樊冲. 基于动态集成加权概率 RF 的门诊量预测[J]. 智能计算机与应用,2024,14(5):209-214. DOI:10.20169/j.issn.2095-2163.240529

基于动态集成加权概率 RF 的门诊量预测

樊 冲
(锦州市大数据管理中心,辽宁 锦州 121000)

摘 要:医院门诊量本质上是一种具有潜在规律的时间序列,通过对门诊量进行有效分析和预测,可以更加科学、合理地配置医疗资源。针对门诊量波动幅度较大的时间序列预测问题,提出一种基于动态集成加权概率 RF 的门诊量预测方法。首先选择具有强泛化性的随机森林(Random Forest,RF)作为预测模型;并且采用 k 近邻-层次聚类算法对 RF 模型中树的强度进行评估,从中动态选择性能最佳的决策树,提高回归模型的性能;为了提升预测模型的准确率,采用加权概率融合规则代替原始 RF 模型的求平均数的规则。经过与 BP 神经网络和 RF 对比实验结果表明,提出方法可以更加精准地对门诊量进行预测和分析,为医院更好的运营管理提供了重要依据和决策支持。

关键词:门诊量;随机森林;k 近邻-层次聚类;加权概率融合
中图分类号:TP391 **文献标志码**:A **文章编号**:2095-2163(2024)05-0209-06

Outpatient volume prediction based on dynamic integrated weighted probability RF

FAN Chong
(Jinzhou Big Data Management Center, Jinzhou 121000, Liaoning, China)

Abstract: Hospital outpatient volume is essentially a time series with potential laws. Through effective analysis and prediction of outpatient volume, medical resources can be allocated more scientifically and reasonably. Aiming at the time series forecasting problem of large fluctuation of outpatient volume, a forecasting method of outpatient volume based on dynamic integrated weighted probability RF is proposed. Firstly, Random Forest (RF) with strong generalization is selected as the prediction model. Furthermore, K nearest neighbor-hierarchical clustering algorithm is used to evaluate the strength of trees in RF model, and the decision tree with the best performance is dynamically selected to improve the performance of regression model. Then, in order to improve the accuracy of the prediction model, the weighted probability fusion rule is used instead of the average rule of the original RF model. Compared with BP neural network and RF, the proposed method can predict and analyze outpatient volume more accurately, which provides the important basis and decision support for better hospital operation and management.

Key words: outpatient volume; Random Forest; K-nearest neighbor hierarchical clustering; weighted probability fusion

0 引 言

医院门诊量预测在医疗管理中扮演着至关重要的角色,其准确性直接影响着医院资源的合理配置和医疗服务的质量。过去几十年间,随着医疗信息化技术的发展和数据挖掘方法的成熟,预测模型的研究和应用日益深入。然而,门诊量的预测受到多种因素的影响,包括季节性变化、假日效应、突发事件等,使得其预测任务变得复杂且具有挑战性。

在传统的预测方法中,一些基于时间序列分析的模型被广泛应用。例如,桑泉红等学者[1]、焦晨等学者[2]提出了一种基于时序序列模型的医院门诊人次预测方法。该方法利用历史门诊量数据,通过时间序列分析和建模技术,实现对未来门诊量的预测。另外,刘焰等学者[3]、陈辉等学者[4]利用移动平均季节指数法进行门诊量分析与预测,该方法结合了季节性和移动平均的特征,适用于具有周期性变化的门诊量数据。在传统的预测方法中,灰色预测模型也是一种常用的方法之一[5]。灰色预测模型源于灰色系统理论,其特点是能够处理数据样本较少、不完整、不确定和不精确的情况,适用于非线性和非平稳的时间序列数据[6-7]。例如,王琦等学者[8]应用灰色 GM(1,1)预测模型在门诊量预测中取得了一定的成效。该模型通过建立灰色微分方程,对门诊量数据进行拟合和

作者简介:樊 冲(1981-),男,硕士,高级工程师,主要研究方向:数据分析,虚拟化。Email:609729352@qq.com

收稿日期:2023-12-29

预测,能够有效地捕捉到数据的变化趋势和规律性。

除了传统的时间序列模型外[9],近年来深度学习方法在医院门诊量预测中也取得了显著进展。吴磊等学者[5]提出了一种基于深度神经网络的门诊量预测方法,通过构建多层神经网络模型,实现对门诊量数据的端到端学习和预测。此外,张筠莉等学者[7]将现代神经网络技术与传统灰色理论相结合,提出了一种灰色 RBF 神经网络模型用于门诊量的预测,该方法能够充分挖掘数据的非线性特征,提高了预测的准确性和稳定性。

尽管深度学习方法在门诊量预测中取得了一定成效,但也存在一些局限性,例如需要大量的数据进行训练,模型复杂度高,训练时间长等[10-12]。针对这些问题,本文提出了一种基于动态集成加权概率随机森林(DEWPRF)的门诊量预测方法。DEWPRF 方法综合了随机森林的集成学习思想和加权概率的预测策略,能够在保持预测准确性的同时,降低模型的复杂度和训练时间,具有较好的实用性和可行性。

本文的研究主要基于前人在医院门诊量预测领域的研究成果,通过对传统方法和深度学习方法的分析和比较,提出了一种新的门诊量预测方法。具体来说,本文的研究目标就是提出基于 DEWPRF 的门诊量预测方法,该方法在保持预测准确性的同时,降低了模型的复杂度和训练时间。并在真实的门诊量数据集上进行实验验证,评估 DEWPRF 方法的预测效果和性能。最后分析 DEWPRF 方法的优缺点,并探讨其在实际应用中的潜在价值和改进方向。

综上所述,本文旨在为医院门诊量预测领域的研究提供新的思路和方法,促进医疗管理的智能化和信息化发展,为人们提供更加便捷高效的医疗服务。

1 基于随机森林的门诊量预测模型

由于医院门诊量具有强随机性与波动性,因此选择具有强泛化性的随机森林(Random Forest,RF)作为门诊量预测模型[13]。RF 是一种树状结构的集成学习方法,通过将多个决策树进行组合来做出预测,既能减少方差,又能保持低偏差近似不变。RF 的关键在于 2 个随机选择过程的应用。首先是 bootstrap aggregation,也称为 bagging。在这个过程中,RF 通过对原始训练数据进行置换随机抽样,生成若干个新的训练集,然后基于这些新训练集构建多个决策树,并将得到的预测结果进行组合。其次是随机子空间选择,这个过程在每个决策树节点上随机选择一个特征子集进行训练。这样做的目的是为了减少各个随机树之间的相关性,从而提高整体模型的泛化能力。

通过使用多个回归树来构建基于 RF 的门诊量预测模型,且每个树通过使用单独的回归来获得其自己的预测结果。基于 RF 的门诊量预测模型的预测流程如图 1 所示。在统计学中,bootstrapping 可以指任何依赖于随机抽样替换的测试或度量,该技术允许使用随机抽样方法来估计抽样分布,并且适合于小样本的情况。基于 RF 门诊量预测的实现过程如下:

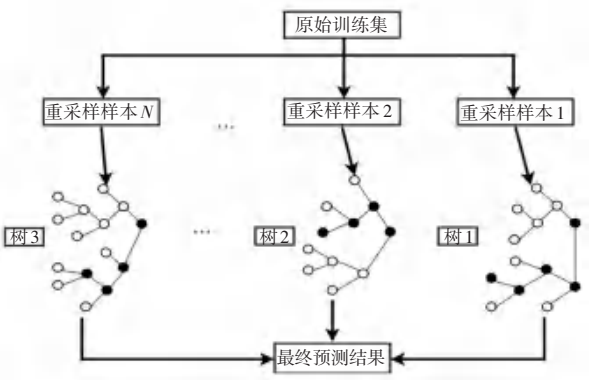


图 1 基于 RF 的门诊量预测模型的预测流程

Fig. 1 Prediction flow chart of outpatient volume prediction model based on RF

步骤 1 构建原始训练样本集,其中个数为 N,输入变量数为 m,该样本将作为生长树的训练集;

步骤 2 通过 bootstrap 方法,重复抽样 *ntree* 次,每次随机生成一个子训练集。因此,在进行 *ntree* 次 bootstrap 抽样后,得到了编号为 *ntree* 的随机森林中的 *ntree* 个子训练集;

步骤 3 在为每个非叶节点选择变量之前,从所有特征中随机选择一定数量的特征,将其作为当前决策树的分裂特征,并选择最佳的一个来分裂节点。每次拆分的变量数用 *mtry* 表示,$mtry \leqslant M$;

步骤 4 在树的生长过程中无需进行修剪,以最大程度促进树的生长;

步骤 5 RF 中的每棵树获取一个预测结果,并且门诊量的最终预测结果与每棵树的预测结果有关;

步骤 6 假设集合 S 包含 k 种特征值,每种特征值生成一个子节点,节点 i 的基尼系数计算如下:

$$Gini(i) = 1 - \sum_{j=1}^{h} [p(j/i)]^2 \tag{1}$$

其中,h 是节点 i 的类型数量,$p(j/i)$ 是节点 i 上类型编号 j 的相对频率。

集合 S 的拆分索引为:

$$Gini_{\text{split}}(S) = \sum_{i=1}^{r} \frac{s_i}{s} Gini(i) \tag{2}$$

其中，r 表示集合 S 中的记录类型；s_i 表示节点 i 上的记录号；S 表示集合 S 的总记录号。

节点分裂将在基尼指数最大程度降低处停止，此过程会遍历所有变量和节点。

最后，使用每个随机生成的树进行回归。通过对每棵树的所有预测结果使用合成投票，来决定最终的预测结果，即：

$$c = \underset{C}{\operatorname{argmax}}\left\{ \sum_{k=1}^{N} I[h(x,\theta_k) = c] \right\} \tag{3}$$

其中，c 是预测结果；N 是决策树的数量；$h(x,\theta_k)$ 是决策树；x 是训练样本；θ_k 是随机独立的特征向量。$I(A)$ 是指示函数，且当条件 A 为真时，$I(A)$ 的值为 1；当条件 A 为假时，$I(A)$ 的值为 0。

2　基于动态集成加权概率 RF 的门诊量预测模型

基于 RF 的预测模型具有较强的泛化性，在对波动性较小的对象进行预测时具有一定的优势。然而，门诊量具有强随机性与波动性，若直接使用原始 RF 模型进行门诊量预测，其性能具有一定的局限性，无法进行高精度的有效预测。为了获取性能更优的门诊量预测模型，本文对 RF 模型进行 2 方面改进：

(1)综合考虑测试样本之间的差异、随机森林中个体树的准确性、以及随机森林中树的多样性问题，基于动态集成选择，采用 k 近邻-层次聚类(k Nearest Neighbors-Hierarchical Clustering，KNN-HC)算法，对 RF 模型中树的强度进行评估，并从每个聚类结果中动态选择性能最佳的决策树，以此来对回归模型性能进行优化；

(2)为了提升 RF 模型的预测准确率，采用加权概率融合规则代替原始 RF 模型的求平均数的规则。

2.1　基于 KNN-HC 的动态集成选择

针对一组新的测试数据点 $\boldsymbol{x}_{\text{new}}$，计算其与样本集 $\boldsymbol{v}_i \in R^m (i = 1,2,\cdots,p)$ 之间的距离。P 在测试集 $\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_p\}$ 中，并且使用欧几里德距离作为距离度量，其计算公式如下：

$$ed(i) = \sqrt{(\boldsymbol{x}_{\text{new}} - \boldsymbol{v}_i)(\boldsymbol{x}_{\text{new}} - \boldsymbol{v}_i)^{\mathrm{T}}},\quad i = 1,2,\cdots,p \tag{4}$$

按照升序排序，选择前 k 个样本作为 x 的 k 个近邻，用于评估随机森林中树的强度，并进一步用于动态集成选择。这样的近邻选择可以帮助了解样本在数据空间中的局部结构，从而更好地评估随机森林模型的表现。

当 $\boldsymbol{\Psi}$ 确定时，集合 $\boldsymbol{\Psi}$ 中任意 2 棵树 c_i 和 c_j 之间的差异 d_{ij} 可以通过不一致度量来计算。不一致度量通常可以使用一致性度量的互补形式。这样的度量能够量化集成模型中不同树之间的差异程度，进而指导动态集成选择的过程，提高模型的稳健性和泛化能力。计算公式具体如下：

$$d_{ij} = \frac{b + c}{a + b + c + d} \tag{5}$$

其中，a 表示 $\boldsymbol{\Psi}$ 中 2 棵树正确回归的样本数；b 和 c 分别表示同时被一棵树正确回归和被另一棵树错误回归的样本数；d 表示被 2 棵树错误回归的样本数。

进而获得差异性矩阵 $\boldsymbol{D}$：

$$\boldsymbol{D} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1N} \\ d_{21} & d_{22} & \cdots & d_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N1} & d_{N2} & \cdots & d_{NN} \end{bmatrix} \tag{6}$$

其中，$\boldsymbol{D}$ 为对称矩阵，即 $d_{ij} = d_{ji}$。

以差异性矩阵 $\boldsymbol{D}$ 作为距离矩阵，可以得到这 N 棵树的层次聚类结果。选择每个聚类结果中性能最佳的决策树，从而形成样本数据 $\boldsymbol{x}_{\text{new}}$ 所选树的集合 EoC，并通过计算集合 EoC 中每棵树的预测准确性，确定权重向量 $\boldsymbol{W} = [w_1 \quad w_2 \quad \cdots \quad w_S]$：

$$w_i = \frac{acc_i}{\sum_{i=1}^{S} acc_i}, i = 1,2,\cdots,S \tag{7}$$

其中，w_i 表示第 i 棵树的权重；acc_i 表示 EoC 中第 i 棵树属于 $\boldsymbol{\Psi}$ 的准确率；S 表示样本 $\boldsymbol{x}_{\text{new}}$ 中树的总数。

2.2　加权概率融合规则

当测试样本 $\boldsymbol{x}_{\text{new}}$ 选取 S 棵决策树时，可以用加权概率融合策略代替取平均数规则来获取最终的预测结果。首先，通过测试集 $\boldsymbol{V}$ 对所选决策树的性能进行评估，并将每棵树的预测结果进行记录，其表示形式如下：

$$\boldsymbol{CM}^s = \begin{bmatrix} N^s_{11} & N^s_{12} & \cdots & N^s_{1M} \\ N^s_{21} & N^s_{22} & \cdots & N^s_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ N^s_{M1} & N^s_{M2} & \cdots & N^s_{MM} \end{bmatrix} \tag{8}$$

其中，$\boldsymbol{CM}^s$ 表示新随机森林中第 s 棵树的预测结果矩阵；M 表示测试集总类别数；S 表示回归树的数量；N^s_{ij} 表示测试集中门诊量预测结果。

基于贝叶斯理论，计算回归树预测结果的概率公式如下：

$$P^s(F_i \mid \boldsymbol{x}_{\text{new}}) = \frac{P^s(\boldsymbol{x}_{\text{new}} \mid F_i) P^s(F_i)}{\sum_{i=1}^{M} P^s(\boldsymbol{x}_{\text{new}} \mid F_i) P^s(F_i)} \tag{9}$$

其中，$s = 1,2,\cdots,S$。

如果 $\boldsymbol{x}_{\text{new}}$ 第 s 棵树的预测结果为 F_j，即 $E_k(\boldsymbol{x}_{\text{new}}) = j$，此时条件概率计算公式为：

$$P^s(\boldsymbol{x}_{\text{new}} \mid F_i) = \frac{N_{ij}^s}{\sum_{j=1}^{M} N_{ij}^s},\ i = 1,2,\cdots,M \quad (10)$$

其中，对于后验概率的计算，采用与先验概率 $P^s(F_i)$ $(i = 1,2,\cdots,M)$ 相同的值。

综上所述，通过加权概率随机森林，可以计算得到数据点 $\boldsymbol{x}_{\text{new}}$ 的预测结果，类别 F_i 的总概率为：

$$OP(F_i \mid \boldsymbol{x}_{\text{new}}) = \boldsymbol{W} \times \begin{bmatrix} P^1(F_i \mid \boldsymbol{x}_{\text{new}}) \\ P^2(F_i \mid \boldsymbol{x}_{\text{new}}) \\ \cdots \\ P^S(F_i \mid \boldsymbol{x}_{\text{new}}) \end{bmatrix} \quad (11)$$

最后，通过概率加权获取最终的门诊量预测结果。

基于对原始 RF 模型的改进，得到研究所需的基于动态集成加权概率 RF 的门诊量预测模型，其预测流程见图 2，主要步骤的阐释论述如下。

步骤 1 给定一个训练集 $\boldsymbol{X} = \{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$，其中 $\boldsymbol{x} \in R^m$，n 为训练集中的样本个数。然后，利用 N 个 C4.5 决策树 $\{t_1,t_2,\cdots,t_N\}$ 构建初始的原始随机森林模型；

步骤 2 基于 KNN-HC 的动态集成选择，选择前 k 个样本作为 x 的 k 个近邻，评估随机森林中树的强度，并利用这些近邻信息进行动态集成选择，获取各簇中性能最优的决策树用于构建回归模型；

步骤 3 采用加权概率融合规则，获取门诊量预测结果。

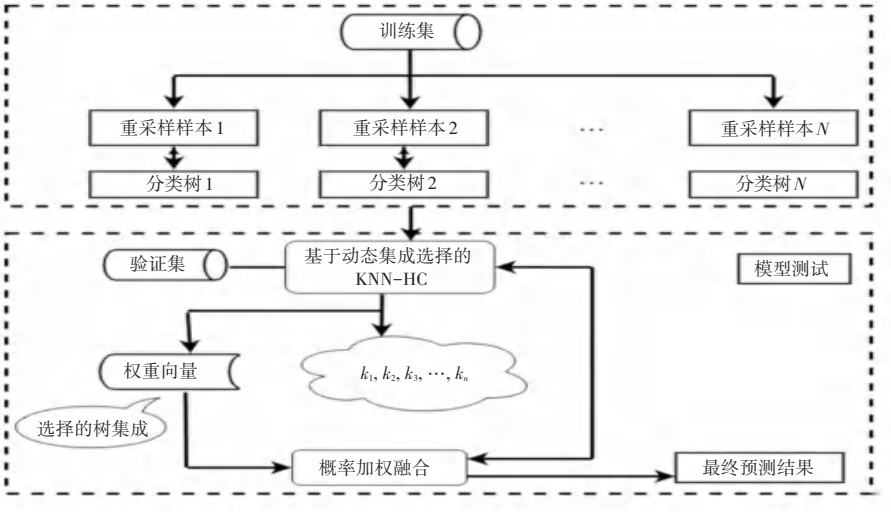


图 2 基于动态集成加权概率 RF 的门诊量预测流程

Fig. 2 Flowchart of outpatient volume prediction based on dynamic integrated weighted probability RF

3 算例分析

本文门诊量预测研究所需的数据为某三甲医院 2020 年全年门诊量的实测数据，数据分辨率为 15 min。进行门诊量预测时，将待预测日前 5 天的门诊量作为预测模型的输入，同时将门诊量实测数据集按照 4∶1 的比例分为训练集和测试集。为了验证提出的预测模型性能的优势，本研究采用绝对值平均误差（Mean Absolute Percentage Error，*MAPE*）、相对平均绝对误差（Relative Mean Absolute Error，*rMAE*）、相对均方根误差（Relative Root Mean Square Error，*rRMSE*）3 个预测精度指标进行评估，3 个指标的计算公式如下所示：

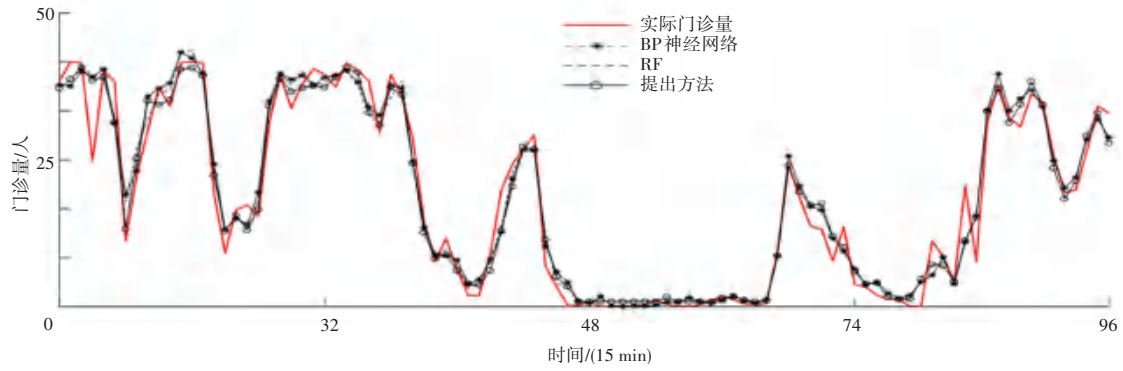
$$MAPE = \frac{1}{m}\sum_{i=1}^{m}\frac{|f_i - f_i'|}{f_i} \quad (12)$$

$$rRMSE = \frac{\sqrt{\frac{1}{m}\sum_{i=1}^{m}(f_i - f_i')^2}}{\frac{1}{m}\sum_{i=1}^{m} f_i} \times 100\% \quad (13)$$

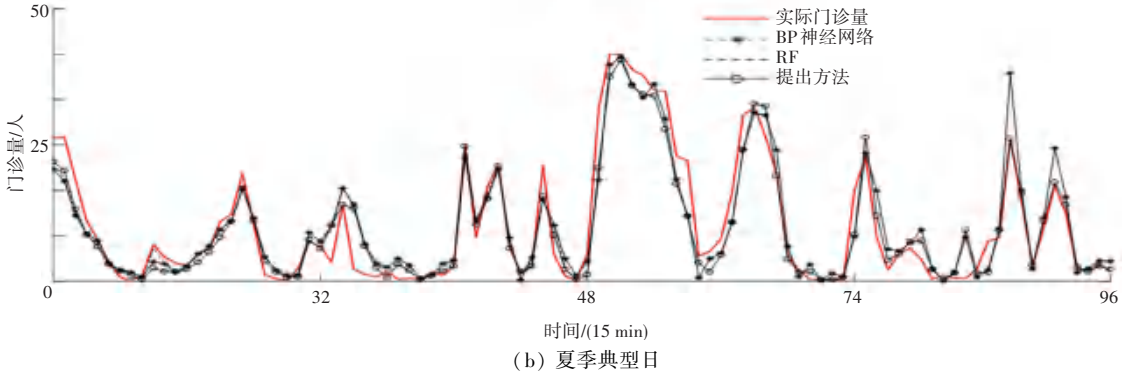
$$rMAE = \frac{\sum_{i=1}^{m}|f_i - f_i'|}{\sum_{i=1}^{m} f_i} \times 100\% \quad (14)$$

其中，f_i、f_i' 分别表示门诊量的真实值与预测值。

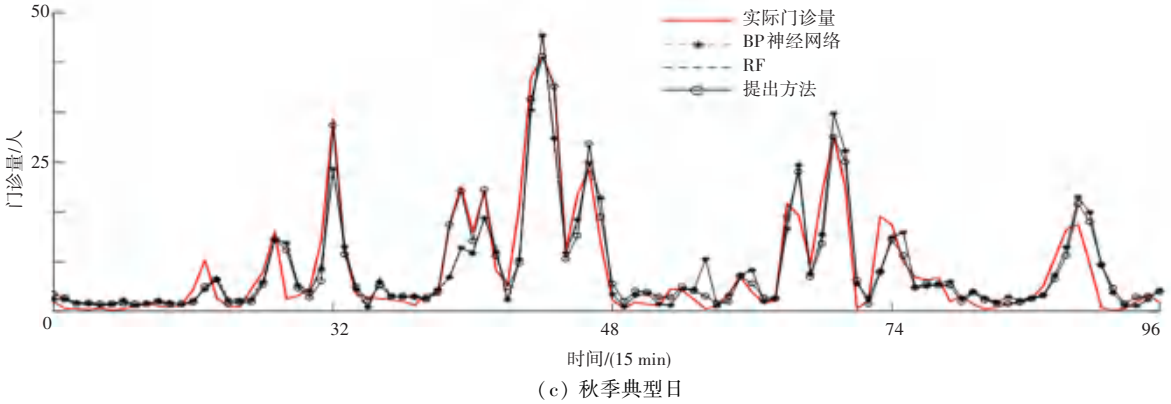
为了验证本文提出方法的有效性，将提出方法与 BP 神经网络模型及 RF 模型进行对比实验。采用 3 个预测模型分别进行门诊量预测后，从 4 个季节中各选择一个典型日展示各模型的门诊量预测结果，其结果如图 3 所示。



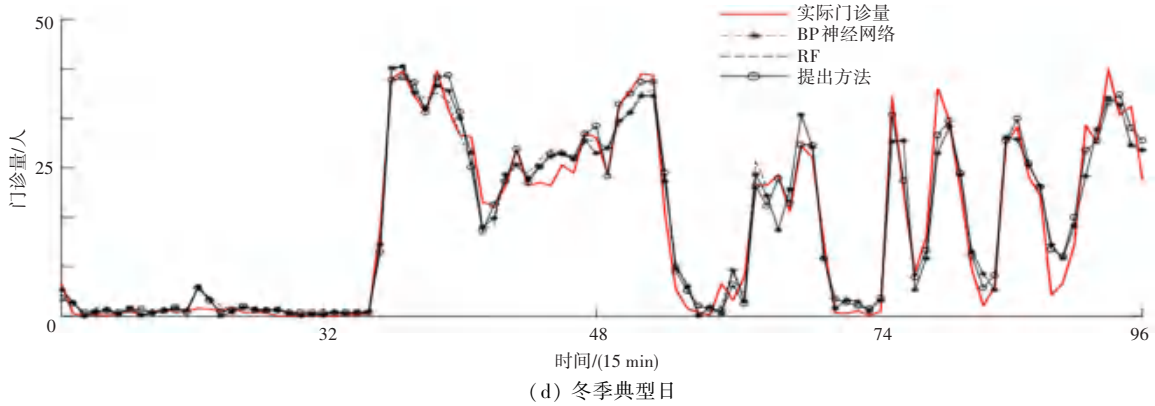
（a）春季典型日



（b）夏季典型日



（c）秋季典型日



（d）冬季典型日

图 3　各模型门诊量预测结果对比图

Fig. 3　Comparison of outpatient volume prediction results of each model

从图 3 可以看出：在 4 个季节中，本文提出方法获取的门诊量预测结果与实际值更贴合，并且预测效果优于对比预测模型，表明了所提方法的有效性。

为了进一步量化对比方法和提出方法门诊量预

测模型的性能,计算各季节典型日门诊量预测结果的 *MAPE*、*rMAE*、*rRMSE* 指标值,计算结果见表 1。从表 1 展示的门诊量预测结果可以看出,在春夏秋冬四个季节中,关于预测结果 *MAPE*、*rRMSE* 和 *rMAE* 评价指标的最优值,均为提出方法获取。同时,充分说明了提出方法在进行预测时,模型展现出较强的泛化性,有效地克服了门诊量强随机性对预测精度的影响,验证了提出方法的有效性。

表 1 各模型门诊量预测结果评价指标值

Table 1 Evaluation index values of outpatient volume prediction results of each model

季节	模型	评价指标		
		rRMSE/ MW	*MAPE* /%	*rMAE* /%
春季	BP 神经网络	5.439 1	5.46	2.12
	RF	6.496 0	6.89	2.56
	提出方法	3.162 9	3.57	1.89
夏季	BP 神经网络	5.462 8	6.04	5.68
	RF	3.547 0	4.65	3.95
	提出方法	3.236 1	3.23	2.16
秋季	BP 神经网络	7.215 4	7.57	5.24
	RF	6.250 8	5.84	4.26
	提出方法	5.192 6	4.12	3.14
冬季	BP 神经网络	5.935 3	4.95	3.84
	RF	4.998 2	5.12	4.58
	提出方法	2.809 6	2.98	2.17

4 结束语

针对医院门诊量预测的强非线性和复杂性问题,本研究提出了基于 RF 的医院门诊量预测模型。并且进一步采用 k 近邻-层次聚类算法和加权概率融合规则对 RF 模型性能进行优化。同时,经对比验证分析,验证了本文提出方法获取的门诊量预测结果精度更高。本次研究构建的门诊量预测模型,可以对医院运营管理提供科学的理论支撑和依据,同时也能够为医院管理者提供有效的决策支持。

参考文献

[1] 桑泉红,徐培文. 基于时序序列模型预测医院门诊人次[J]. 中国医院统计,2022,29(1):25-28.

[2] 焦晨,黄艳然,赵钦风,等. 时间序列分解模型在山东省糖尿病门诊量预测中的应用[J]. 中国农村卫生事业管理,2021,41(2):93-97.

[3] 刘焰,卢萍萍. 基于移动平均季节指数法的门诊量分析及预测[J]. 医学信息,2021,34(23):156-158.

[4] 陈辉,周雄辉,朱燕,等. 移动平均季节指数法在预测门诊量和出院人数中的运用[J]. 中国卫生统计,2012,29(2):312.

[5] 吴磊,徐凯. 基于深度神经网络的医院门诊量预测[J]. 微型电脑应用,2021,37(7):108-110,130.

[6] 唐路,宋萍,谢冰珏,等. 基于 R 语言的 ARIMA 乘积季节模型对重庆某儿童医院门诊量的预测分析[J]. 医学信息,2021,34(11):19-22.

[7] 张筠莉,杨祯山. 现代医院门诊量的灰色 RBF 神经网络预测[J]. 计算机工程与应用,2010,46(29):225-228.

[8] 王琦,郑静,吴清香,等. 灰色 GM(1,1)预测模型在门诊量预测中的应用[J]. 中国医院管理,2007(2):26-27.

[9] 胡蓉. 某院门诊量动态分析及预测[J]. 中国卫生事业管理,2010,27(S1):39-41.

[10] 许崇伟,沈俊学,邓光璞,等. 医院门诊量影响因素及预测方法[J]. 中国卫生经济,2015,34(3):74-75.

[11]杨旭华,钟楠祎. 基于深度信念网络的医院门诊量预测[J]. 计算机科学,2016,43(S2):26-30.

[12]黄美林,刘世科,胡丹标,等. 时间序列分解法在预防接种门诊接种量预测的应用[J]. 中国疫苗和免疫,2017,23(6):681-684.

[13]张珊珊,尚莉丽. 基于灰色系统理论的中医医院门诊工作量分析及预测研究[J]. 合肥学院学报(自然科学版),2013,23(4):24-28.

[14]马春柳,刘海霞,李小升,等. 灰色预测模型 GM(1,1)在医院门诊量预测中的应用[J]. 中国病案,2012,13(12):23-25.

[15]孔超. 基于灰色预测模型的门诊量预测—以上海市浦东新区门诊总量为例[J]. 中国卫生资源,2008(6):267-268,277.

[16]王鑫,吴际,刘超,等. 基于 LSTM 循环神经网络的故障时间序列预测[J]. 北京航空航天大学学报,2018,44(4):772-784.

[17]朱乔木,李弘毅,王子琪,等. 基于长短期记忆网络的风电场发电功率超短期预测[J]. 电网技术,2017,41(12):3797-3802.

[18]张群,唐振浩,王恭,等. 基于长短时记忆网络的超短期风功率预测模型[J]. 太阳能学报,2021,42(10):275-281.

[19]吕鑫,慕晓冬,张钧,等. 混沌麻雀搜索优化算法[J]. 北京航空航天大学学报,2021,47(8):1712-1720.

[20]张帅可,罗萍萍. 基于混合分布模型的风电功率超短期预测误差分析[J]. 电力科学与技术学报,2020,35(5):111-118.