DOI:10.16781/j.0258-879x.2018.02.0226

·短篇论著·

极端学习机模型在张家口市手足口病发病率预测中的应用

杨　旭[1]，张　倩[2*]
1. 张家口市疾病预防控制中心，张家口 075000
2. 衡水市疾病预防控制中心，衡水 053000

[摘要]　目的　探讨极端学习机（ELM）模型在手足口病发病率预测中的应用，并与神经网络模型进行比较。**方法**　收集 2008 年 5 月至 2017 年 7 月张家口市手足口病月发病率资料，并组成具有 111 个数据的时间序列，随机选择数据集中 75% 的数据进行学习建模，剩余 25% 作为预测的检验数据，以对 2 种模型的预测效果进行验证。**结果和结论**　ELM 学习的平均相对误差（MRE）为 0.05，预测的 MRE 为 0.07；神经网络学习的 MRE 为 0.09，预测的 MRE 为 0.12。ELM 模型的学习效果和预测效果优于神经网络模型，可以提高预测的精度，具有较高的实用价值。

[关键词]　手足口病；极端学习机；神经网络；发病率；预测
[中图分类号]　R 195.4　　**[文献标志码]**　A　　**[文章编号]**　0258-879X(2018)02-0226-05

Application of extreme learning machine model in prediction of hand-foot-and-mouth disease incidence in Zhangjiakou city

YANG Xu[1], ZHANG Qian[2*]
1. Center for Disease Control and Prevention of Zhangjiakou, Zhangjiakou 075000, Hebei, China
2. Center for Disease Control and Prevention of Hengshui, Hengshui 053400, Hebei, China

[Abstract]　Objective　To explore the application of extreme learning machine (ELM) model in predicting the incidence of hand-foot-and-mouth disease, and to compare the difference between ELM model and neural network model. **Methods**　The monthly incidence data of hand-foot-and-mouth disease from May 2008 to Jul. 2017 in Zhangjiakou were collected and formed a time series with 111 data. To validate and evaluate the prediction performance of the two models, 75% of the randomly selected dataset were used to train model and the remaining 25% were used as testing data for prediction. **Results and conclusion**　The mean relative errors (MREs) of learning and prediction based on ELM model were 0.05 and 0.07, respectively. The MREs of learning and prediction based on neural network model were 0.09 and 0.12, respectively. The learning and prediction effects of ELM model are better than neural network model. It can improve the accuracy of prediction and has high application value.

[Key words]　hand-foot-and-mouth disease; extreme learning machine; neural network; morbidity; prediction
[Acad J Sec Mil Med Univ, 2018, 39(2): 226-230]

手足口病是由多种肠道病毒感染引起的传染病，多发于儿童群体[1]。目前研究发现，引起手足口病的病毒种类多达 20 余种，其中以肠道病毒 71 型（EV71）[2]和柯萨奇 A16 型[3]最为常见，前者可引起重症病。手足口病危害大，传播快，其防治任重道远。张家口市地处华北平原，是手足口病多发地区，发病率忽高忽低，所以有必要建立精确可靠的手足口病发病率监测预测系统，为手足口病防治工作奠定基础。

传染病发病率预测模型主要有灰色预测模型[gray forecast model，GM（1，1）][4]、自回归综合移动平均模型（autoregressive integrated moving average model，ARIMA）[5]、支持向量机模型[6]和神经网络（neural network，NN）模型等。GM（1，1）[7]和 ARIMA 模型[8]通过对数据进行变换处理，形成平稳的时间序列，由于手足口病发病率

[收稿日期]　2017-09-14　　**[接受日期]**　2017-10-10
[基金项目]　衡水市科技计划自筹经费项目(2016014001Z). Supported by Science and Technology Self-financing Project of Hengshui (2016014001Z).
[作者简介]　杨　旭，主管医师．E-mail: 271359212@163.com
*通信作者(Corresponding author). Tel: 0318-2811026, E-mail: hsjkchenchao@163.com

复杂多变，结果误差较大；支持向量机模型虽然通过核函数可以处理非线性问题，但是对于回归问题参数较多，精度不高；NN[9]模型学习效率较低，难以取得全局最优解。

极端学习机（extreme learning machine，ELM）是一种基于广义 NN 逼近原理建立的新型 NN[10-11]，它在随机给定神经元输入权值与偏差的基础上，将传统 NN 训练问题转化为求解线性方程组，并根据广义逆矩阵理论，以解析方式直接计算出其输出权值的最小二乘解，从而完成网络训练过程。相比于传统 NN，ELM 由于具有计算原理简单、训练速度快和泛化能力强的优点，是解决时间序列预测问题的有力工具。但其在发病率预测领域极端学习的应用少见报道。本研究将 ELM[12]用于手足口病发病率预测，并结合实际数据，以预测误差作为目标函数选取最佳参数进行学习，以期进一步提高模型精度，为张家口市手足口病的预防与控制工作提供参考依据。

1　ELM 模型

1.1　ELM 回归原理　假设训练集为 $\{(X_k,Y_k)\}_{k=1}^{N}$，且包含 L 个隐层神经元 $f(.)$ 的 ELM 回归模型如下：

$$\begin{cases}\sum_{i=1}^{L}\beta_i f(X_1;a_i,b_i)=Y_1\\ \sum_{i=1}^{L}\beta_i f(X_2;a_i,b_i)=Y_2\\ \vdots\\ \sum_{i=1}^{L}\beta_i f(X_N;a_i,b_i)=Y_N\end{cases}\qquad(1)$$

（1）式中 N 为训练样本的数量，X_k、Y_k 分别为第 k 个训练样本的特征向量和响应值，即输入和输出，β_i 为连接第 i 个神经元的输出权值，$f(.)$ 为神经元激活函数，$\alpha_i=[\alpha_{i1}\ \alpha_{i2}\ \cdots\ \alpha_{ih}]$为连接第 i 个神经元的输入权值，b_i 为第 i 个神经元的偏差，将（1）式改写为矩阵的形式：

$$H\beta=Y\qquad(2)$$

（2）式中，H 为神经元矩阵，Y 为输出向量，β 为输出权值。其值分别为：

$$H=\begin{bmatrix}f(X_1;a_1,b_1) & f(X_1;a_2,b_2) & \cdots & f(X_1;a_L,b_L)\\ f(X_2;a_1,b_1) & f(X_2;a_2,b_2) & \cdots & f(X_2;a_L,b_L)\\ \cdots & \cdots & & \cdots\\ f(X_N;a_1,b_1) & f(X_N;a_2,b_2) & \cdots & f(X_N;a_L,b_L)\end{bmatrix}=\begin{bmatrix}h_1\\ h_2\\ \cdots\\ h_N\end{bmatrix}$$

$$Y=\begin{bmatrix}Y_1 & Y_2 & \cdots & Y_N\end{bmatrix}^T$$

$$\beta=\begin{bmatrix}\beta_1 & \beta_2 & \cdots & \beta_L\end{bmatrix}^T$$

由（2）式可知，ELM 的训练过程等价于求解线性方程。

1.2　参数的选择　ELM 模型中涉及的参数有嵌入维数、隐层神经元数量、输入权值、神经元偏置和输出权值。当嵌入维数和隐层神经元数量确定后，输入权值 α_i 和神经元偏置 b_i 在[－1，1]内随机取值，输出权值 β 利用求取矩阵的 MP 逆获取最小二乘解。

嵌入维数和隐层神经元数量根据预测误差进行选取，首先确定嵌入维数的取值范围为[1, 10]，隐层神经元数量取值范围为[1, 150]。其次将训练样本按数量均分为 10 等份，任意取选 9 份作为训练数据，1 份作为验证数据，依次进行训练，并用验证数据计算预测误差，共进行 10 次，记录每个参数下的平均预测误差，最后根据平均预测误差的变化规律，选择合适的嵌入维数和隐层神经元数量。

2　资料和方法

2.1　资料来源　2008 年 5 月至 2017 年 7 月手足口病月发病率数据来自张家口市传染病网络直报系统，人口数据来源于全国第六次人口普查公布结果。见图 1。

2.2　基于数据的模型求解与评价

2.2.1　基于 ELM 的张家口市手足口病月发病率预测模型　一般来说，嵌入维数和隐层神经元数量越多学习的精度越高，但是这种情况容易出现过拟合现象，严重影响模型的泛化能力。所以在允许的预测误差范围内，选择嵌入维数和隐层神经元数量较少的组合，步骤如下：（1）嵌入维数和神经元数量取值范围设定为 [1，10]×[1，150]，用留一法交叉验证并计算每个参数组合下的平均预测误差。（2）计算最小的平均误差值，记为 min$e$。（3）在平均误差取值范围[1.2×min$e$，1.3×min$e$]内寻找嵌入维数 h 和隐层神经元数量 L 之和最小的组合，作为最佳的 ELM 模型参数。

根据上述步骤，选择嵌入维数为 5，隐层神经元数量为 32。对比的 NN 模型采用相同的嵌入维数和隐层神经元数量，激活函数采用 Sigmoid 函数，采用梯度下降法进行训练，学习速率为 0.05，最大训练次数为 5 000 次，设置平均相对误差（mean relative error, MRE）目标界限为 0.065。

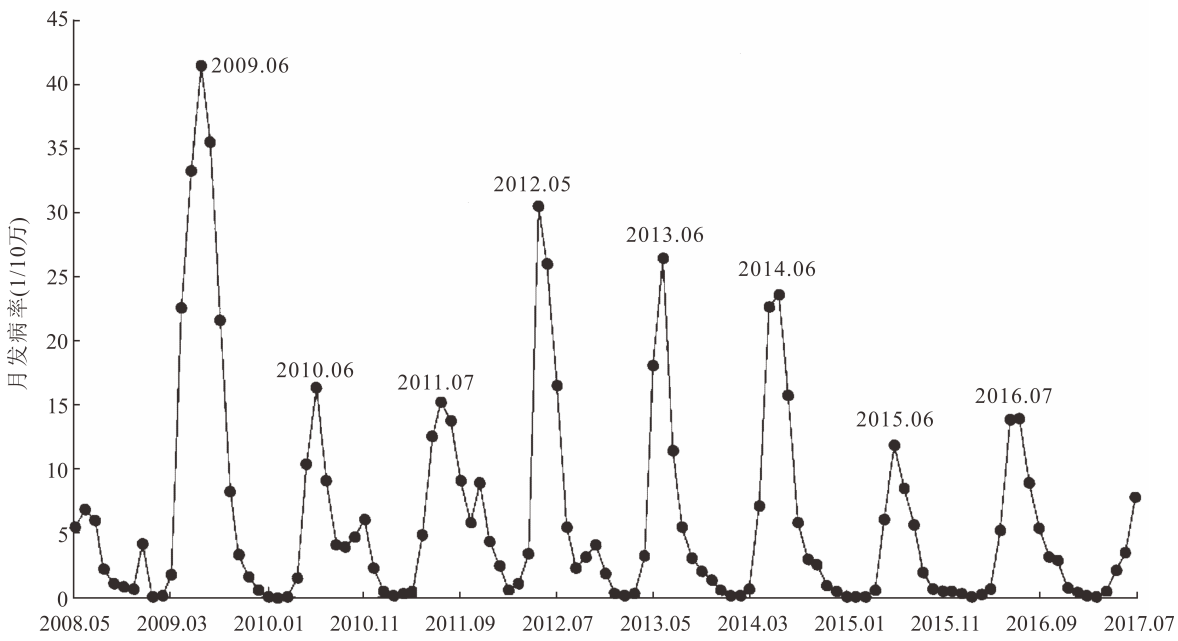

图 1　张家口市手足口病月发病率时序图

2.2.2　模型评价指标　模型学习效果使用后验差检验方法[13]进行评价，即对模型拟合的残差进行统计学分析。首先计算发病率序列的均值（$\bar{x}$）和方差（S_0），如下：

$$\bar{x}=\frac{1}{N}\sum_{i=1}^{N}x(i)$$

$$S_0=\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}[x(i)-\bar{x}]^2}$$

其次计算预测残差的均值（$\bar{\Delta}$）和方差（S_1）：

$$\Delta(i)=\left|x(i)-\hat{x}(i)\right|$$

$$\bar{\Delta}=\frac{1}{N}\sum_{i=1}^{N}\Delta(i)$$

$$S_1=\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}[\Delta(i)-\bar{\Delta}]^2}$$

由此计算标准差的比值 $C=S_1/S_0$ 和决定系数 $R^2=|S_0^2-S_1^2|/S_1^2$　C 值越小、R^2 值越大，说明模型学习的精度越高。

学习的精度高不能说明模型的预测效果好，预测效果主要体现为模型的泛化能力，即对未知数据的适应性。采用 MRE 对模型预测效果进行评价。MRE 越小说明模型预测精度越高。计算公式如下：

$$\mathrm{MRE}=\frac{1}{N}\sum_{k=1}^{N}\left|\frac{\hat{x}(i)-x(i)}{x(i)}\right|$$

式中，$\hat{x}(i)$ 为第 i 个预测值，$x(i)$ 为第 i 个真值。采用相对误差（relative error，RE）对单次预测结果进行评价，计算公式如下：

$$\mathrm{RE}=\left|\frac{\hat{x}(i)-x(i)}{x(i)}\right|\times 100\%$$

2.3　统计学处理　用 Matlab 2014a 软件建立 ELM 和 NN 模型，并对学习效果和预测效果进行定量分析。

3　结　果

3.1　预测结果对比　在构成的手足口病月发病率样本集 $\{(X_k,Y_k)\}_{k=1}^{N}$ 中随机选择 75% 训练模型，剩余 25% 作为验证数据。预测数据是指根据真实的手足口病月发病率预测下一个月的发病率。由图 2 可知，拟合和预测数据基本与真实数据吻合，真值均处于预测置信区间范围内，说明张家口市手足口病月发病率是可以预测的。

以 S_1、C、R^2 和 MRE 4 个指标对模型进行评价，ELM 模型在学习方面性能优于 NN 模型。ELM 模型预测的 MRE 也优于 NN 模型。见表 1。

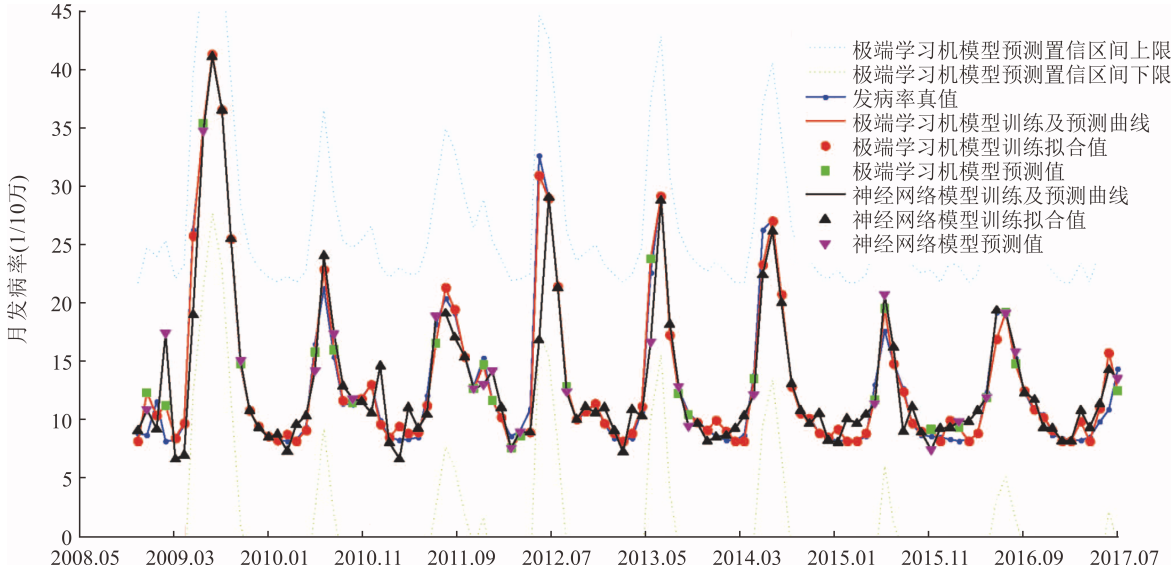

图 2　极端学习机模型和神经网络模型训练预测曲线

表 1　极端学习机（ELM）和神经网络（NN）模型对比

模型	模型训练评价指标				预测评价指标
	S_1	C	R^2	MRE	MRE
ELM	0.25	0.12	32.31	0.05	0.07
NN	0.54	0.38	15.86	0.09	0.12

S_1: 预测残差的均方差; C: 后验差比值; R^2: 决定系数; MRE: 平均相对误差

3.2　多步预测结果对比　多步预测是利用当前张家口市手足口病月发病率数据，逐步向后预测多个月的发病率。本研究将 2017 年 4 月至7 月的手足口病月发病率作为验证数据，在其余手足口病月发病率数据中选择 75% 作为学习数据集。多步预测结果如表 2 所示。

多步预测效果主要反映了模型的泛化能力和对未知数据的解释能力，由表 2 可知 ELM 的多步预测结果与真实值比较接近。而 NN 第 2 步预测的结果偏离真实值较大，后续预测结果与真实值相差较大。从对未知数据的预测能力来看，ELM 模型比 NN 模型较为理想。

表 2　ELM 和 NN 模型用于 2017 年 4 月至 7 月手足口病月发病率多步预测结果对比

月份	真实值	ELM 模型			NN 模型		
		预测值	相对误差 (%)	MRE	预测值	相对误差 (%)	MRE
4 月	0.46	0.48	4.35	0.13	0.54	17.39	0.29
5 月	2.09	2.46	17.70		2.92	39.71	
6 月	3.46	3.87	11.85		4.68	35.26	
7 月	7.79	9.08	16.56		9.69	24.39	

ELM: 极端学习机; NN: 神经网络; MRE: 平均相对误差

4　讨　论

自 2008 年以来，张家口市手足口病发病率呈现先上升后下降、波动逐渐规律、最后趋于稳定的状态。众所周知，手足口病发病率受到错综复杂的因素影响，包括流动人口、卫生条件、生活环境、气候等，由于相关资料的检测和收集不充分，又难以判断与手足口病的关系，所以本研究假定错综复杂的影响因素体现在历史数据中，因此对历史发病率建立模型以预测未来数据。

将手足口病月发病率作为时间序列进行处理，因受多因素的影响，数据复杂多变，难以用解析的函数对其进行逼近，所以传统的 GM（1，1）和 ARIMA 模型拟合精度不高，而且建立模型比较复杂，误差较大。而后发展了 NN 和核函数的方法，逼近非线性函数的能力大幅提高，效果比较显

著，但是 NN 模型采用误差反向传播的机制对模型参数进行求解，隐函数为非线性函数，其导数只在中心值的附近呈现近似的线性关系，并且线性关系的范围会随中心值的不同差异较大，所以选择不同的优化步长对模型参数的求解影响较大，使 NN 难以收敛或收敛于局部最小解的情况时有发生，在模型建立过程中参数求解过程比较复杂、耗时，且精度没有 ELM 高。ELM 模型本质上为一单隐层 NN，继承了 NN 逼近非线性函数的优点，但其训练机制与 NN 不同，利用求神经元矩阵的 MP 逆获取最小二乘解，学习速度快且精度高。

本研究利用留一折的方法计算不同参数下的平均预测误差，依据此误差矩阵获取最佳的嵌入维数和隐层神经元数量，所建立的 ELM 模型不仅训练结果较好，预测精度也得到了提高。训练精度仅为为 NN 模型的一半（0.05 vs 0.09），而一步预测精度也比 NN 高出约 70%（0.07 vs 0.12）。从多步预测结果来看预测值与真实值较为接近。

本研究显示，ELM 模型适合张家口市手足口病发病率的拟合与预测，这对指导公共卫生人员依据疫情提前做好防控工作并制定有效防控策略具有重大意义。

[参 考 文 献]

[1] 单宝英. 不同病原所致手足口病临床特征分析[J]. 中国现代医学杂志,2016,26:119-122.

[2] 邵勤,刁玉巧. EV71 感染手足口病患儿血清内皮素-1 水平测定及其临床意义[J]. 中国现代医学杂志,2013,23:91-93.

[3] 刘莹莹,于秋丽,苏通,赵文娜,谢赟,齐顺祥,等. 2011—2015 年河北省手足口病流行特征及病原特征分析[J]. 中华疾病控制杂志,2017,21:151-155.

[4] 杨永利,毛赛彩,薛源,田翔宇,施学忠. GM(1,1)和趋势外推模型在我国艾滋病发病率预测中的应用[J]. 中国卫生统计,2014,6:952-954.

[5] 夏菁,张华勋,林文,裴速建,孙凌聪,董小蓉,等. ARIMA 模型在疟疾发病率预测中的应用[J]. 中国血吸虫病防治杂志,2016,2:135-140.

[6] 徐学琴,王瑾瑾,马晓梅,刘颖,杨梦利,闰国立,等. 基于支持向量机模型的河南艾滋病发病率预测[J]. 中国现代医学杂志,2017,12:93-95.

[7] 张倩,陈超. 改进的 GM(1,1)模型在衡水市乙肝发病率预测中的应用[J]. 现代预防医学,2017,44:1925-1928,1937.

[8] 杨召,叶中辉,赵磊,薛庆元,梁淑英,王重建. ARIMA-BPNN 组合预测模型在流感发病率预测中的应用[J]. 中国卫生统计,2014,31:16-18.

[9] 曾海燕,解合川,任钦,张兴裕,李晓松. 径向基函数神经网络在甲型病毒性肝炎发病率预测中的应用初探[J]. 现代预防医学,2013,24:4489-4492.

[10] HUANG G B, ZHU Q Y, SIEW C K. Extreme learning machine: a new learning scheme of feedforward neural networks[J]. Proc Int J Conf Neural Netw, 2004(2): 985-990.

[11] HUANG G B, ZHU Q Y, SIEW C K. Extreme learning machine: theory and applications[J]. Neurocomputing, 2006, 70(1/2/3): 489-501.

[12] 何星,王宏力,陆敬辉,姜伟. 基于优选小波包和 ELM 的模拟电路故障诊断[J]. 仪器仪表学报,2013,34:2614-2619.

[13] 王永斌,郑瑶,柴峰,李向文,田珍榛,袁聚祥. 基于周期分解的 ARIMA 模型在甲肝发病率预测中的应用[J]. 现代预防医学,2015,42:4225-4229.

[本文编辑] 尹　茶