时间序列分析与机器学习方法在预测肺结核发病趋势中的应用*

付之鸥[1,2]　周　扬[3]　陈　诚[3]　郑洪伟[4]　宋　伟[2]　李　苑[5]　陆　伟[3△]　彭志行[1△]

【提　要】 **目的**　研究时间序列分析与机器学习方法在预测肺结核发病趋势中的应用。**方法**　使用江苏省 2009 – 2018 年肺结核月度发病率数据,构建时间序列分析(ARIMA 模型)、机器学习方法(支持向量回归(SVR)、BP 神经网络)和两者的组合方法(ARIMA-SVR、ARIMA-BPANN)共 5 种预测模型,分析评价各模型预测性能。**结果**　ARIMA、SVR、BP 神经网络、ARIMA-SVR、ARIMA-BPANN 均方误差分别为 0. 0356、0. 0364、0. 0384、0. 0329、0. 0336;平均相对误差分别为 5. 76%、6. 19%、6. 20%、5. 63%、5. 70%。**结论**　时间序列分析优于机器学习方法,而二者组合模型预测效果优于单独方法,ARIMA-SVR 模型在江苏省肺结核发病趋势预测分析中具有较好的应用价值。

【关键词】 时间序列分析　机器学习　肺结核　预测

Application of Time Series Analysis and Machine Learning Methods in Predicting the Incidence of Tuberculosis

Fu Zhiou, Zhou Yang, Chen Cheng, et al (*Department of Epidemiology and Health Statistics, Public Health College, Nanjing Medical University* (211166), *Nanjing*)

【**Abstract**】 **Objective**　To study the application of time series analysis and machine learning in predicting the incidence trend of tuberculosis. **Methods**　Using the monthly incidence data of tuberculosis in Jiangsu Province from 2009 to 2018, five prediction models were constructed, including time series analysis (ARIMA), machine learning (support vector regression, BP neural network) and their combination methods (ARIMA-SVR, ARIMA-BPANN). The prediction performance of each model was analyzed and evaluated. **Results**　The mean square errors of ARIMA, SVR, BP neural network, ARIMA-SVR and ARIMA-BPANN were 0. 0356, 0. 0364, 0. 0384, 0. 0329 and 0. 0336 respectively, and the average relative errors were 5. 76%, 6. 19%, 6. 20%, 5. 63% and 5. 70%, respectively. **Conclusion**　Time series analysis is superior to machine learning method, and the combined model is superior to the single method. ARIMA-SVR model has a good application value in the prediction and analysis of tuberculosis incidence trend in Jiangsu Province.

【**Key words**】 Time series analysis; Machine learning; Tuberculosis; Prediction

2017 年世界范围内有 1000 万结核病新发病例,造成 130 万人死亡,中国在 30 个高结核病负担国家中排名第二,占世界病例的 9%[1]。过去二十年来,我国的结核病防控取得了较大进展,但结核病依然是当前威胁居民健康的主要公共卫生问题[2-3]。为积极响应 WHO 提出的“2035 年终止结核病流行”这一目标[4-5],改善现有结核病疫情监测评估体系,实现“精准预防”成为了当前防控工作的关键[6-7]。本文以江苏省 2009 – 2018 年月度肺结核疫情资料为基础,分别使用传统时间序列分析方法(ARIMA 模型)、机器学习方法(支持向量回归(SVR)、BP 神经网络)和二者的组合方法(ARIMA-SVR、ARIMA-BPANN),建立 5 种肺结核发病率的预测模型,检验评价各方法的预测性能,旨在为江苏省结核病预测预警体系建设及防治策略的制定提供科学依据和有效参考。

对象与方法

1. 资料来源

江苏省 2009 – 2018 年肺结核发病资料来源于中国疾病预防控制信息系统——结核病管理信息系统;江苏省各年人口学资料来源于《江苏统计年鉴——2018》。

2. 研究方法

(1) ARIMA 模型

ARIMA 模型是时间序列分析预测方法中较常用的一种模型。该模型充分考虑了事物自身发展的延续性和周期性,目前已大量应用于肺结核的预测研究[8-9]。公式可简记为 ARIMA$(p,d,q)\times(P,D,Q)_S$。其中 p,d,q 分别表示自回归阶数、差分阶数和移动平均阶数;P,D,Q 分别表示季节自回归阶数、季节差分阶数、季节移动平均阶数,S 表示周期步长。建模步骤:

①数据平稳化:对 2009 – 2017 年江苏省月度肺结核发病率时间序列做单位根检验,若检验结果 $p>$

* 基金项目:十三五传染病科技重大专项(2018ZX10715 – 002);国家自然科学基金(81673275);深圳市科技创新计划项目(JCYJ20160427155352873);江苏省优势学科建设项目。

1. 南京医科大学公共卫生学院流行病与卫生统计学系(211166)
2. 哈尔滨市宾县卫生健康局(150400)
3. 江苏省疾病预防控制中心(210009)
4. 中国水利水电科学研究院(100048)
5. 深圳市宝安区疾病预防控制中心(518101)
△通信作者:陆伟,E – mail:jsjkmck @ 163. com;彭志行,E – mail:zhihangpeng@ 126. com

0.05,则序列不平稳,一般采用差分或对数变换方法使序列平稳。

②模型识别:根据平稳化后序列的自相关图(ACF)和偏自相关图(PACF)初步识别模型可能的阶数。

③模型参数估计和检验:尝试建立不同阶数组合的 ARIMA 模型。在各参数有统计学意义的基础上,若残差的 Ljung-Box 统计量 $p>0.05$ 则为白噪声,通过检验。再根据拟合优度值(R^2)、AIC 值最小原则综合确定最优模型。

④评价模型预测效果:比较 2018 年 1 月 - 2018 年 12 月肺结核实际发病率与预测发病率的均方误差(MSE)和平均相对误差(MAPE),验证模型预测效果。

(2)支持向量回归(SVR)

支持向量回归(support vector regression,SVR)是一种基于统计学理论的机器学习算法[10]。该算法采用结构风险最小化原则和核函数思想,将低维空间的非线性问题转化成高维空间的线性问题,通过求解凸二次规划问题,得到理论上的全局最优解。算法步骤:

①数据的归一化:将原始发病率数据归一化到[0,1]区间,归一化公式为:

$$p_i' = \frac{p_i - \min p_i}{\max p_i - \min p_i}$$

其中 p_i'为第 i 个月肺结核发病率 p_i 的归一化处理值。

②输入输出维数和样本集的设置:针对 2009 - 2017 年江苏省月度肺结核发病率数据,使用前 3 年历史同期肺结核发病率(输入值)预测下 1 年同期肺结核发病率(输出值)。将 2012 - 2016 年共 60 个样本作为训练集,2017 年的 12 个样本作为测试集。

③模型的建立和检验:研究采用的回归算法为 epsilon 型,核函数为径向基核函数(radial basis);使用网格搜索法(grid search)初步确定核函数(gamma)和成本(cost)的参数值,建立 SVR 模型。以测试集的实际输出值和期望输出值的 MSE 检验模型的有效性,并反复尝试,将参数调整至最佳。

④评价模型预测效果:采用最终确定的 SVR 模型预测 2018 年 12 个月的数据,经反归一化处理后,计算与真实值的 MSE 与 MAPE,验证模型预测效果。

(3)BP 神经网络

BP(back propagation)神经网络是一种按照误差逆向传播算法训练的多层前馈神经网络,自产生以来已成为传染病预测领域最为常用的机器学习算法之一[11]。该网络由输入层、隐含层和输出层组成,通过将正向传播产生的误差信号反向传导,进而修正每个隐含层的各神经元权重,使误差逐渐降低至精度要求。算法步骤:

①数据归一化:同 SVR 模型。

②输入输出维数和样本集的设置:同 SVR 模型。

③确定网络结构和函数设置:设定网络层数及各层节点数 3 - M - 1,其中输入层节点数为 3,输出层神经元数为 1,M 为隐含层神经元数,依次取 3 ~ 9;将隐含层、输出层神经元的激活函数和学习函数分别设定为:"tansig"、"purelin"、"ADAPTgdwm"。

④网络的建立和检验:用训练集训练网络,测试集测试网络,反复执行第③ ~ ④步,以测试集的实际输出值和期望输出值的 MSE 最小为标准,找出最佳网络。

⑤评价模型预测效果:同 SVR 模型。

(4)ARIMA-SVR 和 ARIMA-BPANN 组合模型

将时间序列分析和机器学习方法组合,用 ARIMA 模型拟合肺结核发病率时间序列的线性部分,分别使用 SVR 和 BP 神经网络拟合肺结核发病率时间序列的非线性部分,即 ARIMA 模型的残差,以期提高预测的精度和泛化能力。预测结果的表达式为:

$$\hat{P}_t = \hat{L}_t + \hat{e}_t$$

其中,$\hat{P}_t$ 为组合模型的预测结果,$\hat{L}_t$ 为 ARIMA 模型预测值,$\hat{e}_t$ 为 SVR 或 BP 神经网络对 ARIMA 模型残差的预测值。两种组合模型均参照前述时间序列分析及机器学习方法的操作步骤进行建模、拟合及预测。

3. 软件使用

数据整理以及模型的建立、评价均使用 R 3.6.0 软件实现。应用 forecast、tseries 包建立时间序列模型;应用 e1071 包建立 SVR 模型;应用 AMORE 包建立 BP 神经网络模型。

结　果

1. 江苏省肺结核流行特征概述

2009 - 2018 年江苏省共报告肺结核病例 357512 例,肺结核发病呈现出逐年下降的趋势(对数线性模型:$b=-0.06$,$t=-29.56$,$p<0.0001$)。2018 年发病人数较 2009 年下降 42.8%,年估计百分比变化 EAPC 为 -13.51%,下降速度高于全国平均水平[12]。年均发病率为 45.07 例/10 万。季节分解显示,肺结核发病率存在明显的季节性,每年 3 月、12 月高发,见表 1,图 1、图 2。

表 1　江苏省 2009 - 2018 年肺结核发病趋势

年份	发病例数	发病率(/10 万)
2009	46359	59.36
2010	43249	54.96
2011	39589	50.12
2012	39099	49.37
2013	36256	45.67
2014	35104	44.10
2015	32828	41.16
2016	30120	37.66
2017	28402	35.37
2018	26506	32.92

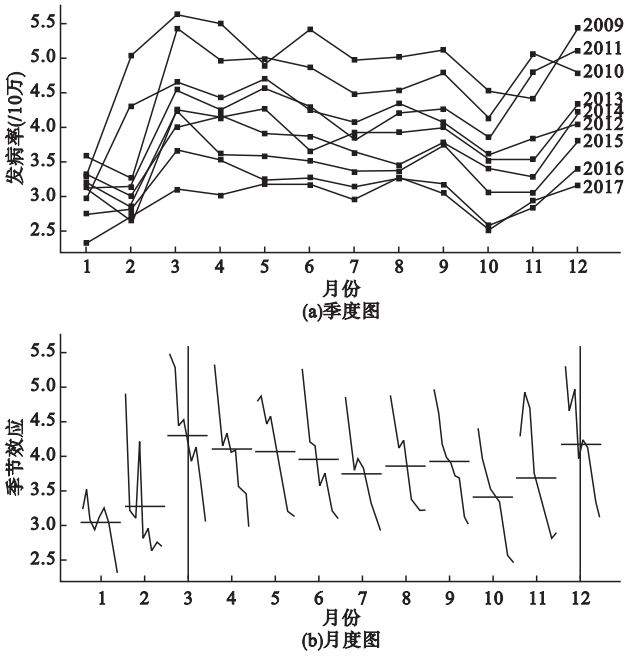

图 1　江苏省 2009－2018 年肺结核年度发病情况的季度图(a)和月度图(b)

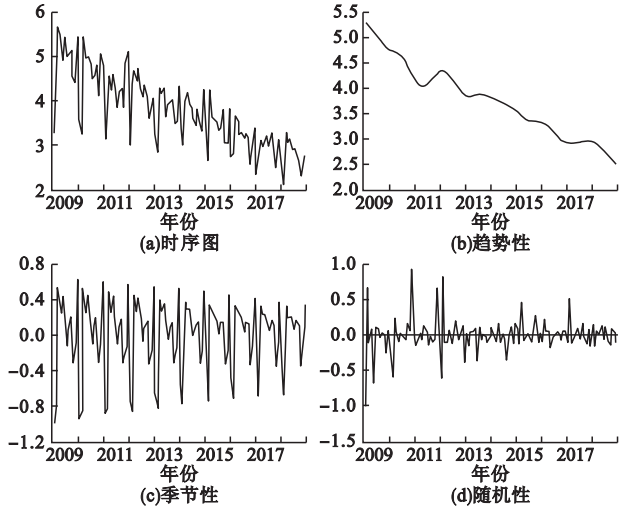

图 2　江苏省肺结核发病率季节分解

2. 基于肺结核发病率预测模型的实证研究

(1)ARIMA 模型

对原始肺结核发病率数据进行 1 阶差分和 12 步季节差分后即可保证序列平稳(图 3),ADF 单根性检验结果显示 Dickey-Fuller = －22.263(p <0.05)。将参数从低阶至高阶进行逐个试凑,分别建立 46 个 ARIMA 模型,根据前述模型筛选标准,最终选定 ARIMA(10,1,1)$(0,1,1)_{12}$的疏系数模型为最优模型。该模型系数均显著非零(表 2),对其残差值进行检验,L－B 统计量 Q = 21.15(p =0.3883 >0.05),为白噪声序列,R^2 =0.6860,AIC =66.41。拟合情况如图 4 所示。

表 2　ARIMA(10,1,1)$(0,1,1)_{12}$模型的参数检验

模型参数	回归系数	标准误	t 值	p
AR(10)	－0.3462	0.0764	－4.5306	<0.001
MA(1)	－0.9734	0.1151	－6.2737	<0.001
SMA(12)	－0.7678	0.1101	－6.9714	<0.001

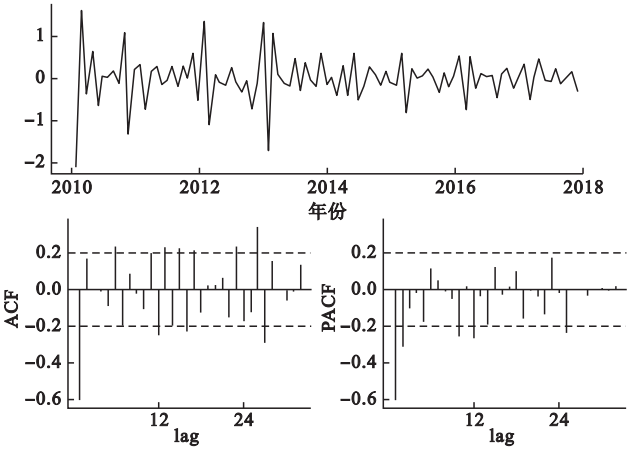

图 3　1 阶差分及 12 步季节性差分图及其 ACF 和 PACF 图

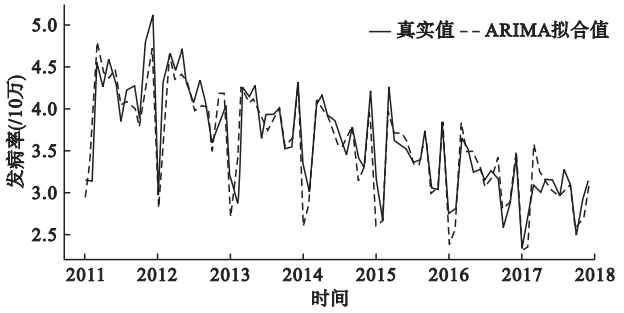

图 4　ARIMA(10,1,1)$(0,1,1)_{12}$模型拟合图

(2)支持向量回归(SVR)

使用 SVR 对肺结核发病率数据进行拟合,根据网格搜索结果初步选定参数值(图 5),经多次模拟检验发现,当 SVR 参数:cost = 1e + 04, gamma = 1e －05, epsilon =0.1 时,测试集 MSE 最小(MSE =0.0047)。拟合情况如图 6 所示。

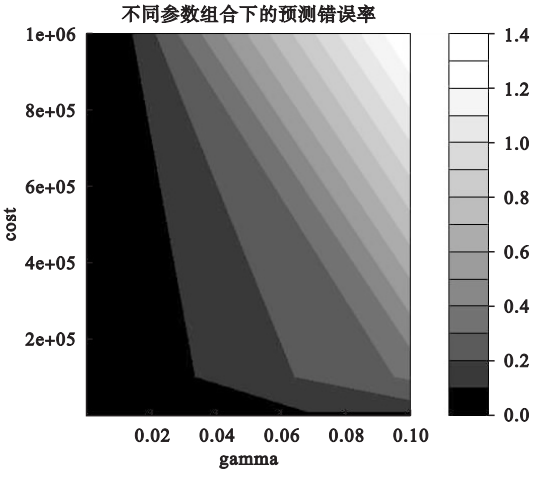

图 5　SVR 参数的网格搜索结果

(3)BP 神经网络

分别建立结构为 3－3－1、3－4－1、3－5－1、3－6－1、3－7－1、3－8－1、3－9－1 的 7 个 BP 神经网络,用肺结核发病率的训练样本集对每一个网络进行训练,测试集样本仿真(预测)结果显示(图 7),当网络结构为 3－4－1,最大训练次数为 1000 次,学习效率为 0.01 时,测试集 MSE 最小(MSE =0.0049)。拟

合情况如图 8 所示。

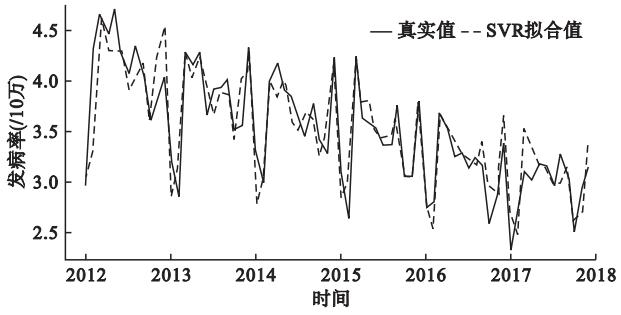

图 6 SVR 模型拟合图

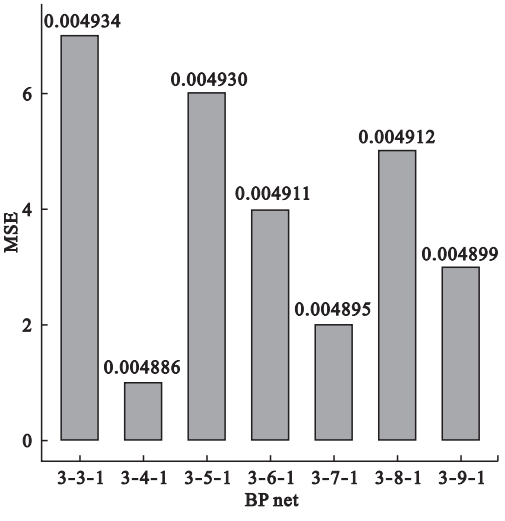

图 7 7 种 BP 神经网络结构的均方误差

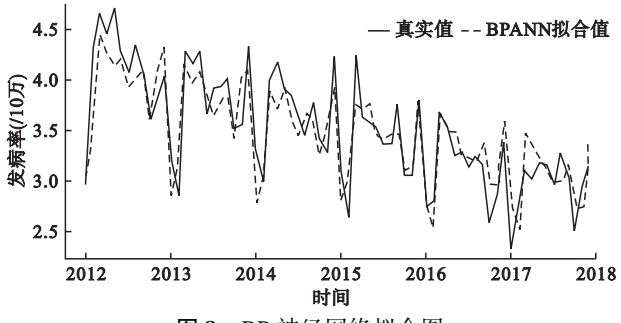

图 8 BP 神经网络拟合图

(4)ARIMA-SVR 组合模型

使用 SVR 对 2009 – 2017 年江苏省肺结核发病率 ARIMA(10,1,1)(0,1,1)$_{12}$模型残差值进行拟合。根据网格搜索结果初步选定参数值(图 9),经多次模拟检验发现,当残差 SVR 参数:cost = 10, gamma = 1e – 6, epsilon = 0.1 时,测试集 MSE 最小(MSE = 0.0111)。将 SVR 残差预测值加上 ARIMA 预测值得到 ARIMA-SVR 结核发病率的预测值。拟合情况如图 10 所示。

(5)ARIMA-BPANN 组合模型

分别建立结构为 3 – 3 – 1、3 – 4 – 1、3 – 5 – 1、3 – 6 – 1、3 – 7 – 1、3 – 8 – 1、3 – 9 – 1 的 7 个 BP 神经网络,用 ARIMA(10,1,1)(0,1,1)$_{12}$模型残差值的训练样本集对每一个网络进行训练,测试集样本仿真(预测)结果显示(图 11),当网络结构为 3 – 3 – 1,最大训练次数为 1000 次,学习效率为 0.01 时,测试集 MSE 最小(MSE = 0.0108)。将 BP 神经网络残差预测值加上 ARIMA 预测值得到 ARIMA-BPANN 肺结核发病率的预测值,拟合情况如图 12 所示。

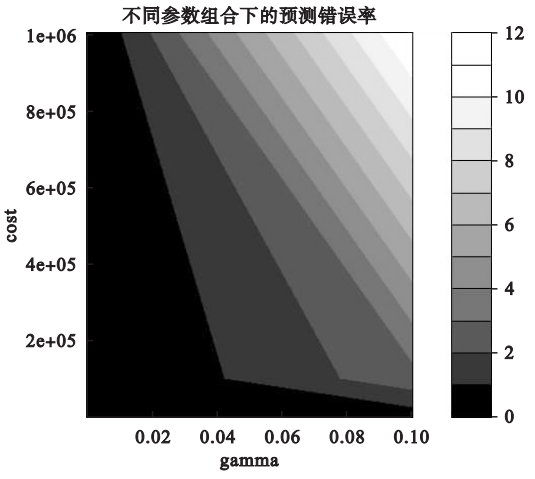

图 9 ARIMA-SVR 参数的网格搜索结果

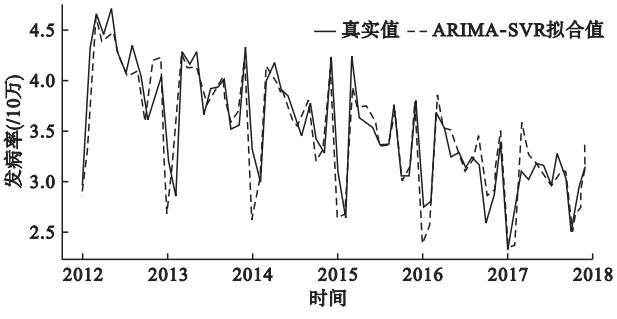

图 10 ARIMA-SVR 组合模型拟合图

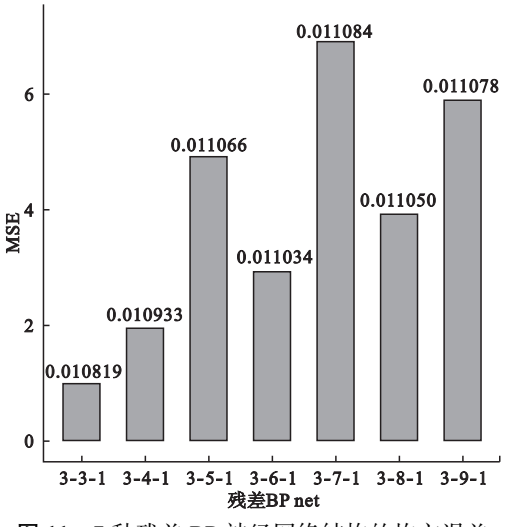

图 11 7 种残差 BP 神经网络结构的均方误差

3.5 个模型的比较

使用 5 种模型对 2018 年 12 个月的肺结核发病率数据进行预测,并计算误差,结果见表 3,图 13。ARIMA、SVR、BP 神经网络、ARIMA-SVR、ARIMA-BPANN 均方误差分别为 0.0356、0.0364、0.0384、0.0329、0.0336;平均相对误差分别为 5.76%、6.19%、6.20%、5.63%、5.70%。可见,组合模型

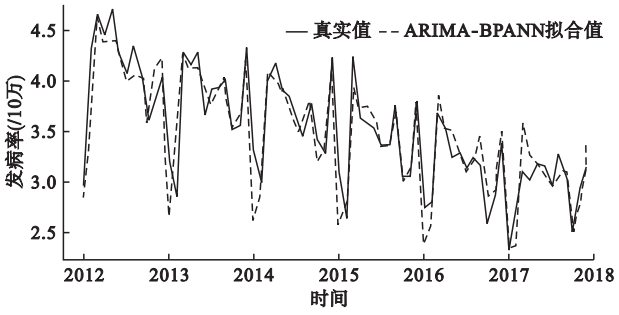

图 12　ARIMA-BPANN 组合模型拟合图

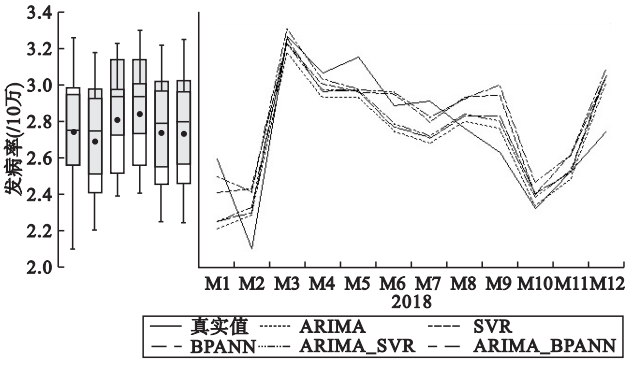

图 13　2018 年真实值和各模型预测值的比较

表 3　5 个模型泛化能力的比较

日期	实际值	预测值				
		ARIMA	SVR	BPANN	ARIMA-SVR	ARIMA-BP
2018 年 1 月	2.59	2.20	2.41	2.49	2.25	2.24
2018 年 2 月	2.10	2.28	2.42	2.41	2.32	2.29
2018 年 3 月	3.26	3.18	3.23	3.30	3.22	3.25
2018 年 4 月	3.06	2.93	3.00	3.02	2.97	2.96
2018 年 5 月	3.15	2.92	2.95	2.97	2.97	2.97
2018 年 6 月	2.88	2.74	2.94	2.96	2.78	2.77
2018 年 7 月	2.91	2.67	2.79	2.82	2.71	2.70
2018 年 8 月	2.76	2.79	2.93	2.92	2.84	2.83
2018 年 9 月	2.63	2.75	2.94	2.99	2.79	2.82
2018 年 10 月	2.32	2.33	2.39	2.46	2.38	2.40
2018 年 11 月	2.52	2.48	2.61	2.61	2.53	2.52
2018 年 12 月	2.75	3.01	3.05	3.09	3.05	3.05
MSE		0.0356	0.0364	0.0384	0.0329	0.0336
MAE		1.88	1.94	1.93	1.80	1.83
MPAE%		5.76	6.19	6.20	5.63	5.70

(ARIMA-SVR < ARIMA-BPANN)预测效果最好,时间序列分析(ARIMA)预测结果优于机器学习方法(SVR < BPANN)。

讨　论

本文利用江苏省 2009 – 2018 年月度肺结核发病率数据尝试建立不同类型、不同复杂程度的预测模型,包括 ARIMA 模型、SVR、BP 神经网络、ARIMA-SVR、ARIMA-BPANN。5 种模型均方误差和平均相对误差的均值分别为 0.354 和 5.9%,总体预测效果较好。我们发现 5 种模型无论是拟合值还是预测值,主要在 2 月与真实数据产生了一定偏差。从图 1(a)可以看出,每年 2 月的肺结核发病率趋势并不稳定,可能是因为 2 月正值中国春节前后,会出现大范围、高密度的人员流动与聚集,从而导致趋势拟合效果不佳。另外,本研究还对 SVR、BP 神经网络及其残差模型尝试了其他输入结构的建模:用前 12 个月的肺结核发病率预测下 1 个月的发病率。拟合预测效果均差于使用历史同期发病率的输入结构,可能是由于历史同期输入结构能在一定程度上反映肺结核发病率的短期趋势和同期效应,弱化了时间序列过程中的随机性,从而拟合预测效果更为理想。

从 3 种单独模型对江苏省 2018 年 12 个月肺结核发病率的预测情况可以看出,ARIMA 模型考虑了肺结核发病率时间序列的自相关性和季节规律,能够有效提取肺结核发病率数据中的线性信息,较好地预测了肺结核发展趋势,预测精度也是三者中最好的。Zheng YL 等人[13]曾在 ARIMA 模型的基础上,加入自回归条件异方差(ARCH)模型来处理时间序列的异方差性,成功预测了新疆结核病发病率。ARIMA 模型结构简单,可解释性强,对于总体趋势稳定,有明显规律的数据有很好的拟合能力。但其非线性映射能力较差;只能进行短期预测;不适合处理波动大、无规律的数据;且在纳入协变量时无法有效解决数据间的共线性问题,现阶段已不能满足大数据背景下的传染病预测需求。SVR 和 BPANN 等机器学习方法能够有效拟合非线性和不规则的数据,善于处理大量的协变量,拥有较强的学习和泛化能力[14-16]。Mollalo A 等人[17]曾收集 278 个探索变量,应用人工神经网络对美国结核病的地理分布进行建模,研究表明,空间建模中的机

器学习技术可应用于美国大陆的结核病发病率研究。当前机器学习方法仍存在模型结构、计算过程复杂;内部函数及参数设定没有统一的选定方法和标准,在一定程度上仍需凭借经验选取;其内部运行的机理不易理解,无法对拟合预测结果进行有效的流行病学解释等缺点。通过本研究可以看出,使用单独的肺结核发病率数据进行拟合,机器学习的预测效果较传统时间序列分析方法稍差。而时间序列分析和机器学习的组合方法则能充分发挥二者优点,弥补各自不足[18-19]。我们的研究结果也证实,时间序列分析和机器学习的组合模型能够充分提取肺结核发病率数据中的线性、非线性信息,在解释能力、学习能力、泛化能力以及预测精准性上都得到了明显提高。

这项研究有三个局限性:(1)肺结核的发生发展过程受公共卫生政策、人口流动、气象条件等多种因素影响,如果将这些因素纳入模型,在更高维度上拟合数据,可以为肺结核疫情的分析、评估及预测提供更多的思路。(2)本次研究范围为整个江苏省,即假定肺结核的传播条件及流行强度在整个区域是相同的。如果能够对整体数据进行空间范围的细化,建立各市、县的独立模型,其预测结果将更具指导意义。(3)虽然我们对 BP 神经网络和 SVR 及其残差模型的输入维度及参数进行了不同尝试,但仍未涵盖所有可实现的预测模型。在今后的研究中,应尽可能地对输入变量组合和参数进行全面测试,并选择具有最大解释能力的最佳模型。

综上,任何一种方法都是对实际系统的抽象和简化,难免带有局限性和片面性,并不能始终作为最好的方法。在实际应用中,应采用最合适的方法而非最复杂的方法。在选择模型前要全面分析数据结构,如果拟合效果不理想,应针对原有模型的不足进行改进或尝试不同方法的组合。对于本研究而言,时间序列分析和机器学习方法的组合模型预测结果最优;SVR 模型相较于 BP 神经网络,其结构和参数设置相对简单,能够更好地解决小样本和过度学习等问题[11,20],应用 ARIMA-SVR 模型对江苏省肺结核发病趋势进行预测是明智的选择。

参 考 文 献

[1] World Health Organization. Global tuberculosis report 2018. [2019-06-19]. https://www.who.int/tb/publications/global_report/zh/.

[2] 王黎霞,成诗明,陈明亭,等. 2010 年全国第五次结核病流行病学抽样调查报告. 中国防痨杂志,2012,34(8):485-508.

[3] Huynh GH, Klein DJ, Chin DP, et al. Tuberculosis control strategies to reach the 2035 global targets in China: the role of changing demographics and reactivation disease. BMC Med, 2015, 13(1): 88.

[4] Lonnroth K, Migliori GB, Abubakar I, et al. Towards tuberculosis elimination: an action framework for low-incidence countries. Eur Respir J, 2015, 45(4): 928-952.

[5] Houben R, Menzies NA, Sumner T, et al. Feasibility of achieving the 2025 WHO global tuberculosis targets in South Africa, China, and India: a combined analysis of 11 mathematical models. Lancet Glob Health, 2016, 4(11): e806-e815.

[6] Zhang G, Huang S, Duan Q, et al. Application of a hybrid model for predicting the incidence of tuberculosis in Hubei, China. PloS one, 2013, 8(11): e80969.

[7] Chae S, Kwon S, Lee D. Predicting Infectious Disease Using Deep Learning and Big Data. Int J Environ Res Public Health, 2018, 15(8): 1596.

[8] Wah W, Das S, Earnest A, et al. Time series analysis of demographic and temporal trends of tuberculosis in Singapore. BMC public health, 2014, 14: 1121.

[9] 谢赐福,王孝君,熊姿,等. SARIMA 模型在长沙市肺结核发病预测中的应用. 中国卫生统计,2018,35(6):859-862.

[10] Vapnik V. The Nature of Statistical Learning Theory-2nd Edition. Springer, 2000, 69-91.

[11] 刘文东,吴莹,艾静,等. BP 神经网络在痢疾发病趋势预测中的应用研究. 中国卫生统计,2012,29(6):801-804.

[12] 言晨琦,王瑞白,刘海灿,等. ARIMA 模型预测 2018-2019 年我国肺结核发病趋势的应用. 中华流行病学杂志,2019,40(6):633-637.

[13] Zheng YL, Zhang LP, Zhang XL, et al. Forecast model analysis for the morbidity of tuberculosis in Xinjiang, China. PloS one, 2015, 10(3): e0116832.

[14] Maletic JI, Marcus A, Grzymala-Busse JW, et al. Data Mining and Knowledge Discovery Handbook. Kybernetes, 2005, 33(7): 809-835.

[15] 徐学琴,裴兰英,王瑾瑾,等. 基于支持向量机的麻疹发病率预测研究. 中华疾病控制杂志,2017,21(5):528-530.

[16] 徐学琴,孙宁,徐玉芳. 基于 BP 神经网络的河南省甲乙类法定报告传染病预测研究. 中华疾病控制杂志,2014,18(6):561-563.

[17] Mollalo A, Mao L, Rashidi P, et al. A GIS-Based Artificial Neural Network Model for Spatial Distribution of Tuberculosis across the Continental United States. Int J Environ Res Public Health, 2019, 16(1): 157.

[18] Zou JJ, Jiang GF, Xie XX, et al. Application of a combined model with seasonal autoregressive integrated moving average and support vector regression in forecasting hand-foot-mouth disease incidence in Wuhan, China. Medicine(Baltimore), 2019, 98(6): e14195.

[19] Li Z, Wang Z, Song H, et al. Application of a hybrid model in predicting the incidence of tuberculosis in a Chinese population. Infect Drug Resist, 2019, 12: 1011-1020.

[20] 谢骁旭,袁兆康. 基于 R 的江西省肺结核发病率 ARIMA-SVM 组合预测模型. 中国卫生统计,2015,32(1):160-162.

(责任编辑:张 悦)