数据驱动的心血管疾病门诊量多步组合预测研究

顾福来[1]，白朝阳[1,2]，郭林霞[1]，刘晓冰[1,2]，孙永亮[1]
（1. 大连理工大学经济管理学院，辽宁 大连 116024；
2. 大连理工大学制造管理信息化技术国家地方联合工程实验室，辽宁 大连 116024）

摘　要　精准的心血管门诊量预测是实现医院医生需求计算、医疗设备分配和管理的重要基础。本文基于心血管门诊量时间序列数据，采用迭代策略进行多步长时间序列预测，为降低多步长预测带来的误差累积问题和数据非平稳、非线性的特征，提出改进集合经验模态分解方法，并结合反向传播神经网络（back propagation neural network，BPNN）建立组合预测模型，达到更优的预测结果。最后以某医院心血管疾病月门诊量进行预测对比分析，实验结果表明，该组合预测模型对心血管门诊量多步长预测准确率较高，证实了模型的有效性。

关键词　心血管门诊量，多步预测，改进集合经验模态分解，BPNN

中图分类号　G311

1　引言

心血管疾病的患病率及就诊量持续上升，其已是我国致死人数最多的疾病[1]。提前预测心血管疾病的门诊量是门诊管理的基础，通过统计预测数据制订相应的工作计划，能有效避免主观盲目性，为医院管理决策提供客观依据，且具有较强可操作性[2]。在现有关于医疗患者预测的文献中，组合预测方法主要包含基于权重的组合预测模型和基于数据特征驱动分解集成方法。Luo 等[3]在每日时间序列上建立季节性 ARIMA（SARIMA）模型，然后在周时间序列上建立单指数平滑模型，并对模型进行改进以建立组合模型。该模型预测了成都一家医院两个内科门诊的每日就诊数据，实验结果具有较低的残差方差和较小的残差均值。王玖等[4]对某医院时间序列数据构建优化组合预测模型，并与趋势拟合法和 ARIMA 模型的预测结果进行比较分析，表明组合预测模型比单一预测模型具有更好的预测性能。Garg 等[5]考虑医院门诊量时间序列数据的事件离散化、基于频率密度的划分时间序列和优化方式创建模糊逻辑关系，并设计出基于模糊时间序列的模型，实验表明该模型具有较好的预测性能。Hadavandi 等[6]针对医院门诊量预测问题，通过人工神经网络（artificial neural networks，ANN）模型和遗传算法构建组合预测模型，提高门诊预测精度。Xu 等[7]对急诊部门的病人人次进行预测，提出了一种组合自回归综合移动平均线性回归（ARIMA-LR）预测方法，将 ARIMA 和逻辑回归（logistics regression，LR）顺序结合起来，具有捕获季节性趋势和预测变量影响的能力，实验表明 ARIMA-LR 混合模型在预测准确性方面优于现有模型。张筠莉和杨祯山[8]将针对医院门诊量时间序列构建出 RBF 神经网络和灰色 GM（1，1）的组合预测模型，获得了更好的预测结果。Huang 和 Wu[9]构建出经验模态分解（empirical mode decomposition，EMD）和通过粒子群算法优化的反向传播人工神经网络的组合预测模型，仿真表明，对于预测门诊就诊量所提出的方法具有更好的性能。门诊量时间序列数据通常是复杂的，组合预测模型可以更好地拟合不同的数据特征。朱顺痣等[10]针对医院门诊量时间序列具有的线性和非线性双重特点，构建神经网络和 ARMAX

两者的组合预测模型，分析厦门市医院门诊量时间序列数据的特征，研究发现组合预测模型能够较好地捕获时间序列数据的线性和非线性特征，表明组合预测模型比单一预测模型具有更好的预测性能。Zack 等[11]通过对 11 709 名患者临床参数建立了一个随机森林回归模型（即机器学习）来估计事件发生的时间，以识别经皮冠状动脉介入治疗（percutaneous coronary intervention，PCI）后死亡或因充血性心力衰竭（congestive heart failure，CHF）再次住院风险的患者，结果表明机器学习有潜力识别数据集中的复杂非线性模式，从而提高模型的预测能力。陈渝和任正军[12]针对医院门诊量时间序列，构建出 EMD-LSTM 组合预测模型，实验表明，分解和集成的思想具有更好的预测结果。

上述研究中的组合预测方法基本为单步预测（未来单个时间点的预测值）。在实际应用中，对中长期趋势的多步预测（未来连续多个时间点的预测）具有重大意义，可为医院的疾病长期预防管理工作提供重要基础。但随着步长的增加，建模难度急剧增加，易出现误差累积等问题[13]。同时，心血管疾病是一种影响因素较为复杂的疾病种类，门诊外地病人多、诊疗项目多、住院病人多，同时还受人口、行为及个体原因（高血压、吸烟、糖尿病等）等因素共同影响。由于影响因素的复杂性和多样性，以及缺乏关于这些疾病的概率性发作的知识和外源性因素，医院患者人次的时间序列数据呈现出非平稳、非线性的特征，因此，基于以往的时间序列预测模型预测心血管病门诊量的准确度还有待提高。

针对上述问题，本文利用集合经验模态分解（ensemble empirical mode decomposition，EEMD）方法在处理非平稳、非线性的复杂时间序列上的优势[14]，通过对复杂的数据进行平稳化预处理，进而将复杂的数据分解成一组性能较好、特征尺度差异较大的本征模态函数（intrinsic mode function，IMF）。在分解成若干个平稳化序列后，可以保留其主要数据特征，降低数据的预测难度，进而通过 BPNN 学习每个平稳化序列并预测，同时结合迭代策略多步预测心血管疾病的门诊量。经实验验证，该组合预测方法可以降低多步预测带来的误差累积问题的影响，达到更优的预测效果。

2 组合预测模型

2.1 EEMD

EEMD 方法[15]建立在 EMD 技术的基础上，通过将某一白噪声数据添加到原始数据当中，利用白噪声频谱均匀分布和零均值特性，使得原始数据当中的具有干扰的噪声被抵消甚至完全消除。EEMD 可以抑制 EMD 过程中出现的端点效应，能够根据数据的自有特征自动将原始数据更有效地分解成若干个 IMF，且每个 IMF 分量都含有原始数据在不同时间尺度下的局部特征信息。其思想就是将复杂的时间序列数据作为一组信号，对原始信号采用信号分解方法，将其分解成多尺度的若干特征简单信号，然后针对每一组特征简单的信号依据其特征选择合适的预测方法，最终将所有简单特征信号的预测结果集成为最后的预测值。因此，EEMD 方法在处理多尺度、非线性、非平稳数据上具有非常明显的优势，目前广泛应用在石油价格[16]、股指波动[17]、风速[18]、物料需求[19, 20]等方面预测，能很好地将复杂数据分解成多个简单平稳序列。

EEMD 的原理：假设数据包含原始数据和噪声两个部分，为避免模态混叠现象，将在原始数据中加入相关的白噪声序列，通过对多次分解得到的 IMF 值进行总体平均，从而抵消加入的白噪声，使结果更逼近实际值，具体算法如下。

（1）在原始数据 $x(t)$ 中，加入等长度的有限幅值的高斯白噪声 $w_i(t)$，可得

$$x_i(t)=x(t)+w_i(t),\ i=1,2,\cdots,N \tag{1}$$

其中，$x_i(t)$为第 i 次加入高斯白噪声 $w_i(t)$ 后的数据；N 为加入高斯白噪声的总次数。

（2）确定序列 $x_{\max}(t)$ 的所有极值点。采用三次样条插值方法拟合极大值点形成上包络线 $w_i(t)$ 和极小值点形成下包络线 $x_{\min}(t)$，则均值为

$$m_{11}(t)=\left[x_{\min}(t)+x_{\max}(t)\right]/2 \tag{2}$$

（3）令 $h_{11}(t)=x_i(t)-m_{11}(t)$，如果 $h_{11}(t)$ 满足 IMF 定义条件，则 $c_{11}(t)=h_{11}(t)$，如果不满足，则将 $h_{11}(t)$ 作为 $x_i(t)$ 重复以上过程，直到满足，得到第一个 IMF 分量：

$$c_{11}(t)=h_{1k}(t)-m_{1k}(t) \tag{3}$$

对序列 $x_i(t)$ 剩余差值序列 $r_1(t)$ 重复以上过程，直到 $c_{1n}(t)$ 满足终止条件。

（4）最后得到若干个 IMF 分量和一个总体趋势的序列 $r_i(t)$，则多次添加高斯白噪声并经 EMD 分解后，$x_i(t)$ 可写为

$$x_i(t)=\begin{bmatrix}x_1(t)\\x_2(t)\\\cdots\\x_i(t)\end{bmatrix}\xrightarrow{\text{EMD}}\begin{bmatrix}c_{1,1}(t) & c_{1,2}(t) & \cdots & c_{1,j}(t) & r_1(t)\\c_{2,1}(t) & c_{2,2}(t) & \cdots & c_{2,j}(t) & r_2(t)\\\cdots & \cdots & & \cdots & \cdots\\c_{i,1}(t) & c_{i,2}(t) & \cdots & c_{i,j}(t) & r_i(t)\end{bmatrix} \tag{4}$$

其中，$c_{i,j}(t)$ 为第 i 次加入白噪声后，分解得到的第 j 阶 IMF 分量，$r_i(t)$ 为第 i 次加入白噪声后，分解得到的余量。

（5）假设叠加白噪声的次数为 N，则多次分解取平均后，得到 EEMD 最终的 IMF 分量为

$$c_j(t)=\frac{1}{N}\sum_{i=1}^{N}c_{i,j}(t) \tag{5}$$

（6）EEMD 分解后，原始数据最终可表示为

$$x(t)=\sum_{j=1}^{M}c_j(t)+r_m(t) \tag{6}$$

其中，M 为 EEMD 方法针对实验的数据的特征利用自动分解得到的 IMF 个数，$r_m(t)$ 为最终剩余分量。

对于采用 EEMD 分解得到的 IMF 或趋势项，N 次加入的白噪声经过求均值后相互抵消了，即由于现实中噪声的随机性和模拟的白噪声的随机性，两者产生叠加效果，就会使噪声对冲，但是结果还是随机的，消除了分解过程中产生的随机误差。

但是在 EEMD 分解时容易产生端点效应，即时间序列数据的端点并不是极值点，使用三次样条插值法使数据形成上、下包络线时，导致端点处的数据出现下包络线超过上包络线，最终导致端点处的波形失真。如果待处理的数据时间尺度大，失真现象还会向中间部分延伸，造成最终的 IMF 准确性受到很大影响，严重时甚至会使最终分解出的 IMF 失去意义。因此，本文采用极值延拓法进行处理，抑制端点效应的产生[21]，优化分解数据的能力，提高预测的效果。

2.2 改进 EEMD

端点延拓的目的是确保上、下包络都与端点相交，以便有与每一个数据点相对应的局部平均值。上、下包络是由极大值和极小值联结而成的，因此只要对极大值和极小值进行延拓，而不必对数据本身进行延拓。极大值和极小值是相间分布的，考虑到样条插值的要求，所以只要在数据左、右两端分别延拓两个极大值和两个极小值即可。

计算过程：

设时间序列为 $x(t)$，t 是时间序列 $x(t)$ 对应的时间信息，Δt 为采样时间长度，$x(t)$ 存在 M 个极大值和 N 个极小值，则时间序列的极值对应的下标为 (I_m, I_n)，对应的时间信息为 (T_m, T_n)，对应的函数值为 (U, V)。

计算时间序列左端第一个特征波包含的数据点数 k_1：

$$k_1 = \begin{cases} I_m(2) - I_m(1), & \text{当} I_m(1) < I_n(1) \\ I_n(2) - I_n(1), & \text{当} I_m(1) > I_n(1) \\ 2\left|I_m(1) - I_n(1)\right|, & \text{当} m = n = 1 \end{cases} \tag{7}$$

向外延拓两个极值位置为 (T_m, T_n) 和函数值 (U, V)，

$$x_m(0) = x_m(1) - k_1 \Delta x \tag{8}$$

$$U(0) = U(1) \tag{9}$$

$$x_n(0) = x_n(1) - k_1 \Delta x \tag{10}$$

$$V(0) = V(1) \tag{11}$$

1. 同理，计算时间序列右端延拓的位置为 (T_m, T_n) 和函数值 (U, V)。

2. 当端点的数值比近端点的第一个极大值大或极小值小时，要进行特殊的处理，认为端点处的值为极值点，以避免数据落到包络线之外，左端公式如下：

$$T_m(0) = t_1,\ U(0) = x_1,\ \text{当} x_1 > U(1) \tag{12}$$

$$T_n(0) = t_1,\ V(0) = x_1,\ \text{当} x_1 < U(1) \tag{13}$$

以端点的一个特征波为依据进行延拓，分别在时间序列数据两端增加两个极大值和两个极小值，从而使原始数据被延长的包络线所限制，有效地抑制了端点效应，使得 EEMD 分解得到合理的各个 IMF 模态。

以本文实验选取的心血管疾病门诊量数据为例，对该数据进行分解，表 1 是部分时间序列通过 EEMD 和极值延拓法改进的 EEMD 分解对端点效应的影响对比。

表 1 EEMD 和改进 EEMD 对比

原始数据	EEMD	改进 EEMD
7 088	40.94	0.64
9 701	25.07	15.24
10 681	20.89	3.32
9 199	13.26	17.18
10 645	30.96	1.70
9 681	10.53	2.15
10 287	2.49	8.31
10 632	14.29	19.10
…	…	…
12 141	5.26	16.18

续表

原始数据	EEMD	改进 EEMD
11 542	21.68	17.37
10 689	10.22	5.17

从表 1 中可以看出两者的分解都会产生一定的误差，形成一定的前后变化趋势，改进的 EEMD 方法并不是每一个数据点的结果都优于未改进，如第 4、8 个点等，原因可能是极值延拓法仅是在序列两端增加极值，并未改变原始序列的特征，此外，这些时间节点由于受季节性影响，波动频繁，加入白噪声序列后，部分分解结果误差可能会大于直接分解结果误差，因此不能明显地看出极值延拓法改进的 EEMD 要优于 EEMD 分解的结果，基于此计算两者的平均变化趋势，计算公式如下：

$$\theta=\sum_{j=1}^{N}\frac{\left|\sum_{i=1}^{m}\mathrm{IMF}_i+\mathrm{RES}-x\left(t_j\right)\right|}{x\left(t_j\right)} \tag{14}$$

分别计算两者趋势变化的平均值，可以发现$\theta\geqslant 0$。如果$\theta=0$，则表明 EEMD 不会产生任何端点效应。θ越大，端点效应的影响越大。通过计算，EEMD 的值为 0.099 6，而改进 EEMD 的值为 0.083 6，可以看出，基于极值延拓法的集合模态分解的θ值结果最小，对端点效应的抑制结果较好，证明了所提方法的有效性。

2.3 BPNN

BPNN 是由 Rumelhart 于 1986 年提出的一种利用误差反向传播训练算法的多层网络模型。1989 年，Hecht-Nielson 证明了具有输入层、隐含层和输出层的三层 BP 网络（图 1）可以完成任意 N 维到 M 维的映射，对于非线性复杂时间序列具有良好的拟合能力，且具有计算简单和无须先验假设数据的属性[22]。

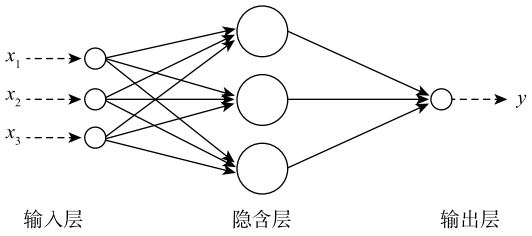

图 1　三层神经网络结构图

2.4 组合预测模型的建立

时间序列数据的分析与预测，其基本思想都是收集历史数据进行统计分析，通过模型表征历史数据之间的规律泛化到未来的时间段内。本文的心血管疾病门诊量调研数据是非线性、非平稳的时间序列数据，如果直接对该数据进行预测，结果会产生较大偏差。基于此，提出组合预测模型[23]，首先为避免分解时间序列数据存在端点效应，采用极值延拓法对时间序列向外延拓极值点，能够抑制 EEMD 分解过程中的端点效应问题，得到新的时间序列数据并进行分解，得到包含各个时间尺度特征的 IMF 和一个趋势项（RES），这样将原本复杂的时间序列分解成多个简单的平稳性序列，各个分量之间的频率不同，也就是说各个分量包含的特征信息是不一样的。之后将每个分量输入

BPNN 中进行组合预测，提高预测精度，该组合预测模型的结构图如图 2 所示。

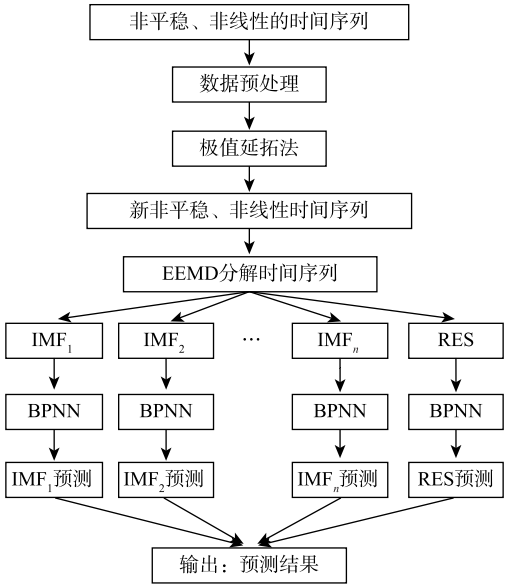

图 2 组合预测模型结构图

3 数值检验

3.1 实验数据

本文以某大型三甲医院心血管月门诊量数据为研究对象，从该院 Oracle 数据库 HIS 挂号系统中获取了从2012年1月至2017年12月共72个月的门诊量数据，并且该数据库中不存在数据缺失的现象。经初步处理后，通过传统的统计方法得到了数据的基本特征（表 2），以及心血管门诊量折线图（图 3）。通过图表可以发现，心血管疾病门诊量序列的样本近似于正态分布，极差和标准差都相对较大，可推断出数据的波动性很大，说明每月的门诊量可能因各种因素影响存在较大的波动，符合非平稳、非线性的复杂时间序列特征。

表 2 心血管门诊量描述性统计

最大值	最小值	平均值	中位数	偏度	峰度	标准差
13 408	7 088	10 696	10 683	−0.330	0.822	1 239.74

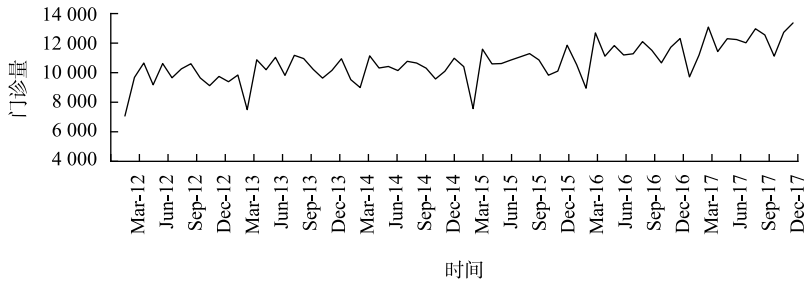

图 3 2012~2017 年心血管门诊量

3.2 多步预测方法

本文采用迭代策略对时间序列进行多步预测，迭代策略通过预测模型进行最小化单步预测并计算误差，然后迭代地运用该模型进行多步预测，在这个过程中，前期的预测值作为输入参与后期的预测，如图 4 所示，时间序列长度为$N=72$，设置长度为 w 的时间窗口，每一次预测后，都将时间窗口向后滑动一个，将预测值加入其中并保证时间窗口长度不变，选取第二次预测的训练样本，反复操作，完成多步预测。

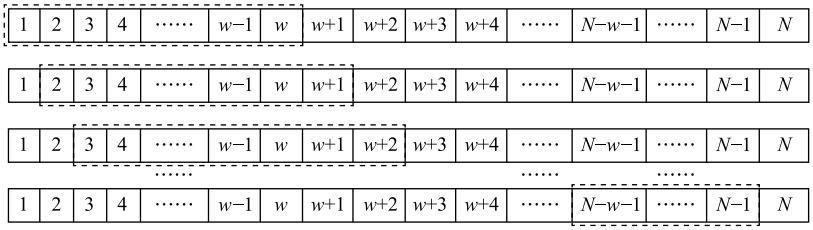

图 4 迭代策略解释图

由于预测值作为输入参与接下来的预测过程，迭代策略易产生误差累积的问题，即随着预测步长的增加，积累的预测误差会越来越大，从而明显地降低模型的多步预测能力，因此，采用 EEMD 分解−组合的预测原理，先使复杂数据简单化，弥补上述问题。

3.3 实验流程与评价指标

实验流程（图 5）显示：首先输入心血管门诊时间序列，确定预测步长，并将数据划分为训练集和测试集，将训练集进行分解与模型训练，预测下一个值，最后采用迭代策略进行多步预测，并进行评价指标计算。

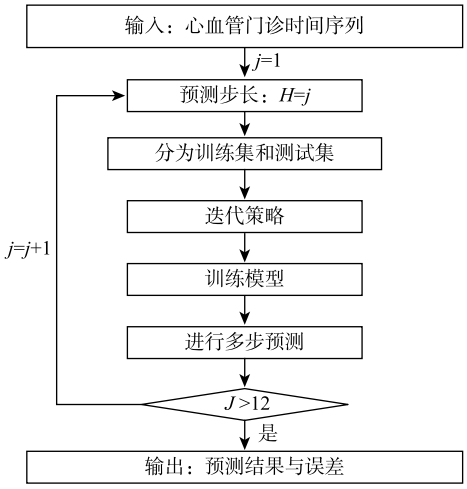

图 5 预测流程图

为检验最终预测结果，本文引入均方根误差（root-mean-square error，RMSE）、平均绝对百分比误差（mean absolute percentage error，MAPE）、平均绝对误差（mean absolute error，MAE）对预测效果进行评价，计算公式为

$$\mathrm{RMSE}=\sqrt{\sum_{i=1}^{n}\left(y-y_i\right)^2/n} \tag{15}$$

$$\mathrm{MAPE}=1/n\sqrt{\sum_{i=1}^{n}\left(\left|y-y_i\right|\right)/y} \tag{16}$$

$$\mathrm{MAE}=1/n\sum_{i=1}^{n}\left|y-y_i\right| \tag{17}$$

3.4 实验结果与对比分析

由于该医院通常以五年为其医疗管理规划的时间节点，因此本文依据上述的实验步骤，以 2012~2016 年共计 60 个月的心血管疾病的月门诊量作为初始学习集。时间窗口长度过小会影响预测精确度，因此将 w=60 作为初始的滑动窗口，将此作为基础进行改进的 EEMD 分解，以减小后续的误差累积。如图 6 所示，初始的学习集，经极值点延拓后，共计 64 个数据点，从上到下依次为：原始数据、IMF_1、IMF_2、IMF_3、IMF_4、IMF_5 和 RES，可以看出 IMF 分量在零点上下波动，具有明显的周期性变化且波动趋势各不相同。随着时间的变化，每个 IMF 分量表现出强弱不一的非均匀性变化，在年际（一年内）尺度上，存在大概 3 个月（IMF_1）、6 个月（IMF_2）和 12 个月（IMF_3）的周期变化，在年代（多年内）尺度上，存在大概 30 个月（IMF_4）和 42 个月（IMF_5）的周期性变化，表明这些 IMF 分量的不同时间尺度的周期性波动含有心血管疾病门诊量的外在影响因素周期性变化，而且也包含一些影响因素的非线性的反馈作用；也可以看出 RES 分量呈现递增的趋势，分解的效果良好，而且也验证了心脏病的门诊量呈现增加趋势。

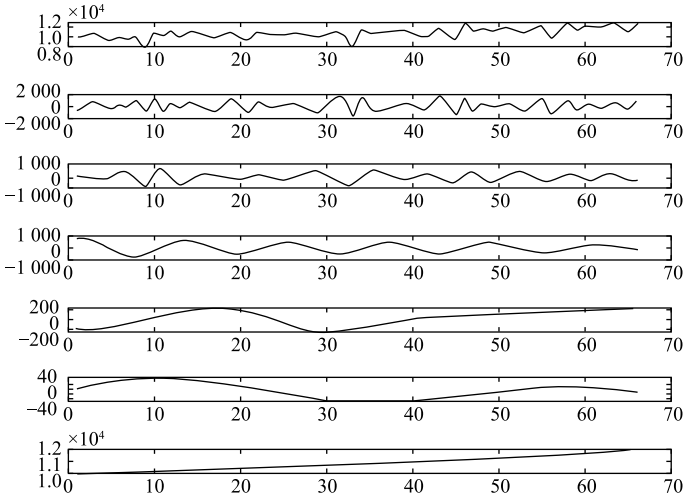

图 6 改进 EEMD 分解图

为抑制 EEMD 分解过程中的端点效应，采用极值点延拓法后将会分解得到 64 × 16 的矩阵，而实际数据并没有两端的四个极值点，因此后续步骤的预测，去掉因四个极值点而产生的矩阵，即使用 60×6 的矩阵进行预测。本文统计了初始窗口的统计指标进行说明，如表 3 所示。

表 3 改进 EEMD 分解结果描述统计

分量	最大值	最小值	峰度	偏度	均值
IMF_1	1 522.45	−1 633.42	−0.33	0.07	−7.60

续表

分量	最大值	最小值	峰度	偏度	均值
IMF_2	716.80	−728.28	−0.06	−0.25	10.27
IMF_3	851.85	−713.96	−1.08	0.08	47.65
IMF_4	353.08	−313.75	−0.76	0.36	−32.86
IMF_5	307.45	−238.29	−1.59	−0.16	52.00
RES	11 928.9	9 234.07	−1.08	0.37	10 347.51

从表 3 中还可以看出 IMF_1 波动频率较高；IMF_2 和 IMF_3 波动频率居中，数据的离散程度也较为接近且与原时间序列数据的相关性较强；IMF_4 和 IMF_5 波动频率较低，数据的离散程度也较为接近；从 EEMD 分解可以得到，IMF_1、IMF_2、IMF_3、IMF_4、IMF_5 及残差项对于原始时间序列均有隐含的统计学解释，能够说明其波动原因随某一特征而变化，最后从表和图都可以看到残差项（RES）的趋势是长期上升，也反映出实际情况，长期来看心脏病门诊量每年都在增加。

最后，将各分量输入 BPNN 预测模型中得到对应预测值，对各子序列的预测结果进行线性累加集成，以获得初始时间序列数据的预测值。通过迭代策略，预测未来一年（预测步长 H=12）的门诊量，结果如表 4 所示。

从表 4 和图 7 可以看出，随着时间的增加，预测误差也在变大，但在中间阶段预测误差在变小，主要原因可能是 2017 年 6 月国家提出医疗改革政策等，因此造成了门诊量发生了变化，而模型不能有效地识别这一政策，造成这段时间预测误差变化较大，但总体来说，预测误差随着预测步长的增加而累积，各项指标呈现先上升后下降之后再上升的趋势，而从表 5 的统计指标来看，并没有因这一现象而造成各种指标出现明显大的改变。

表 4　预测实验结果表

时间/月份	真实值/人次	预测值/人次	误差
1	9 733	10 143.57	−410.13
2	11 214	11 839.46	−625.03
3	13 127	12 398.69	728.69
4	11 465	10 683.92	781.40
5	12 329	11 505.72	823.32
6	12 270	12 148.47	121.38
7	12 057	11 974.62	81.93
8	13 013	12 158.91	853.70
9	12 588	12 158.91	428.66
10	11 141	12 240.04	−1 099.38
11	12 760	12 524.7	235.43
12	13 408	12 300.79	1 106.86

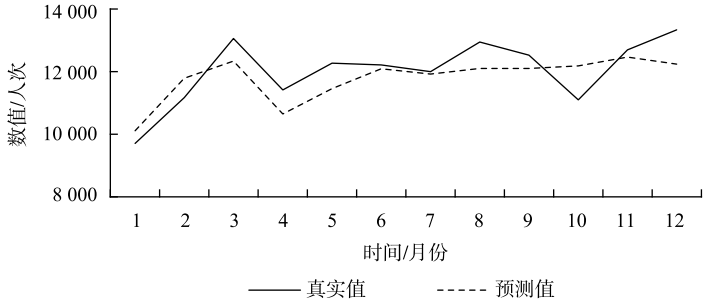

图 7 预测实验结果比较图

表 5 预测实验结果评价表

步长	MAE	MAPE	RMSE
1	410.57	4.22%	410.57
2	518.01	4.90%	529.04
3	588.11	5.11%	602.83
4	636.35	5.54%	651.97
5	673.74	5.77%	689.65
6	581.70	4.97%	631.51
7	510.37	4.36%	585.49
8	553.34	4.63%	625.41
9	539.53	4.50%	606.74
10	595.48	5.03%	672.39
11	562.74	4.74%	645.01
12	608.11	5.04%	695.36

本文将 ARIMA、BPNN、EEMD-BPNN 和改进 EEMD-BPNN 四个模型，以步长为 3、6、10 分别进行对比，其结果如表 6 所示。从整体来看，采用迭代策略进行预测，各步长的结果相对比较稳定；在步长为 6 时，因政策等各种不稳定因素的变化，各种模型在该处拟合结果较好，各项指标数据在下降，同时影响到对未来步长的预测；最后可以看出，相较于其他几种预测模型，本文提出的改进 EEMD-BPNN 模型的各项评价指标更优，对降低迭代策略对未来多步长的预测结果产生的误差累积问题效果更优，具有较好的预测性能。

表 6 不同模型对比评价表

方法	指标	Step=3	Step=6	Step=10
ARIMA	MAE	1 464.69	1 058.37	780.82
	MAPE	13.8%	9.9%	7.1%
	RMSE	1 640.04	1 328.51	1 054.31
BPNN	MAE	1 419.63	1 335.26	1 400.28
	MAPE	12.9%	11.5%	11.7%
	RMSE	1 661.35	1 565.34	1 626.46

续表

方法	指标	Step=3	Step=6	Step=10
EEMD-BPNN	MAE	780.23	800.12	813.67
	MAPE	7.41%	6.97%	7.87%
	RMSE	897.45	845.23	906.23
改进 EEMD-BPNN	MAE	588.11	581.70	595.48
	MAPE	5.11%	4.97%	5.03%
	RMSE	602.83	631.51	672.39

4 总结

本文采用迭代策略对心血管门诊量多步长时间序列预测问题进行研究，为降低数据非平稳性对多步长预测结果的影响，利用极值延拓法抑制 EEMD 的端点效应，将原始序列分解为若干 IMF 分量和趋势项，更好地表征数据信息，并基于 BPNN 对时间序列分量数据进行预测，得到最终的预测结果。本文以某医院心血管门诊量为研究对象，并与 ARIMA、BPNN 预测模型进行效果对比，验证了改进 EEMD-BPNN 组合预测的有效性，能为医院对门诊量预测提供参考模型；同时也对医疗资源管理和分配、医护人员的安排、就诊路径研究提供支撑；为医院诊疗工作实现智能化、细致化和高效化提供重要的理论基础。

参 考 文 献

[1] 胡盛寿，高润霖，刘力生，等.《中国心血管病报告 2018》概要[J]. 中国循环杂志，2019，34（3）：209-220.

[2] 白贺伊. 基于数据挖掘理论的心血管疾病预警方法建模[J]. 信息技术，2020，44（2）：53-57.

[3] Luo L，Luo L，Zhang X L，et al. Hospital daily outpatient visits forecasting using a combinatorial model based on ARIMA and SES models[J]. BMC Health Services Research，2017，17（1）：469.

[4] 王玖，韩春蕾，栾奕昭. 组合预测在医院门诊量预测中的应用[J]. 中国卫生统计，2012，29（6）：881-883，886.

[5] Garg B，Beg M M S，Ansari A Q. A new computational fuzzy time series model to forecast number of outpatient visits[C]. IEEE，2012.

[6] Hadavandi E，Shavandi H，Ghanbari A，et al. Developing a hybrid artificial intelligence model for outpatient visits forecasting in hospitals[J]. Applied Soft Computing，2012，12（2）：700-711.

[7] Xu Q N，Tsui K L，Jiang W，et al. A hybrid approach for forecasting patient visits in emergency department[J]. Quality and Reliability Engineering International，2016，32（8）：2751-2759.

[8] 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.

[9] Huang D Z，Wu Z H. Forecasting outpatient visits using empirical mode decomposition coupled with back-propagation artificial neural networks optimized by particle swarm optimization[J]. PLoS ONE，2017，12（2）：1-17.

[10] 朱顺痣，王大寒，何亚男，等. 基于时间序列模型的医院门诊量分析与预测[J]. 中国科学技术大学学报，2015，45（10）：795-803.

[11] Zack C J，Senecal C，Kinar Y，et al. Leveraging machine learning techniques to forecast patient prognosis after percutaneous coronary intervention[J]. JACC-Cardiovascular Interventions，2019，12（14）：1304-1311.

[12] 陈渝，任正军. 融合 EMD 与 LSTM 神经网络的门诊量预测模型研究[J]. 软件导刊，2019，18（3）：133-138.
[13] Franses P H，Legerstee R. A unifying view on multi-step forecasting using an autoregression[J]. Journal of Economic Surveys，2010，24（3）：389-401.
[14] Wang Y M，Gu J Z，Zhou Z L，et al. Diarrhoea outpatient visits prediction based on time series decomposition and multi-local predictor fusion[J]. Knowledge-Based Systems，2015，88：12-23.
[15] Wu Z H，Huang E. A study of the characteristics of white noise using the empirical mode decomposition method[J]. Proceedings of the Royal Society A：Mathematical，Physical and Engineering Sciences，2004，460（2046）：1597-1611.
[16] Zhu B Z，Shi X T，Chevallier J，et al. An adaptive multiscale ensemble learning paradigm for nonstationary and nonlinear energy price time series forecasting[J]. Journal of Forecasting，2016，35（7）：633-651.
[17] 李合龙，冯春娥. 基于 EEMD 的投资者情绪与股指波动的关系研究[J]. 系统工程理论与实践，2014，34（10）：2495-2503.
[18] Wang S X，Zhang N，Wu L，et al. Wind speed forecasting based on the hybrid ensemble empirical mode decomposition and GA-BP neural network method[J]. Renewable Energy，2016，94：629-636.
[19] 白朝阳，宋林杰，李晓琳. 基于 EMD-PSO-LSSVR 的物料需求组合预测模型[J]. 统计与决策，2018，34（18）：185-188.
[20] 白朝阳，胡子涵，刘晓莹. 面向装备制造业的非平稳时间序列需求组合预测方法[J]. 信息与控制，2017，46（4）：495-502.
[21] 马小刚，王永平，杜历. 一种求解信号包络曲线端点值的新方法[J]. 噪声与振动控制，2018，38（4）：159-164.
[22] 史忠植. 神经网络[M]. 2 版. 北京：高等教育出版社，2009.
[23] 凌立文，张大斌. 组合预测模型构建方法及其应用研究综述[J]. 统计与决策，2019，35（1）：18-23.

Research on Data-driven Multi-step Combined Forecast of Cardiovascular Disease Outpatient Volume

GU Fulai[1]，BAI Zhaoyang[1, 2]，GUO Linxia[1]，LIU Xiaobing[1, 2]，SUN Yongliang[1]
（1. School of Economics and Management，Dalian University of Technology，Dalian 116024，China；
2. National and Local Joint Engineering Laboratory for Manufacturing Management Information Technology，Dalian 116024，China）

Abstract Precise cardiovascular outpatient volume forecasting is an important basis for realizing outpatient management such as doctor demand calculations and medical equipment management and allocation. In this paper，based on the time series data of cardiovascular outpatient volume，multi-step time series prediction is carried out by iterative strategy. In order to reduce the error accumulation caused by multi-step prediction and the non-stationary and non-linear characteristics of data，an improved ensemble empirical mode decomposition method is proposed，and a combined prediction model is established with back propagation neural network （BPNN）to achieve better prediction results. Finally，the monthly outpatient volume of cardiovascular diseases is used for prediction and comparative analysis. The experimental results show that the combined prediction model has high accuracy in predicting the cardiovascular outpatient volume with multiple steps，which proves the effectiveness of the model.

Keywords cardiovascular disease outpatient volume，multi-step prediction，ensemble empirical mode decomposition，BPNN

作者简介

顾福来（1981—），男，大连理工大学博士研究生，主要研究方向为医疗大数据、机器学习，E-mail：GFL128@126.com。

白朝阳（1978—），男，大连理工大学经济管理学院副教授、硕士生导师，主要研究方向为精益医疗、医疗大数据，E-mail：baizhaoyang@dlut.edu.cn。

郭林霞（1996—），女，大连理工大学硕士研究生，主要研究方向为医疗大数据、机器学习，E-mail：3029085728@qq.com。

刘晓冰（1956—），男，大连理工大学经济管理学院教授、博士生导师，主要研究方向为精益医疗、医疗大数据，E-mail：xbliu@dlut.edu.cn。

孙永亮（1995—），男，大连理工大学硕士研究生，主要研究方向为医疗大数据、机器学习，E-mail：2239314954@qq.com。