trajectory in monthly number of measles cases reported and predicted with LSTM neural network model in Hebei province: 2017 - 2020

3 讨　论

传染病的防控一直是人类需要面对和解决的公共卫生难题，传染病发病趋势预测则是疫情防控的重要环节，对制定相关预防和控制措施具有重要的参考意义。近年来，人工智能成为疾病预测领域的研究热点，加入隐藏层的 ANN 可以逼近任意非线性映射，避开了复杂的参数估计程序，解决了变量之间的关系不能精确地用函数来表达的问题[8]。基于神经网络模型的预测算法在传染病预测领域取得了较线性模型更优的预测效果，使得其越来越受到欢迎[9－10]。

本研究基于 LSTM 神经网络模型探讨其在麻疹疫情发病趋势预测上的可行性。本研究视窗长度分析结果显示，视窗长度为 3 时模型的预测效果最好，提示河北省 2017 — 2020 年麻疹发病与近 3 个月的情况关系密切。因此，在分析麻疹疫情时，可重点关注近 3 个月的相关麻疹影响因素情况，例如，近 3 个月内的气候变化情况和人员流动性情况等。程宁等[11] 以不同视窗长度对数据进行了预处理，结果发现使用视窗长度分析能够有效提升 LSTM 神经网络模型的预测效果，与本研究结果一致。本研究应用 LSTM 神经网络模型对河北省 2017 — 2020 年麻疹发病趋势进行了拟合预测，结果显示模型预测的趋势与实际趋势基本一致，且 RMSE 和 MAE 值除 2018 年外均＜10（即预测全省每月发病数与实际发病数的差值平均＜10 例），提示 LSTM 神经网络模型的预测效果较好，存在一定的实际应用价值。韩天齐等[12] 和倪茹玉等[13] 分别基于 LSTM 神经网络模型对中国麻疹发病情况进行了预测，研究结果均表明 LSTM 神经网络模型在麻疹发病预测上具有一定的适用性，与本研究结果一致。本研究模型评价指标 RMSE 和 MAE 的值均小于李顺勇等[14] 采用 LSTM 神经网络模型进行肺结核发病数预测时的 RMSE 值（9 287.70）和 MAE 值（6 851.17），这可能与本研究采用双流 LSTM 神经网络模型有关。与既往研究不同，本研究采用全局趋势预测

模型和局部数量预测模型相结合的方式进行预测，此方法能综合考虑疾病的长期发生规律和近期发生情况对疾病的影响，提升了模型的预测效果。

综上所述，LSTM 神经网络模型在河北省麻疹疫情发病趋势中的预测效果较好，可用于麻疹发病趋势研判和风险评估，为科学预测麻疹发病情况提供统计学方法的参考。但本研究预测的河北省 2018 年麻疹发病趋势的误差略大，这可能与 2018 年整体发病趋势与往年不同有关，提示 2018 年河北省麻疹的发病可能受一些因素的影响，同时也表明本研究所提出的模型算法架构存在一定的优化空间。麻疹属于呼吸道传染病，其发生与发展受多种因素（人群、气象、政策等）的影响。例如 2010 年河北省在全省范围内对 < 5 岁适龄儿童开展含麻疹类疫苗查漏补种、应急接种等遏制麻疹流行“集中活动”以及对全省 8 月龄～14 岁儿童开展含麻疹类疫苗的强化免疫活动，2011 — 2012 年河北省麻疹发病数明显下降，但 2013 年出现疫情反弹[15 – 16]；2020 年受新型冠状病毒感染疫情的影响，公众坚持佩戴口罩，大大减少了呼吸道传染病的传播，2020 年河北省麻疹发病数较低，麻疹的发生与流行受到影响，这都可能在一定程度上影响模型的预测效果。随着近期新型冠状病毒感染疫情防控措施的放开，人群聚集性逐渐增加，呼吸道传染病的发病率出现上升趋势，因此对于麻疹疫情的防控仍不容松懈。下一阶段研究可考虑将混杂因素纳入模型建立多因素预测模型以更加全面准确地对麻疹发病趋势进行预测，从而采取更有针对性的措施减少麻疹的发生，保护人群健康。

参考文献

[1] 李兰娟, 任红, 高志良, 等. 传染病学 [M]. 9 版. 北京: 人民卫生出版社, 2018: 81.

[2] 刘倩倩, 唐林, 温宁, 等. 中国 2020 年麻疹流行病学特征 [J]. 中国疫苗和免疫, 2022, 28(2): 135 – 139.

[3] 赵子平, 许可, 吴莹, 等. 基于深度学习的猩红热流行趋势预测模型研究 [J]. 南京医科大学学报: 自然科学版, 2022, 42(2): 252 – 257, 263.

[4] 赵永翼, 王菲, 申莹. 基于长短期记忆网络的 COVID-19 疫情趋势序列分析预测 [J]. 沈阳师范大学学报: 自然科学版, 2020, 38(6): 525 – 531.

[5] 陈春艳, 陈亿雄, 赵执扬, 等. SARIMA 模型和 LSTM 神经网络在预测深圳市宝安区手足口病疫情中的应用 [J]. 山西医科大学学报, 2022, 53(10): 1302 – 1307.

[6] 杨寒雨, 赵晓永, 王磊. 数据归一化方法综述 [J]. 计算机工程与应用, 2023, 59(3): 13 – 22.

[7] 马停停, 冀天娇, 杨冠羽, 等. 基于长短时记忆神经网络的手足口病发病趋势预测 [J]. 计算机应用, 2021, 41(1): 265 – 269.

[8] 杨文静, 杜然然, 吕章艳, 等. 人工智能在疾病预测研究中可视化分析 [J]. 中国公共卫生, 2021, 37(5): 871 – 874.

[9] 陈亿雄, 李苑, 刘小明, 等. 长短记忆神经网络在流行性感冒暴发预测中的应用 [J]. 江苏预防医学, 2019, 30(6): 622 – 625.

[10] 刘振球, 严琼, 左佳鹭, 等. EMD-BP 神经网络在传染病发病趋势和预测研究中的应用 [J]. 中国卫生统计, 2018, 35(1): 152 – 155.

[11] 程宁, 丁长松, 高婉卿, 等. 基于时间窗长短期记忆模型分析新型冠状病毒肺炎疫情 [J]. 中华疾病控制杂志, 2021, 25(5): 577 – 582.

[12] 韩天齐, 宋波. 基于 LSTM 神经网络的麻疹发病率预测 [J]. 电脑与电信, 2018(5): 54 – 57.

[13] 倪茹玉, 胡婉, 张恒川, 等. ARIMA 乘积季节模型与 LSTM 神经网络模型对我国麻疹发病数预测效果的比较 [J]. 现代预防医学, 2023, 50(1): 177 – 182.

[14] 李顺勇, 张钰嘉. LSTM 和 Prophet 模型在肺结核发病数预测中的应用 [J]. 河南科学, 2020, 38(2): 173 – 178.

[15] 丛艳丽, 张富斌, 张振国, 等. 2009 — 2012 年河北省麻疹流行特征及消除麻疹策略 [J]. 职业与健康, 2014, 30(23): 3408 – 3411.

[16] 刘曙光, 王立芹, 刘岩, 等. 河北省麻疹疫情时间序列的预测和预警分析 [J]. 中国卫生检验杂志, 2015, 25(17): 2954 – 2956.