

基于多源数据融合的流行病组合预测方法

李怡琳¹, 苏学峰², 李慧³, 朱梦旒⁴

1. 山西医科大学公共卫生学院, 山西太原 030000; 2. 山西省临汾市人民医院院办公室, 山西临汾 041000; 3. 山西省临汾市人民医院智慧医院办公室, 山西临汾 040000; 4. 山西医科大学管理学院, 山西太原 030000

【摘要】为了改善传统流行病预测模型普遍存在的准确度低、泛化性弱、结构单一、非线性处理能力差、预测时间长等问题, 研究基于多源数据融合的流行病组合预测方法。将采集到的流行病多源数据进行归一化和主成分分析技术特征选择, 结合ARIMA模型、灰色GM模型和BPNN神经网络, 构建一个ARIMA-GM-BPNN流行病组合预测模型, 将前两种预测模型的拟合值作为BPNN神经网络的输入, 进行模型训练, 将数据充分融合后, 综合不同预测模型的优点, 获取最优流行病组合预测模型, 对未来流行病的发病率和发病趋势进行预测。实验表明该方法组合模型拟合效果很好, 流行病发病率预测值与真实值接近, 预测趋势也十分吻合, 提高预测精度和泛化能力, 可以为流行病的预测和防治工作提供可靠的数据支持。

【关键词】流行病; 预测模型; 多源数据融合; 组合预测; 发病趋势

【中图分类号】R318; TP391

【文献标志码】A

【文章编号】1005-202X(2024)02-0258-07

Epidemic prediction method based on multi-source data fusion

LI Yilin¹, SU Xuefeng², LI Hui³, ZHU Mengni⁴

1. School of Public Health, Shanxi Medical University, Taiyuan 030000, China; 2. Hospital Office, Linfen People's Hospital, Linfen 041000, China; 3. Office of Smart Hospital, Linfen People's Hospital, Linfen 040000, China; 4. School of Management, Shanxi Medical University, Taiyuan 030000, China

Abstract: A combined epidemic prediction method based on multi-source data fusion is presented to address the common problems of low accuracy, weak generalization, single structure, poor nonlinear processing ability, and long prediction time in traditional epidemic prediction models. The collected multi-source epidemic data are normalized and subjected to feature selection using principal component analysis. An ARIMA-GM-BPNN model for pandemic prediction is constructed by combining ARIMA model, grey GM model and BPNN. The fitting values of the first two prediction models are used as inputs to BPNN for model training. After sufficiently integrating the data and combining the advantages of different prediction models, the optimal combined model is obtained and used for forecasting the incidence and trend of epidemics. Experimental results show that the combined model exhibits excellent fitting performance, with predicted incidences and trends consistent with the real conditions. The proposed approach improves prediction accuracy and generalization capabilities, and it can provide reliable data support for epidemic prediction and control.

Keywords: epidemic; prediction model; multi-source data fusion; combined prediction; incidence trend

前言

近年来多种高发的流行病在世界范围内爆发, 对人们的身体健康产生极大的威胁。各种流行病的致病因素、流行趋势以及预防控制措施成为最近医

学界研究的热点问题^[1]。为了尽早发现和遏制各种流行病的发生, 增强抵抗突发疫情的能力, 我国卫生部门建立疫情管理系统, 通过大数据动态分析, 制定相应的流行病预防评价机制, 从而确定具体的防治政策, 因此对于流行病发病趋势进行准确的预测是实现防治的前提^[2]。流行病的传播受遗传因素和环境因素的影响, 因此建立疾病预测模型需要根据流行病学因素进行研究, 从而准确圈定高危人群, 对其采取相应的预防措施^[3]。何宇浩等^[4]采用时空图卷积网络构建流行病预测模型, 利用各城市之间的传

【收稿日期】2023-10-28

【基金项目】山西省重点研发计划(201903D311010)

【作者简介】李怡琳, 在读研究生, 研究方向: 公共卫生, E-mail:

L1104598000@163.com

播相似度进行预测,同时加权处理拓扑图,根据城市网络的空间特征,预测城市的流行病发病趋势。蔡晶等^[5]使用ARIMA模型预测猩红热的发病趋势,采用季节指数、聚类分析和Spearman相关分析等模型,最终拟合的最优模型为(ARIMA)(0, 1, 1)(0, 1, 0)。但是上述两种方法获得的预测结果仍然存在准确度低、泛化能力差的缺点,而且还存在预测模型单一、非线性处理能力差、预测时间长等问题。随着互联网的发展,数据规模也越来越大,我国早已进入大数据时代,多源数据融合技术也在各个领域被广泛使用,将多种数据进行统一和融合,能够提高信息采集的质量和精度,为流行病预测模型的建立提供新的技术支持^[6]。同时为了突破传统预测模型的局限,组合预测模型也在流行病预测领域广泛应用^[7]。

本文提出一种基于多源数据融合的流行病组合预测方法,在数据融合的基础上,构建一个ARIMA-GM-BPNN流行病组合预测模型,可以提高预测结果的准确度和泛化能力。

1 多源数据融合的流行病组合预测

1.1 流行病多源数据采集

随着互联网的发展,网络数据呈爆发式增长,现代信息技术与多种学科交叉,诞生了新的多源数据融合技术,它可以增强信息采集的质量和精度,通过融合多个信息源的数据,将其进行分析、关联、估计、组合等处理,从而获取身份和状态的精确估算,用于更进一步的综合决策^[8]。由于单一的信息来源普遍存在数据不准确、不全面的缺点,因此使用多源数据融合技术分析和处理不同领域或数据源采集的数据资料,可以增加数据采集的精度,通过不同源数据的互补,得到相对准确和完善的判定和估算结果。

影响流行病发病和传播的因素有很多,主要包括患者的年龄、性别、基因分型、家族遗传史、发病时间、诊断时间、临床分型、先证者临床信息等遗传因素和环境因素两大类^[9]。其中,环境因素分为自然环境因素和社会环境因素,自然环境因素包括地理、气候、生物种类和分布等,社会环境因素包括城市分布、人口结构、工作状况、医院分布、经济文化水平等。

关于流行病的数据采集来源^[10],除了世界卫生组织和国家卫生部等公共权威部门公布的统计数据以外,还包括中国疾病-预防控制信息系统、传染病-报告信息管理系统、突发公共卫生事件-管理信息系统等,以及各大医院的病历系统和网上的各种公开医学病例数据库等。同时还可以对各大网站平台进行接口采集,直接抓取各地新闻媒体的数据。结合流行病学的数据库情况和数据需求,本文采用多种

数据采集技术结合的数据采集方案,包括数据采集^[11]、数据存储^[12]和数据管理,为后续的预测分析打好基础,具体技术架构如图1所示。

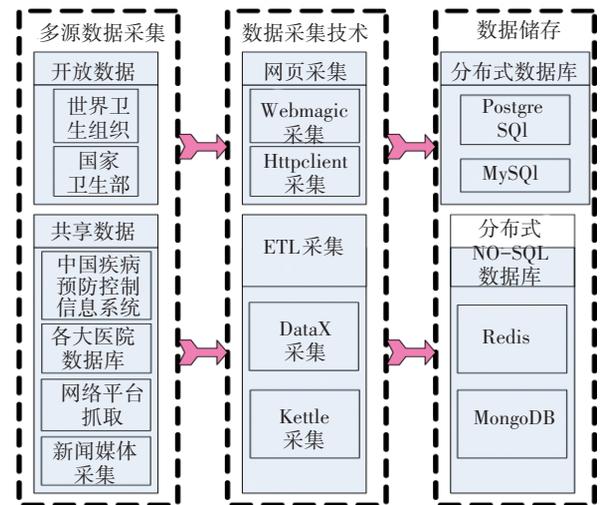


图1 流行病多源数据采集架构

Figure 1 Multi-source data acquisition architecture for epidemics

1.2 流行病多源数据预处理

1.2.1 归一化 在将采集到的流行病多源数据进行集中分析和融合之前,需要统一各个不同类型数据的取值范围,将其进行归一化处理。这样能够使各个维度上的特征在数值上进行对比,从而降低计算复杂度,数据分析的精度也可以提升。本文采用标准化的方式对多源数据进行归一化处理,通过标准分数(Z-score)代表数据,其运算公式如下:

$$z = \frac{(z - \delta)}{\omega} \quad (1)$$

式中,全部样本数据的均值和标准差分别用 δ 和 ω 描述。在均值为0的Z-score标准化当中,原数据的方差在归一化后形成新的数据,各个维度都具有相同价值的量纲,因此全部都属于0均值、1方差的正态分布,有利于后续的分析 and 处理。

1.2.2 主成分分析法(PCA)特征选择 完成归一化处理的流行病多源数据,各种数据的特征维度还是存在很大的差别,为了避免在使用组合预测模型进行流行病预测时出现过拟合现象,导致预测效果不理想,需要对这些数据再次降维^[13],通过特征选择算法,将各个数据源的最佳特征数据筛选出来^[14]。本文在多源数据的降维处理方面,采用PCA方法^[15],它是统计最优原则的一种变换方法,可以在数据降维的过程中,完全保留原始数据的信息,获取新的主元向量,并按照其重要程度进行排序,保留重要的精化部分,删除后面不重要的维数,从而实现降维的目的。

的^[16]。该方法不仅可以将数据预测过程进行简化,还具有用户独立性,最终降维后的结果不与数据以外的任何因素相关。

使用少数的综合变量替代原始流行病多源数据中大量的单个变量即为PCA的主要过程,可以将原始流行病多源数据变量的杂乱信息进行精简,只显示重要的信息,其中各个综合变量之间是相互独立的,没有任何相关性。假设某个数据源采集的原始数据流行病影响因素为 P_1, P_2, \dots, P_q , 设置样本点 n 个,变量 q 个,那么PCA数学模型如下:

$$P = (P_1, P_2, \dots, P_q) = \begin{pmatrix} p_{11} & \dots & p_{1q} \\ \dots & c_{ij} & \dots \\ p_{n1} & \dots & p_{nq} \end{pmatrix} \quad (2)$$

式中,主成分系数用 c_{ij} 表示^[17];主成分顺序用 i 表示, $i = 1, \dots, n$;主成分变量顺序用 j 表示, $j = 1, \dots, q$ 。

上述PCA完成后,即可获得综合变量 k 个,其公式为:

$$\begin{cases} C_1 = c_{11}p_1 + c_{21}p_2 + \dots + c_{q1}p_q \\ C_k = c_{1k}p_1 + c_{2k}p_2 + \dots + c_{qk}p_q \end{cases} \quad (3)$$

将以上公式精简后可得:

$$C = Pc$$

$$C_i = c_{1i}p_1 + c_{2i}p_2 + \dots + c_{qi}p_q, i = 1, \dots, k \quad (4)$$

经过上述计算可得 C_1 为第一主成分, C_2 为第二主成分,以此类推,直到 C_k 为第 k 主成分。分别将各个数据源提取的主成分进行汇总,作为后期流行病预测模型的分析数据。

1.3 基于数据融合的ARIMA-GM-BPNN流行病组合预测模型

本文在多源数据采集和预处理的基础上,结合ARIMA模型^[18]、灰色GM模型^[19]和BPNN神经网络^[20],构建一个基于多源数据融合的ARIMA-GM-BPNN流行病组合预测模型,将前两种预测模型的拟合值作为BPNN神经网络的输入,进行网络训练,将数据充分融合后,综合不同预测模型的优点,获取最佳网络模型,提高预测结果的准确度,组合预测模型构建原理如图2所示。

1.3.1 ARIMA模型 在时间序列分析预测方法当中,使用频率较高的是ARIMA模型,因为它擅长分析各个影响因素发展的周期性和延续性,因此在各种疾病的预测研究领域被广泛应用^[21],其公式如下:

$$\text{ARIMA}(g, h, l) = C \times (G, H, L)_B \quad (5)$$

式中,自回归、差分和移动平均阶数分别用 g, h, l 描述,这三者的季节阶数分别用 G, H, L 描述;周期步长用 B 描述。

ARIMA模型的构建包括以下4个过程:(1)稳定

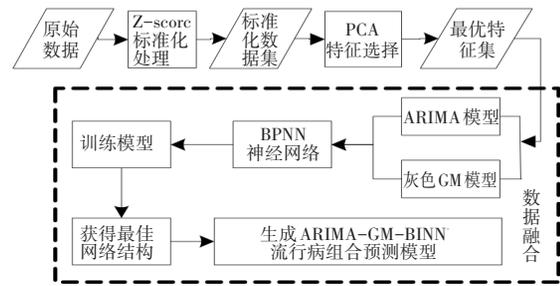


图2 流行病组合预测模型构建原理

Figure 2 Construction principle of combined model for epidemic prediction

数据序列:对流行病的发病率时间序列进行单位根检测,阈值为0.03,如果检测结果小于阈值,则序列稳定,反之序列就不稳定,需要通过取对数变换或差分法对其进行平稳化处理;(2)鉴别模型:将模型需要的阶数进行鉴别,可以参考稳定的序列自相关和偏自相关图,根据二者的分布,初步进行确定;(3)测试和估算模型参数:使用不同的阶数组合构建ARIMA模型^[22],在统计学范围内进行测试,假设残差的统计量阈值为0.04,那么残差大于阈值的就是白噪声,说明测试合格,最佳模型的确定需要参考AIC值最小原则和拟合优度值 R^2 ;(4)模型预测效果评估:通过对比过去几年我国流行病预测目标值与实际目标值之间的差距,验证模型预测的效果,采用平均相对误差(MAPE)和均方误差(MSE)作为评估指标。

1.3.2 灰色预测模型GM 使用少量的样本和部分信息,即可通过构建灰色微分数学模型,预测目标未来的发展规律,该方法也被称为灰色预测模型^[23]。假设某时间段内,某项流行病的发病次数采集的原始时间数据序列为:

$$s^{(0)} = \{s^{(0)}(1), s^{(0)}(2), \dots, s^{(0)}(t)\} \quad (6)$$

式中,原始时间序列观察值总数用 V 描述, $t \in V$ 。

在计算时,采用累加生成运算^[24],通过一次累加可得:

$$s^{(1)} = \{s^{(1)}(1), s^{(1)}(2), \dots, s^{(1)}(t)\}, t \in V \quad (7)$$

则 $s^{(1)}$ 需要遵守的约束为一阶常微分方程:

$$r = \frac{ws^{(1)}}{wt} + fs^{(1)} \quad (8)$$

式中,发展灰数用 f 描述;内生控制灰数用 r 描述;该方程需要遵循的初始约束如下:

$$\text{当 } t = t_0 \text{ 时, } s^{(1)} = s^{(1)}(t_0) \quad (9)$$

那么该方程的解可以表示为:

$$s^{(1)}(t) = \frac{r}{f} + \left[s^{(1)}(t_0) - \frac{r}{f} \right] e^{-f(t-t_0)} \quad (10)$$

将以上结果进行最小二乘估计可得:

$$\hat{U} = \begin{bmatrix} \hat{f} \\ \hat{r} \end{bmatrix} = M^T u (M^T M)^{-1} \quad (11)$$

在时间响应方程当中输入上述两种估计值 \hat{f} 和 \hat{r} ,结果如下:

$$\hat{s}^{(1)}(t+1) = \left[s^{(1)}(1) - \frac{r'}{\hat{f}} \right] e^{-\hat{f}t} + \frac{r'}{\hat{f}} \quad (12)$$

计算得出的响应函数为:

$$S(t+1) = 0.589426e^{0.001572t} - 0.615743 \quad (13)$$

设定拟合值为 $\hat{s}^{(1)}(i)$,采用时间响应方程进行求解,通过累减生成计算进行还原,可得:

$$\hat{s}^{(0)}(i) = s(i) - \hat{s}^{(1)}(i-1), i = 2, 3, \dots, t, t \in V \quad (14)$$

1.3.3 BPNN BPNN是一种多层前馈神经网络,包括输入层、隐含层和输出层,它是按照误差逆向传播算法进行训练的神经网络,能够映射多维函数,具有优秀的模式分类能力^[25]。将上述两种预测模型ARIMA和GM输出的拟合值输入BPNN预测模型当中进行网络训练,通过3层BPNN^[26]的学习,获取相应的组合权重,在学习过程中需要遵守MSE最小原则,详细的BPNN模型如图3所示。

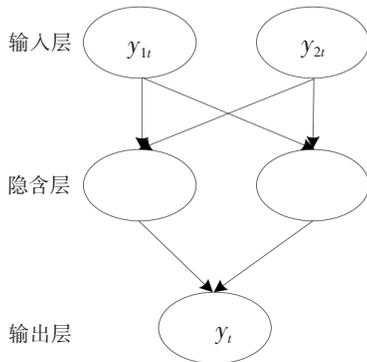


图3 3层BPNN模型

Figure 3 Three-layer BPNN model

BPNN模型使用的传递函数为S型,表示为:

$$f(x) = \frac{1}{1 + e^{-x}} \quad (15)$$

采用反向误差函数,表达如下:

$$I = \frac{\sum_i (a_i + o_i)^2}{2} \quad (16)$$

式中,期望值用 a_i 描述;运算输出值用 o_i 描述。

为了实现 e 值最小的目标,需要将权值和阈值进行循环调节,具体步骤如下:(1)模型训练:将两种模型ARIMA和GM输出的预测数据进行整合和归一化处理,输入BPNN当中进行学习,将实际的数据归一化后作为网络的输出,从而获得训练模型;(2)进

行目标预测:在BPNN当中,输入步骤(1)获取的结果数据,对目标进行预测,分为3个层面的工作:输入层:通过 y_{1t} 层,优化ARIMA模型的预测结果;通过 y_{2t} 层,优化灰色GM模型的预测结果。隐含层:在网络构成的中间部分确定神经元的数量,数量太多会导致运算量增加,频繁引发拟合问题,数量太少会导致运算效果不理想。本文根据多次实验获得神经元最佳数量计算公式为:

$$v = \sqrt{E + D} + A \quad (17)$$

其中,神经元输入和输出数量分别用 E 和 D 描述,常数0~10用 A 描述。输出层:通过 y_t 层输出基于BPNN进行数据融合后的组合预测结果;(3)使用滚动目标预测方法,定期更新前一时间段的历史数据,重新对目标进行预测,从而保持数据的实时性,不断提升对未来趋势预测的准确率。

1.3.4 ARIMA-GM-BPNN 组合预测模型 为了更好地提升流行病组合预测模型的准确率,本文将时间序列分析预测模型、灰色预测模型和BPNN 3种方法进行组合,得到ARIMA-GM-BPNN组合预测模型。结合各个预测模型的优点,将多个单一模型的预测数据进行融合,用时间序列分析预测模型ARIMA拟合流行病预测目标时间序列的线性部分,通过灰色预测模型拟合流行病预测目标时间序列的非线性部分,利用BPNN对两者的预测结果进行学习和优化,最终提高组合预测模型的精度和泛化能力。组合模型的预测结果表示为:

$$\hat{F}_t = \hat{Q}_t + \hat{K}_t \quad (18)$$

式中,ARIMA模型的预测结果用 \hat{Q}_t 表示;GM预测模型的预测结果用 \hat{K}_t 表示。

2 结果与分析

2.1 实验对象与数据采集

本文选择流行病中的麻疹作为预测对象,根据某省2009~2022年的麻疹发病案例和数据资料,预测未来1年麻疹的发病率情况,从而为下一步的麻疹防治提供可靠的参考依据。通过中国疾病预防控制中心信息系统下属传染病报告信息管理系统进行麻疹病例的个案数据统计,数据采集的标准如下:(1)采集范围为某省境内的全部居民;(2)发病日期为2009年1月1日至2022年12月31日;(3)全部经过审核的确诊病例和临床诊断。

麻疹爆发的疫情数据由中国疾病-预防控制信息系统下属突发公共卫生事件-管理信息系统提供,数据采集的标准如下:(1)事件发生时间为2009年1月1日至2022年12月31日;(2)严重等级属于一般及以上级别

事件;(3)属于已经审核完毕的事件,并已经结案。

在麻疹发病率的预测当中,人口数量也是很重要的一个影响因素,为了准确计算发病率,人口统计数据采用国家统计局发布的官方数据。本文采用 R3.6.0 软件进行数据整理和构建预测模型,在构建时间序列模型时,使用 Forecast/tseries 包工具,通过 AMORE 包构建 BPNN 模型。

2.2 流行病组合模型预测实验

为了验证本文方法在流行病预测方面的效果,根据某省 2009~2022 年的麻疹发病案例和数据资料,进行 2023 年度某省麻疹发病率的预测实验,通过多种途径进行多源数据采集,得到 2009~2022 年某省麻疹发病率和发病人数的统计数据,如图 4 所示。

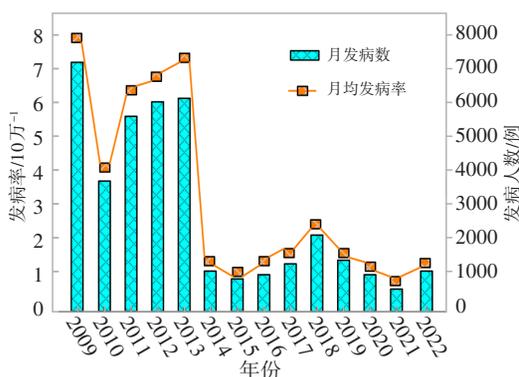


图 4 2009~2022 年某省麻疹发病率和发病人数

Figure 4 Incidence and number of measles cases in a province from 2009 to 2022

从图 4 可以看出,某省在 2009~2022 年的麻疹报告总病例为 25 387 例,月均发病率为 4.02/10 万。其中发病率比较突出的年份为 2009~2013 年,在这之后麻疹的发病率降低很多,直到 2015 年降到 1/10 万以下,之后虽然有反弹的趋势,但是增加并不是很多,在 2021 年甚至达到 0.5/10 万的最低水平,但是 2022 年又有所回升。

实验将麻疹发病率历史数据中的前 2009~2021 年数据作为预测模型构建的训练集数据,将 2022 年的数据作为模型验证的测试集数据,将采集到的多源数据进行预处理,通过归一化和 PCA 特征选择后,构建 ARIMA-GM-BPNN 麻疹发病率组合预测模型,具体步骤如下:(1)构建 ARIMA(1, 0, 1)模型,根据原始数据,输出 ARIMA 预测值 X_1 ; (2)构建灰色 GM(1, 1)模型,根据原始数据,输出灰色预测值 X_2 ; (3)构建 ARIMA-GM-BPNN 组合模型,分别建立 6 个不同结构的 BPNN,用 X_1 和 X_2 值分别对这些网络进行训练,根据不同结构神经网络的 MSE 结果(图 5),选择误差最小的 2-4-1 网络结构为最佳结构,其学习效

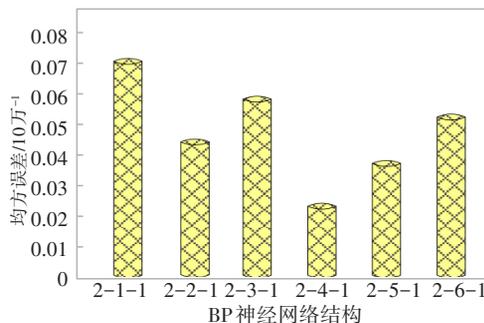


图 5 不同结构神经网络的 MSE

Figure 5 Mean square error of neural networks with different structures

率为 0.02,最高训练次数为 1 500 次。

将 X_1 和 X_2 输入最佳 BPNN 当中,输出值是真实数据 X ,使三者共同形成一个数据传输组合 $((X_1, X_2), X)$ 。其中有 15 个神经元处在神经网络的隐含层当中,输入层的两个神经元就是 ARIMA 模型和 GM 模型预测数据,输出节点就是真实数据。使用组合预测模型进行麻疹发病率的预测,最终得到组合预测模型的拟合值。将拟合值与真实值进行对比分析,ARIMA-GM-BPNN 组合预测模型拟合效果如图 6 所示。从图 6 可以看出,组合预测模型的拟合值与当时的真实值数据之间基本上都是趋于吻合的,说明本文方法构建的 ARIMA-GM-BPNN 组合预测模型拟合效果很好,预测的数据与真实值相差很小,准确度很高。

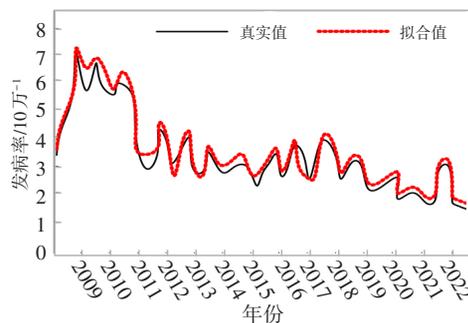


图 6 组合预测模型拟合图

Figure 6 Fitting diagram of combined prediction model

采用最佳 ARIMA-GM-BPNN 组合预测模型,预测某省 2023 年的麻疹发病率,实验结果如图 7 所示。从图 7 可以看出,2023 年某省麻疹发病率预测最高的月份为 5 月,4 月和 12 月次之,根据每个月的发病率进行运算,可以获得 2023 年度的麻疹预测月均发病率为 2.17/10 万,说明该省今年麻疹病流行的因素较多,预防麻疹病的防控措施需要进一步加强。

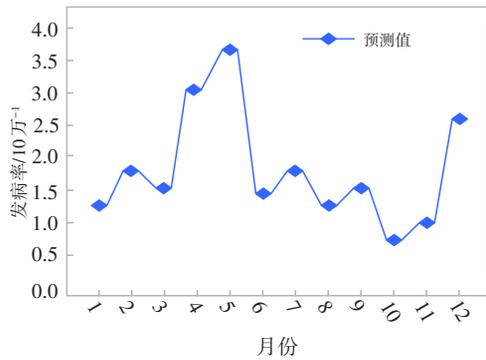


图7 某省2023年麻疹发病率预测趋势图

Figure 7 Forecast trend of measles incidence in 2023 in a province

2.3 不同模型预测效果对比

为了进一步测试本文方法的预测精度和泛化能力,分别使用传统的ARIMA预测模型、GM预测模型、SVR回归预测模型,以及本文方法预测2022年度某省的麻疹发病率,并将所得预测结果与真实值进行比较,统计误差数据,实验结果如表1所示。从表1可以看出,相比于其他传统的预测模型,本文方法预测的结果准确度最高,准确率达98.27%,同时MSE和MAPE都是最小的,分别为0.0311/10万和3.67%,说明本文方法的预测精度最高,效果最好,结合不同模型的优点,在线性和非线性数据的处理上,能力均有所增强,因此无论是在学习能力、泛化能力以及预测精度方面都实现大幅度提升。

表1 不同模型预测精度和误差比较

Table 1 Comparison of prediction accuracy and error among different models

模型方法	准确率/%	MAPE/%	MSE/10万 ⁻¹
ARIMA 预测模型	88.32	8.78	0.0639
GM 预测模型	95.21	5.47	0.0478
SVR 回归预测模型	92.89	6.89	0.0569
本文方法	98.27	3.67	0.0311

3 结论

本文将流行病的多源数据进行采集和预处理后,构建一个基于数据融合的ARIMA-GM-BPNN流行病组合预测模型,从而提高预测结果的准确度。实验表明本文方法构建的ARIMA-GM-BPNN流行病组合预测模型拟合效果很好,预测值与真实值非常接近,提高预测精度和泛化能力,可以为流行病的预测和防治工作提供可靠的数据支持。

【参考文献】

- [1] Ahmed I, Goufo E, Yusuf A, et al. An epidemic prediction from analysis of a combined HIV-COVID-19 co-infection model via ABC-fractional operator[J]. Alex Eng J, 2021, 60(3): 2979-2995.
- [2] 付之鸥, 鲍昌俊, 李中杰, 等. 基于“大数据”的流感预警研究进展[J]. 中华流行病学杂志, 2020, 41(6): 975-980.
Fu ZO, Bao CJ, Li ZJ, et al. Progress of research regarding the influenza early warning system, based on "Big Data"[J]. Chinese Journal of Epidemiology, 2020, 41(6): 975-980.
- [3] 邹娇娇, 王雷, 陈琦, 等. 2012-2018年湖北省流行性腮腺炎流行病学特征和短期发病预测[J]. 中国疫苗和免疫, 2020, 26(1): 45-48.
Zou JJ, Wang L, Chen Q, et al. Epidemiology of mumps during 2012-2018 in Hubei and prediction of short-term mumps incidence[J]. Chinese Journal of Vaccines and Immunization, 2020, 26(1): 45-48.
- [4] 何宇浩, 郑贤伟. 基于CIST-GCN的流行病数据分析与预测[J]. 现代信息技术, 2022, 6(14): 30-34.
He YH, Zheng XW. Analysis and prediction of epidemic data based on CIST-GCN[J]. Modern Information Technology, 2022, 6(14): 30-34.
- [5] 蔡晶, 黄淑琼, 杨雯雯, 等. 2010-2018年湖北省猩红热流行病学特征及发病趋势预测[J]. 中华疾病控制杂志, 2020, 24(2): 134-138.
Cai J, Huang SQ, Yang WW, et al. Epidemiological characteristics and trend prediction of scarlet fever in Hubei Province from 2010 to 2018[J]. Chinese Journal of Disease Control & Prevention, 2020, 24(2): 134-138.
- [6] 雷秀娟, 张文祥, 刘恋, 等. 基于多数据融合的circRNA-疾病关联关系预测[J]. 中国科学(信息科学), 2021, 51(6): 927-939.
Lei XJ, Zhang WX, Liu L, et al. Prediction of circRNA-disease associations based on multiple bi-ological data[J]. Science in China (Information Sciences), 2021, 51(6): 927-939.
- [7] 庞维庆, 何宁, 罗燕华, 等. 基于数据融合的ABC-SVM社区疾病预测方法[J]. 浙江大学学报(工学版), 2021, 55(7): 1253-1260.
Pang WQ, He N, Luo YH, et al. ABC-SVM disease prediction method based on data fusion in community health care[J]. Journal of Zhejiang University (Engineering Science), 2021, 55(7): 1253-1260.
- [8] 吕仲琪, 董卓达, 刘晓丽, 等. 多源异构数据特征的智能结构化方法仿真[J]. 计算机仿真, 2022, 39(7): 451-455.
Lü ZQ, Dong ZD, Liu XL, et al. Intelligent multi-source heterogeneous structured data characteristics simulation[J]. Computer Simulation, 2022, 39(7): 451-455.
- [9] 李若瞳, 饶慕蓁, 傅晏红, 等. 2030年中国慢性阻塞性肺疾病的疾病负担预测与危险因素控制效果模拟[J]. 中华流行病学杂志, 2022, 43(2): 201-206.
Li RT, Rao ZZ, Fu YH, et al. Prediction on the burden of disease of chronic obstructive pulmonary disease and simulation of the effectiveness of controlling risk factors in China by 2030[J]. Chinese Journal of Epidemiology, 2022, 43(2): 201-206.
- [10] 李展鹏, 邹孝付, 苏雍贺, 等. 基于FPGA的多源异构数据并行可配置采集方法[J]. 计算机集成制造系统, 2021, 27(4): 1008-1020.
Li ZP, Zou XF, Su YH, et al. Configurable acquisition method of multi-source heterogeneous data based on FPGA[J]. Computer Integrated Manufacturing Systems, 2021, 27(4): 1008-1020.
- [11] Pertuz T, Malechmir A, Bos J, et al. Broadband seismic source data acquisition and processing to delineate iron oxide deposits in the Bltberget mine-central Sweden[J]. Geophys Prospect, 2022, 70(1): 79-94.
- [12] Schmitter S, Grundmann S, John K, et al. Reynolds stress tensor and velocity measurements in technical flows by means of magnetic resonance velocimetry Data acquisition in the flow field behind the sudden expansion of the FDA benchmark nozzle and identification of systematic errors[J]. Tech Mess, 2022, 89(3): 201-209.
- [13] Fujiwara T, Shilpika F, Sakamoto N, et al. A visual analytics framework for reviewing multivariate time-series data with dimensionality reduction[J]. IEEE Trans Vis Comput Graph, 2021, 27(2): 1601-1611.
- [14] 刘瑞, 闫敏敏, 魏嵘, 等. 儿童法定传染病流行特征及发病趋势预测分析[J]. 中国医院管理, 2023, 43(2): 46-48.
Liu R, Yan MM, Wei R, et al. Epidemic features and incidence trend of notifiable infectious diseases for children[J]. Chinese Hospital Management, 2023, 43(2): 46-48.

- [15] Sebro R, Garza-Ramos C. Support vector machines are superior to principal components analysis for selecting the optimal bones' CT attenuations for opportunistic screening for osteoporosis using CT scans of the foot or ankle[J]. *Osteoporosis Sarcopenia*, 2022, 8(3): 112-122.
- [16] 徐先峰, 赵依, 龚美, 等. 基于特征降维和组合模型的短期电力负荷预测[J]. *计算机仿真*, 2022, 39(4): 66-70.
Xu XF, Zhao Y, Gong M, et al. Short-term power load forecasting based on dimensionality reduction and combined model[J]. *Computer Simulation*, 2022, 39(4): 66-70.
- [17] Ramirez-Estrada A, Mena-Cervantes VY, Mederos-Nieto FS, et al. Assessment and classification of lignocellulosic biomass recalcitrance by principal components analysis based on thermogravimetry and infrared spectroscopy[J]. *Int J Environ Sci Te*, 2022, 19(4): 2529-2544.
- [18] Abotaleb M, Makarovskikh T, Yonar A, et al. State of the art in wind speed in England using BATS, TBATS, Holt's Linear and ARIMA model[J]. *Mausam*, 2022, 73(1): 129-138.
- [19] Yang X, Ding D, Luo Y. Non-equidistance fractional order accumulation grey model NFGM (1, 1) and its application[J]. *IAENG International Journal of Computer Science*, 2021, 38(3 Pt 2): 672-677.
- [20] Zhang C, Tian YX, Fan ZP. Forecasting sales using online review and search engine data: a method based on PCA-DSFOA-BPNN[J]. *Int J Forecasting*, 2022, 38(3): 1005-1024.
- [21] 白哲, 罗云云, 周智博, 等. 基于机器学习算法的大于胎龄儿风险预测模型[J]. *中华流行病学杂志*, 2021, 42(12): 2143-2148.
- Bai X, Luo YY, Zhou ZB, et al. Development and evaluation of a machine learning prediction model for large for gestational age[J]. *Chinese Journal of Epidemiology*, 2021, 42(12): 2143-2148.
- [22] Li Q, Zhang L, Wang X. Loosely coupled GNSS/INS Integration based on factor graph and aided by ARIMA model[J]. *IEEE Sens J*, 2021, 21(21): 24379-24387.
- [23] 邱劲松, 罗磊. 基于灰色模型的中国职业性耳鼻喉口腔疾病发病预测研究[J]. *环境与职业医学*, 2020, 37(3): 281-284.
Qiu JS, Luo L. Prediction of occupational otorhinolaryngological diseases in China based on grey model[J]. *Journal of Environmental & Occupational Medicine*, 2020, 37(3): 281-284.
- [24] Khan MA, Haque A, Kurukuru V, et al. Stand-alone operation of distributed generation systems with improved harmonic elimination scheme[J]. *IEEE J Em Sel Top P*, 2021, 9(6): 6924-6934.
- [25] 窦嘉铭, 马鸿雁. 基于多神经网络组合的建筑能耗预测研究[J]. *计算机仿真*, 2022, 39(5): 438-443.
Dou JM, Ma HY. Prediction of building energy consumption based on ensemble artificial neural networks[J]. *Computer Simulation*, 2022, 39(5): 438-443.
- [26] 刘金培, 张了丹, 朱家明, 等. 非结构性数据驱动的混合分解集成碳交易价格组合预测[J]. *运筹与管理*, 2023, 32(3): 149-154.
Liu JP, Zhang LD, Zhu JM, et al. Unstructured data driven carbon price combined forecast based on hybrid decomposition-integration[J]. *Operations Research and Management Science*, 2023, 32(3): 149-154.

(编辑:陈丽霞)