doi:10. 16055/j. issn. 1672-058X. 2021. 0002. 009

基于非结构数据和 EMD-WTS 二层分解的 AQI 组合预测方法*

刘金培[1,2], 张了丹[1**], 丁　蓉[1], 汪　漂[1], 罗　瑞[1]

(1. 安徽大学 商学院,合肥 230601;
2. 北卡罗莱纳州立大学 工业与系统工程系,美国 罗利 27695)

摘　要:针对具有高复杂性与非平稳性的空气质量指数(AQI)时间序列,提出一种融合非结构数据的 EMD-WTS 二层分解组合预测模型;首先,筛选百度指数关键词并提取对应数据,运用局部线性嵌入算法(LLE)对之降维;其次,对 AQI 历史序列与降维结果进行经验模态分解(EMD)与重构;接着,对所得高频项进行小波分解(WT)与重构;然后,运用 Holt 指数平滑法、支持向量回归(SVR)与人工神经网络(ANN)分别对二层分解结果与原始低频、趋势项进行组合预测并运用 BP 神经网络集成;最后,叠加集成结果得到 AQI 预测值;对比实验说明预测方法充分利用了多源数据信息,具有较高的预测精度。

关键词:组合预测;空气质量指数;EMD-WTS 二层分解;非结构数据

中图分类号:O212　　**文献标志码**:A　　**文章编号**:1672-058X(2021)02-0056-08

0　引　言

近年来,我国空气污染问题日趋严重[1]。通过监测包括 $PM_{2.5}$、SO_2、NO_2 等在内的空气污染物,空气质量指数(Air Quality Index, AQI)能够全面地反应空气污染状况。因此,AQI 的有效预测有利于政府制定科学的环境保护政策,对于维护居民健康、改善生态环境而言意义重大。

已有预测模型大多仅以历史数据作为输入。然而,由于历史数据的收集与公布通常具有滞后性,模型的预测精度往往受到限制。此时,结合具有即时性的非结构数据能够弥补由历史数据预测带来的预测滞后性,从而增强预测结果的时效性。非结构数据主要指网络搜索数据。张玲玲等[2]将非结构数据融入旅游市场客流量预测模型,Francesco 等[3]结合非结构数据与失业预测模型,陈声利等[4]将非结构数据引入股指期货波动率预测模型。上述研究均证实了非结构数据对于预测效果起正向促进作用,但目前将非结构数据应用于空气质量指数预测的研究仍为少见。由于空气质量相关关键词的网络搜索热度能够反映社会公众对空气质量关注程度的改

收稿日期:2020-03-26;**修回日期**:2020-04-28.

* **基金项目**:国家自然科学基金(71871001, 71901001, 71771001, 71701001);教育部人文社科研究规划基金项目(20YJAZH066);安徽省高校人文社科基金重点项目(SK2019A0013);安徽大学大学生创新训练计划项目(201910357704).

作者简介:刘金培(1984—),男,山东滨州人,教授,博士,从事预测与决策分析研究.

** **通讯作者**:张了丹(1999—),女,浙江温州人,从事预测与决策分析研究. Email: zhangliaod@163. com.

变,可以对空气质量的变化起到一定的预见与解释作用。因此,尝试将网络搜索数据融入 AQI 预测框架,进一步改善预测效果。

非结构数据能弥补历史数据的不足,但同时增加了 AQI 预测模型输入数据的复杂性,同时 AQI 时序具有随机性、非平稳性等特征[5]。为解决上述问题,传统研究多采用单一分解方法对分解数据以有效获取数据所含有效信息,进而提高预测精度。已有研究表明,相较于单一分解方法模型,二层分解模型能够充分地提取数据特征并克服单一分解方法带有模态混叠等固有缺陷的问题,其预测效果更为显著[6]。罗宏远等[7]结合二层分解技术应用于 $PM_{2.5}$ 浓度预测,梁小珍等[8]将二层分解策略应用于航空客运需求预测。上述模型均证实了二层分解方法表现显著优于传统的单一分解策略。在此基础上,运用 EMD- WTS 模型对 AQI 非结构数据与历史数据进行二层分解,以充分刻画数据的细节波动,进而减小预测误差。

数据分解虽有助于提高预测精度,但由于分解后所得序列分别具有不同的特征,因此使用单一预测方法所得结果精度较低。研究表明,使用组合预测模型可以形成模型优势互补,能有效避免单一模型弊端[9]。具有多样性的组合预测模型能够充分利用分解所得维度所含信息,同时适用于具有不同特征的数据。因此,选取 Holt 指数平滑法、支持向量回归(SVR)以及人工神经网络(ANN)3 种预测方法对二层分解所得预测结果开展预测,该组合预测框架同时适用于具有线性或非线性特征的数据,同时包含了传统计量模型与人工智能模型,能够全面考虑分解所得序列的特征,从而进一步提高了预测精度。

综合已有研究,可以发现 AQI 预测仍存在下述问题:(1)已有 AQI 预测方法大多对数据进行单一分解,而 AQI 数据的高复杂性与非平稳性导致该方法无法全面提取数据特征,且单一分解方法带有模态混叠等固有缺陷;(2)少部分研究对 AQI 历史数据进行二次分解,但基于历史数据的预测所得结果具有滞后性,现有 AQI 预测研究对于融入以及如何融入非结构数据以弥补历史数据不足仍缺乏探索;(3)单一预测方法难以同时捕捉分解所得序列的不同特征,此时运用包含多个不同特征预测方法的组合预测模型将显著提升预测效果。

因此,针对已有研究存在的问题,提出一种结合非结构数据的 EMD-WTS 的二层分解 AQI 组合预测方法。首先,基于百度指数“需求图谱”功能等筛选 AQI 相关百度指数关键词,并使用局部线性嵌入(LLE)进行降维。其次,对降维结果与 AQI 历史数据进行 EMD 分解,重构后得到降维结果与 AQI 历史数据的原始高频序列、低频序列与趋势序列。接着,对所得高频序列均进行 WT 分解,重构后得到原高频序列的高、低频与趋势项。然后,对上述所得序列分别运用 Holt、SVR、ANN 进行组合预测并将所得结果输入 BP 神经网络进行集成,集成所得结果相加后得到原高频序列预测值。同时,运用相同的组合预测方法对原始低频与趋势项分别开展预测,得到各自预测结果。最后,将原始高、低频与趋势项预测结果相加,得到融合非结构数据的 AQI 二层分解组合预测结果。为验证上述模型的预测精度,开展仿真及对比实验,证实了本模型预测精度更高、预测效果更为显著。

1　组合预测模型理论与框架

考虑到 AQI 非结构数据与历史数据的高复杂性与非平稳性等特征,提出一种融合非结构数据的二层分解组合预测新框架,具体内容如下。

1.1　非结构数据

1.1.1　非结构数据获取

AQI 网络搜索数据反应了居民及政府对于空气质量的关注度。AQI 相关关键词的网络搜索热度能够反映社会公众对空气质量关注程度的改变,可以对空气质量的变化起到一定的预见与解释作用。相对于谷歌搜索引擎,在我国百度搜索引擎占据更高市场份额,是体现我国居民关注度的重要数据来源。因此,选取百度指数作为非结构数据源,通过百度指数关键词体现居民对于 AQI 的关注度。

综合专家推荐与百度“需求图谱”功能,筛选AQI相关关键词。百度指数“需求图谱”功能展示了关键词与各时期内相关检索词之间的关联强度,能够充分体现出网民的需求。最终,确定了空气质量、$PM_{2.5}$、CO、北京空气质量、大气污染、雾霾等30个最能反映居民对于AQI专注度的百度指数关键词,提取2019-01-01至2020-01-31的指数数据。

1.1.2 LLE降维

由于获取的百度指数维度高且存在信息冗余,直接用于预测将导致模型高度复杂、预测效率低等问题。因此,对非结构数据进行降维十分必要。选用局部线性嵌入(LLE)算法对之降维。LLE原理在于使得降维前后近邻之间的局部线性结构不变,具体步骤如下[10]。

Step 1 根据数据集 $\boldsymbol{X}=[x_1,x_2,\cdots x_n]$ 各点之间的欧氏距离寻找每个样本点 x_i 的 k 个最近邻$\{x_j, j\in J_i\}$,J_i 表示样本点 x_i 的 k 个最近邻点下标集合。

Step 2 计算各点与对应邻域点之间的重构权值 w_{ij}(非邻域点取权重为0),通过最小化重构误差计算权重矩阵 $\boldsymbol{W}$,如式(1)所示:

$$e(\boldsymbol{W})=\sum_{i=1}^{n}\left\|x_i-\sum_{j=1}^{n}w_{ij}x_j\right\|^2 \tag{1}$$

Step 3 最小化降维带来的损失函数,如式(2)所示:

$$\min\ \varphi(Y)=\sum_{i=1}^{n}\left\|y_i-\sum_{j=1}^{n}w_{ij}y_j\right\|^2$$
$$\text{s.t.}\begin{cases}\sum_i y_i=1\\ \dfrac{1}{n}\sum_i y_i y_i^{\mathrm{T}}=I\end{cases} \tag{2}$$

其中,$\boldsymbol{I}$ 为单位矩阵。另取 $\boldsymbol{M}=(I-\boldsymbol{W})^{\mathrm{T}}(I-\boldsymbol{W})$,求得低维嵌入 $\boldsymbol{Y}$ 取 $\boldsymbol{M}$ 的最小 $d+1$ 个特征值对应的特征向量 $v_2,v_3,\cdots,v_{d+1}$,即 $\boldsymbol{Y}=[v_2,v_3,\cdots,v_{d+1}]^{\mathrm{T}}$。

1.2 二次分解与重构

1.2.1 二次分解

由于AQI数据具有高复杂度于非平稳性特征,采用EMD-WTS二层分解模型对之进行数据分解,以全面获取数据特征。

(1) 经验模态分解(EMD)。EMD分解能较好地分解非平稳、非线性时序,有较好的时间与频率分辨率[11]。分解后得到多个频率由高至低排列的IMF分量及一个残差项,如式(3)所示:

$$x(t)=\sum_{k=1}^{N}IMF_k(t)+R_N(t) \tag{3}$$

其中,$x(t)$ 代表原时间序列,$IMF_k(t)$ 代表第 k 个本征模函数,$R_N(t)$ 代表残差项。

(2) 小波分解(WT)。小波分解所得各序列具有单一频率,因而具有更好的稳定性,其定义如式(4)所示[12]:

$$DWT_x(m,n)=\frac{1}{\sqrt{2^m}}\sum_k x_k\psi^*\frac{k-n}{2^m} \tag{4}$$

其中,m 是比例因子,$n=1,2,\cdots,N$ 是采样时间,N 是样本数。

1.2.2 数据重构

根据数据特征对分解结果进行重构将有效降低模型复杂度,提高模型预测效率。首先,按频率由高至低排列分解结果并计算其均值,接着利用t检验确定均值显著偏离0的第 m 个序列,最后叠加第1个至 $m-1$ 个序列获取原始序列的高频项,叠加第 m 至最后一个序列得到低频项[13]。

1.3 组合预测

组合预测由于其方法的多样性能降低预测误差。选取线性与非线性模型、传统计量与人工智能模型搭建组合预测框架,具体包括Holt指数平滑模型、SVR以及ANN。

1.3.1 Holt

Holt指数平滑法适用于含趋势成分的时间序列预测。Holt模型一般形式如式(5)所示。

$$\begin{aligned}S_t&=\alpha x_t+(1-\alpha)(S_{t-1}+T_{t-1})\\ T_t&=\gamma(S_t-S_{t-1})+(1-\gamma)T_{t-1}\\ F_{t+k}&=S_t+kT_t\end{aligned} \tag{5}$$

其中:α,β 为平滑系数,S_t 为第 t 期的指数平滑值,T_t 为第 t 期趋势值,F 为预测值。

1.3.2 SVR

SVR为支持向量机应用之一,通过在高维空间中构造线性决策函数来实现线性回归,并用核函数

代替线性方程中的线性项。SVR 问题公式描述如式(6)所示[14]:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} |f(x_i - y_i)|$$
$$\text{s.t. } |f(x_i) - y_i| \leqslant \varepsilon \tag{6}$$

其中,$\boldsymbol{w}$ 为常向量,$\boldsymbol{C}$ 为常量,(x_i, y_i) 为给定数据训练集,$f(x_i)$ 为映射函数。

1.3.3　人工神经网络

人工神经网络(ANN)算法启发于生物神经系统。类比于人脑学习,ANN 将获得的"知识"存储于神经元之间的权重。ANN 由输入层、隐含层以及输出层搭建而成,搭建时需要确定层数、每层的神经元数量、各层与拓扑网络之间的连接类型[15]。ANN 由于其大规模并行、自组织、自学习等优点于包括预测在内的众多领域里得到了广泛的应用。

1.4　误差评价指标

为检验模型预测精度,通过对比各模型的平均绝对误差(MAE)、误差平方和(SSE)、均方误差(MSE)与平均绝对百分比误差(MAPE)数值来证实本预测模型结果的有效性。各指标计算公式如下所示:

$$P_{\text{MAE}} = \frac{1}{n}\sum_{t=1}^{n} |x(t) - \hat{x}(t)|$$

$$P_{\text{SSE}} = \sum_{t=1}^{n} (x(t) - \hat{x}(t))^2$$

$$P_{\text{MSE}} = \frac{1}{n}\sqrt{\sum_{t=1}^{n} (x(t) - \hat{x}(t))^2}$$

$$P_{\text{MAPE}} = \frac{1}{n}\sum_{t=1}^{n} \left|\frac{x(t) - \hat{x}(t)}{x(t)}\right|$$

1.5　预测框架

根据上述文献梳理和理论基础,提出基于非结构数据和 EMD-WTS 二层分解的 AQI 组合预测框架(图 1),具体步骤如下。

Step 1　根据专家推荐与百度"需求图谱"功能确定百度指数关键词作为非结构数据,并利用 LLE 方法对之降至 2 维得到序列 L_1 与 L_2。

Step 2　对 AQI 历史数据以及 L_1、L_2 分别进行 EMD 分解,而后运用均值检验进行重构得到各自的高频、低频与趋势序列。

Step 3　对各高频序列进行 WT 分解,重构后得其高、低频与趋势项。

Step 4　将所得序列各自输入组合预测模型(其中 ANN 模型的输入为历史数据分解结果以及与之对应的非结构数据数列,Holt 与 SVR 仅以历史数据分解结果为输入),运用 BP 神经网络各自集成得到原高频序列的高、低频与趋势项预测结果 $\hat{y}_1^1$、$\hat{y}_1^2$、$\hat{y}_1^3$。上述结果相加得到了 AQI 高频序列预测结果 y_1。

Step 5　将 AQI 的低频与趋势序列分别输入组合预测模型,运用 BP 神经网络拟合后得到各自预测结果 $\hat{y}_2$、$\hat{y}_3$。而后,$\hat{y}_1$、$\hat{y}_2$ 与 $\hat{y}_3$ 相加得到 AQI 预测结果 $\hat{y}$。

对比现有的 AQI 预测模型,上述预测框架存在如下优势:对 AQI 数据进行二层分解,更加全面地提取数据信息;运用非结构数据弥补历史数据滞后缺点,并运用 LLE 降维以降低模型复杂度,实现了非结构数据的有效利用;对特征各异的数据分解结果进行组合预测。模型既包含线性模型,又包含非线性模型,既包含传统计量模型,又包含人工智能模型,能够同时有效预测具有不同特征的数据。

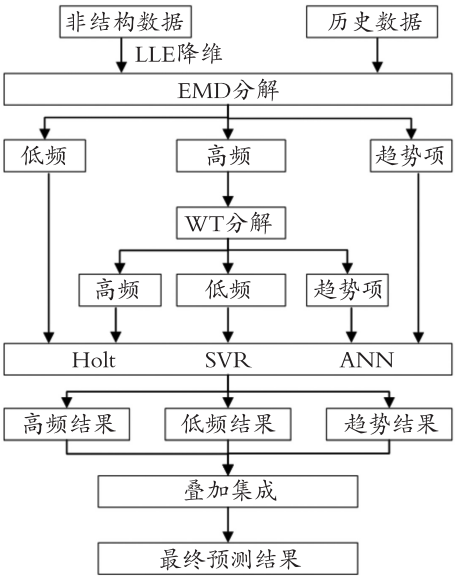

图 1　本模型预测框架

Fig. 1　The forecast framework of this model

2 仿真实验

2.1 数据来源于数据处理

在仿真实验中,以北京市为例,选取 AQI 在 2019-01-01 至 2019-12-31 期间的数据为训练集,在 2020-01-01 至 2020-01-31 期间的数据为测试集。同时,选取北京地域相应时间段的 30 个百度指数关键词的指数数据,并运用 LLE 对之降维,取 $k=6$、$d=2$,得到两个降维后的序列 L_1、L_2。

2.2 EMD-WTS 二层分解与重构

首先,运用 EMD 模型对 2019-01-01 至 2019-12-31 的 AQI 历史数据与非结构数据降维结果进行分解。其中,EMD 的趋势序列用 res. 表示,剩余序列用 IMF 表示。历史数据 EMD 分解结果如图 2 所示。

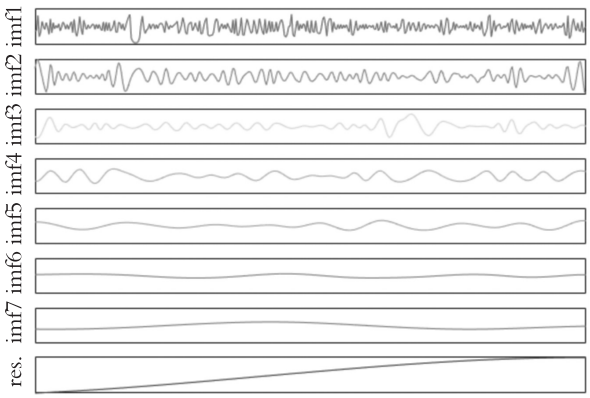

图 2 历史数据 EMD 分解结果

Fig. 2 Decomposition results of historical data based on EMD

接着,对趋势序列以外的各序列按频率由高至低排列,并根据 t 检验所得结果分组累加,得到历史数据高频、低频与趋势序列 y^{11}、y^{12}、y^{13},降维序列的高频、低频与趋势序列 L_k^{11}、L_k^{12}、L_k^{13}($k=1,2$)。重构结果表明,原始高频序列波动性较大,确实需要进一步分解以提取其数据特征。

然后,对原始序列的高频项进行 WT 分解,得到历史数据高频 y^{11} 的高、低频与趋势项 y^{21}、y^{21}、y^{21},降维序列高频 L_k^{11} 的高、低频与趋势项 L_k^{21}、L_k^{22}、L_k^{23}($k=1,2$)。其中,历史数据高频 y^{11} 的分解结果如图 3 所示,图 3 中 d 为 y^{11} 的细节波动,res. 为趋势项。

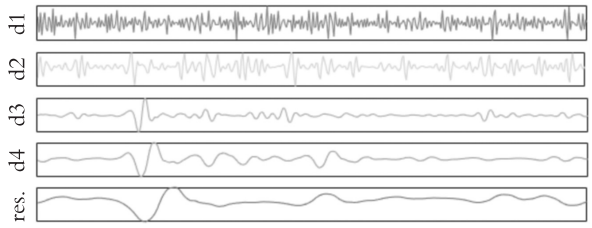

图 3 历史数据高频 WT 分解结果

Fig. 3 Decomposition results of high-frequency data based on WT

2.3 AQI 组合预测

对历史数据二层分解结果 y^{21}、y^{22}、y^{23} 分别运用组合预测模型开展预测,对各自组合预测结果集成得到$\hat{y}^{21}$、$\hat{y}^{22}$、$\hat{y}^{23}$,叠加后得到高频序列 y^{11} 的预测值 $\hat{y}^{11}$。同理,对历史数据低频 y^{12} 与趋势序列 y^{13} 进行组合预测,得到$\hat{y}^{12}$、$\hat{y}^{13}$,叠加得到历史数据 y 最终预测结果$\hat{y}$。具体实验步骤如下。

以高频 y^{21} 的预测过程为例具体展示二层分解的组合预测环节。取 $\alpha=0.4$、$\beta=0.4$ 运用 Holt 对 y^{21} 进行预测,取 $\varepsilon=0.001$ 运用 SVR 对 y^{21} 预测,取输入层、隐含层与输出层分别为 3、1 与 1 搭建 ANN 模型对 y^{21} 与 L_1^{21}、L_2^{21} 进行多输入的预测。最后,取输入层、隐含层与输出层分别为 3、1 与 5 搭建 BP 神经网络以集成上述预测结果,得到$\hat{y}^{21}$。在获取$\hat{y}^{22}$、$\hat{y}^{23}$ 时,为了验证模型的适用性,将组合预测方法中的 SVR 预测方法替换 BP 神经网络预测方法。结果表明本文的组合预测模型具有灵活的适用性。最后叠加$\hat{y}^{21}$、$\hat{y}^{22}$与$\hat{y}^{23}$得到高频预测结果$\hat{y}^{11}$。接着对 y^{12}、y^{13} 进行预测,预测方法同 y^{21} 预测,得到$\hat{y}^{12}$、$\hat{y}^{13}$。叠加$\hat{y}^{11}$、$\hat{y}^{12}$与$\hat{y}^{13}$得到 AQI 最终预测结果$\hat{y}$(图 4)。由图 4 可知,结合非结构数据的 AQI 二层分解组合预测模型不仅能有效地预测 AQI 的升降趋势,而且对 AQI 细节的波动拟合效果好。

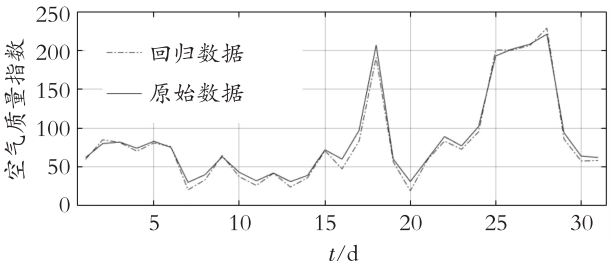

图 4 历史数据预测结果图

Fig. 4 Final forecast results of historical data

2.4　各预测模型预测结果对比

为体现提出的结合非结构数据的二层分解组合预测模型的有效性,实验将对比本模型与其余 5 个模型。其中,模型 1 为二层分解-组合预测模型,对比本模型(即模型 6)未使用非结构数据;模型 2 为非结构数据-EMD-组合预测,对比本模型未进行二层分解;模型 3 与模型 4 分别为非结构数据-二层分解-ANN 模型与非结构数据-二层分解-SVR 模型,对比本模型对分解结果仅单项预测;模型 5 为 EMD-ARIMA,是现有研究预测方法[16]。通过记录各模型的 MAE、SSE、MSE 与 MAPE 指标来体现其预测精度,所得结果如表 1 所示。

表 1　各模型预测精度评价指标对比

Table 1　The comparison of the evaluation index of prediction accuracy of each model

预测模型	MAE	SSE	MSE	MAPE
二层分解-组合预测	2.551 9	23 548	0.387 5	0.044 2
非结构数据-EMD-组合预测	4.597 9	23 257	0.385 1	0.073 2
非结构数据-二层分解-ANN	3.815 0	17 166	0.330 9	0.058 6
非结构数据-二层分解-SVR	44.315 6	1.33×10^{6}	2.920 5	0.597 6
EMD-ARIMA[16]	21.580 4	$3.305\ 1\times10^{5}$	1.451 8	0.301 0
本模型	1.525 4	3 568.6	0.150 9	0.021 8

首先,对比各误差指标数值,发现本模型预测精度显著高于其他模型,体现了本文预测框架的实用性。其次,对各模型做详细对比:模型 1 相比本模型各误差指标值均较高,证实了非结构数据对于预测起显著信息补充作用,融入非结构可以提高预测精度;模型 2 对比本模型预测误差更大,说明二层分解能更为充分地刻画原始数据细节波动;模型 3、模型 4 与本模型的误差数据对比突出了组合预测的重要性,由多种分解方法搭建的组合预测模型确实能吸收各模型优点,进而提高预测精度;模型 5 与本文模型的对比进一步体现了非结构数据、二层分解以及组合预测方法的显著效果,与现有研究方法的对比广泛地证实本模型的有效性。综上,对比实验综合地体现了本模型的有效性,证实了结合非结构数据的二层分解组合预测模型预测效果更为显著。

3　结束语

空气质量指数的精确预测对于维护居民健康、制定合理的环保政策以及改善生态环境具有重要意义。提出一种结合非结构数据的 AQI 二层分解组合预测模型,用非结构数据弥补历史数据预测的滞后性,并通过二层分解更为全面地提取数据包含的信息,最后利用组合预测克服单项预测无法适应不同特征分解结果的缺点,提高了模型的适用性。首先,基于百度指数提取非机构数据并运用 LLE 对之降维。其次,利用 EMD 分解 AQI 历史数据与降维结果,重构后得到 AQI 数据的高、低频与趋势项。接着,对各原始高频序列进行 WT 分解,得其高、低频与趋势序列。然后,对所得序列运用组合预测框架进行预测,并输入 BP 神经网络进行集成,得到各序列组合预测结果。累加上述结果输出原高频序列预测值。相同地,对原始低频与趋势序列进行组合预测与 BP 集成,得到各自预测结果。最后,叠加原高、低频与趋势项得到 AQI 最终预测结果。实验表明,非结构数据的加入对预测框架起显著信息补充作用,有效地提高了预测精度;二层分解相较单一分解更为全面地刻画了数据的细节波动,充分提取了原始数据特征,有助于减小预测误差;包含线性与非线性模型、传统计量与人工智能模型的组合预测体系充分结合各单项预测模型的优点,能同时适用于具有不同特征的分解所得序列,提高了模型预测效果。

参考文献(References):

[1] WU Q L, LIN H X. A Novel Optimal-hybrid Model for Daily Air Quality Index Prediction Considering Air Pollutant Factors[J]. Science of the Total Environment, 2019(683):808—821

[2] 张玲玲,张笑,崔怡雯. 基于聚类方法的百度搜索指数关键词优化及客流量预测研究[J]. 管理评论,2018,30(8):126—137
ZHANG L L, ZHANG X, CUI Y W. Forecasting Tourist Volume Based on Clustering Method with Screening Keywords of Search Engine Data [J]. Management Review, 2018, 30(8):126—137(in Chinese)

[3] D' AMURIF, MARCUCCIA J. The Predictive Power of Google Searches in Forecasting US Unemployment[J]. International Journal of Forecasting, 2017, 33(4):801—816

[4] 陈声利,关涛,李一军. 基于跳跃、好坏波动率与百度指数的股指期货波动率预测[J]. 系统工程理论与实践,2018,38(2):299—316
CHEN S L, GUAN T, LI Y J. Forecasting Realized Volatility of Chinese Stock Index Futures Based on Jumps, Good-bad Volatility and Baidu Index[J]. Systems Engineering- Theory& Practice, 2018, 38(2):299—316(in Chinese)

[5] LI H M, WANG J Z, LI R R, et al. Novel Analysis-forecast System Based on Multi-objective Optimization for Air Quality Index[J]. Journal of Cleaner Production, 2019(208):1365—1383

[6] NA S N, ZHOU J Z, LU C N, et al. An Adaptive Dynamic Short-term Wind Speed Forecasting Model Using Secondary Decomposition and an Improved Regularized Extreme Learning Machine[J]. Energy, 2018(165):939—957

[7] 罗宏远,王德运,刘艳玲,等. 基于二层分解技术和改进极限学习机模型的$PM_{2.5}$浓度预测研究[J]. 系统工程理论与实践,2018,38(5):1321—1330
LUO H Y, WANG D Y, LIU Y L, et al. $PM_{2.5}$ Concentration Forecasting Based on Two-layer Decomposition Technique and Improved Extreme Learning Machine[J]. Systems Engineering-Theory& Practice, 2018, 38(5):1321—1330(in Chinese)

[8] 梁小珍,邬志坤,杨明歌,等. 基于二层分解策略和模糊时间序列模型的航空客运需求预测研究[J]. 中国管理科学,2020,28(2):1—11
LIANG X Z, WU Z K, YANG M G, et al. Air Passenger Demand Forecasting Based on a Dual Decomposition Strategy and Fuzzy Time Series Model [J]. Chinese Journal of Management Science, 2020, 28(2):1—11(in Chinese)

[9] ZHU J M, LIU J P, WU P, et al. A Novel Decomposition-ensemble Approach to Crude Oil Price Forecasting with Evolution Clustering and Combined Model[J]. International Journal of Machine Learning and Cybernetics, 2019(10):3349—3362

[10] 贾晶晶,顾明亮,朱恂,等. 基于流形学习与特征融合的汉语方言辨识[J]. 计算机工程与应用,2015,51(7):233—237
JIA J J, GU M L, ZHU X, et al. Chinese Dialect Identification Based on Manifold Learning and Feature Fusion [J]. Computer Engineering and Applications, 2015, 51(7):233—237(in Chinese)

[11] 王书平,胡爱梅,吴振信. 基于多尺度组合模型的铜价预测研究[J]. 中国管理科学,2014,22(8):21—28
WANG S P, HU A M, WU Z X. Forecasting of Copper Price Based on Multi-scale Combined Model[J]. Chinese Journal of Management Science, 2014, 22(8):21—28(in Chinese)

[12] SUN W, ZHANG C C, SUN C P. Carbon Pricing Prediction Based on Wavelet Transform and K-ELM Optimized by Bat Optimization Algorithm in China ETS: The Case of Shanghai and Hubei Carbon Markets[J]. Carbon Management, 2018, 9(6):605—617

[13] 杨云飞,鲍玉昆,胡忠义,等. 基于EMD和SVMs的原油价格预测方法[J]. 管理学报,2010,7(12):1884—1889
YANG Y F, BAO Y K, HU Z Y, et al. Crude Oil Price Prediction Based on Empirical Mode Decomposition and Support Vector Machines [J]. Chinese Journal of Management, 2010, 7(12):1884—1889(in Chinese)

[14] 甘中学,喻想想,许裕栗,等. 基于周期性ARMA-SVR模型的空调冷热负荷预测[J]. 控制工程,2020,27(2):380—385
GAN Z X, YU X X, XU Y L, et al. Air-conditioning

Cooling and Heating Load Prediction Based on Periodic ARMA-SVR Model[J]. Control Engineering of China, 2020,27(2):380—385(in Chinese)

[15] PKDM F, CAG S, GBL S D. Analysis of The Use of Discrete Wavelet Transforms Coupled with ANN for Short-term Streamflow Forecasting[J]. Applied Soft Computing, 2019(80):494—505

[16] 李勃旭,南西康,郑向东,等. 基于 EMD-ARIMA 模型的地铁门传动系统早期故障预测[J]. 计算机系统应用, 2019,28(9):110—117
LI B X, NAN X K, ZHENG X D, et al. Early Fault Prediction of Metro Door Transmission System Based on EMD-ARIMA Model[J]. Computer Systems & Applications,2019,28(9):110—117(in Chinese)

AQI Combined Forecast Method Based on Unstructured Data and EMD-WTS Two-layer Decomposition

LIU Jin-pei[1,2], ZHANG Liao-dan[1], DING Rong[1], WANG Piao[1], LUO Rui[1]**

(1. School of Business, Anhui University, Hefei 230601, China
2. Edward P. Fitts Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, NC, 27695, USA)

Abstract: To deal with the highly random and unstable sequence of Air Quality Index (AQI), an EMD-WTS two-layer decomposition and unstructured data based combined forecast model is proposed. Firstly, the Baidu index keywords are filtered and the corresponding data is extracted, after which the locally linear embedding (LLE) is applied to reduce the dimensions. Secondly, the empirical modal decomposition (EMD) and reconstruction are carried out on AQI historical sequence and dimension-lowering results. Then, the wavelet transform (WT) is adopted to decompose and reconstruct the gained high-frequency sequence. After reconstruction, the Holt exponential smoothing, support vector regression (SVR) and artificial neural network (ANN) are used to forecast the results of two-layer decomposition and the original low frequency and trend sequence. Subsequently, the forecast results are integrated by BP neural network. Eventually, the gained forecast results above are added up and the final prediction results of AQI are obtained. The comparative experiment's results demonstrate that the forecast model aforesaid can make full use of a variety of data information, and the prediction accuracy is quite high.

Key words: combined forecast; air quality index; EMD-WTS two-layer decomposition; unstructured data

责任编辑:罗姗姗

引用本文/Cite this paper:
刘金培,张了丹,丁蓉,等. 基于非结构数据和 EMD-WTS 二层分解的 AQI 组合预测方法[J]. 重庆工商大学学报(自然科学版),2021,38(2):56—63
LIU J P, ZHANG L D, DING R, et al. AQI Combined Forecast Method Based on Unstructured Data and EMD-WTS Two-layer Decomposition[J]. Journal of Chongqing Technology and Business University (Natural Science Edition),2021,38(2):56—63