针对 PM2.5 单时间序列数据的动态调整预测模型

张熙来 [1] 赵俭辉 [1] 蔡 波 [1]

摘 要 针对细颗粒物 PM2.5 的浓度预测, 本文提出了基于单时间序列数据的动态调整模型. 在动态指数平滑算法中, 指数平滑次数与参数基于样本数据并借助二分查找进行调整. 在动态马尔科夫模型中, 马尔科夫链的残差状态数、隐马尔科夫模型的隐状态数、连续样本数和阈值参数都通过训练数据加以调整. 动态调整模型将指数平滑法和马尔科夫模型有效结合起来, 指数平滑法得到的预测值由马尔科夫模型进行校正, 从而提高预测准确度. 基于大量实际 PM2.5 数据进行测试, 验证了算法的有效性. 并与其他现有的灰色模型、人工神经网络、自回归滑动平均模型、支持向量机等方法进行了对比, 表明所提模型能够得到精度更高的预测结果. 本文模型不局限于 PM2.5 数据, 还可应用于其他类型的数据预测.

关键词 空气质量指数, 指数平滑法, 马尔科夫模型, 动态调整

引用格式 张熙来, 赵俭辉, 蔡波. 针对 PM2.5 单时间序列数据的动态调整预测模型. 自动化学报, 2018, **44**(10): 1790−1798

DOI 10.16383/j.aas.2017.c170026

Prediction Model With Dynamic Adjustment for Single Time Series of PM2.5

ZHANG Xi-Lai[1] ZHAO Jian-Hui[1] CAI Bo[1]

Abstract A prediction model is proposed with dynamic adjustment for single time series of PM2.5 data. In the dynamic exponential smoothing algorithm, the optimal exponent and parameter are determined by sample data and binary search. In the dynamic Markov model, the state number of residual errors from Markov chain, numbers of hidden and observable states, and threshold parameters from hidden Markov model, are all decided dynamically based on training data. The proposed dynamic model combines the two models effectively, and predictions from exponential smoothing are adjusted by Markov model to increase the accuracy. Using a large number of real PM2.5 data, efficiency of the proposed model has been tested. Compared with the existing popular methods, such as gray model, artificial neural networks, auto-regressive moving average, support vector machine, the proposed model can obtain prediction results with the best precision. In addition to PM2.5, the dynamically adjusted prediction model may be used for prediction of other type single time series of data.

Key words Air quality index, exponential smoothing, Markov model, dynamic adjustment

Citation Zhang Xi-Lai, Zhao Jian-Hui, Cai Bo. Prediction model with dynamic adjustment for single time series of PM2.5. *Acta Automatica Sinica*, 2018, **44**(10): 1790−1798

在空气污染中, 污染物颗粒是我国主要监测指标, 如 PM2.5、PM10, 其中 PM2.5 是多环芳烃、重金属等有毒物质的载体, 对人体危害极大[1−2]. 针对它们的监测与预报, 已成为污染防治的大事. 迄今为止, 已有学者对空气污染预测问题展开了研究. Rajasegarar 等[3] 利用多种传感器并设计多种模型来预测特定污染物尤其是 PM10 的浓度. Shaban 等[4] 对空气质量检测及预测进行了研究, 利用支持向量机 (Support vector model, SVM)、M5P 模型树和人工神经网络 (Artificial neural networks, ANN) 来预测城市的 SO_2、O_3、NO_2 浓度, 用于训练的独立变量包括 H_2S、NO_2、O_2、SO_2、温度、湿度、风速等多种因素. Díaz-Robles 等[5] 利用自回归滑动平均模型 (Auto-regressive moving average, ARMA) 和人工神经网络组成的混合模型预测城市中 PM10 的浓度, 用于训练的独立变量包括 PM10、温度、风速、降水量、相对湿度、太阳辐射、压力等. Stadlober 等[6] 以气象和人为参数为多元线性回归日预测模型的输入变量, 例如 PM10、温度、风速、降水量、日期等, 进行 PM10 浓度预测. 事实上, 预测模型的设计应适合数据自身的特点, 截止目前尚难以找到针对 PM2.5 浓度预测的方法. 同时, 与污染颗粒预测相关的重要因素来源很多, 例如车辆废气、工业活动、煤燃烧等, 而收集或测量这些污染数

收稿日期 2017-01-18 录用日期 2017-05-11
Manuscript received January 18, 2017; accepted May 11, 2017
湖北省科技支撑计划 (2014BAA149), 中国空间技术研究院创新基金 (CAST2014), 中央高校基本科研业务费专项资金 (2042016GF0023) 资助
Supported by Hubei Support Plan for Science and Technology (2014BAA149), China Academy of Space Technology (CAST2014), and Fundamental Research Funds for the Central Universities (2042016GF0023)
本文责任编委 郭戈
Recommended by Associate Editor GUO Ge
1. 软件工程国家重点实验室, 武汉大学计算机学院 武汉 430072
1. State Key Laboratory of Software Engineering, School of Computer, Wuhan University, Wuhan 430072

据本身就非常困难. 另外, 某些地区或时段内相关监测数据有时会出现不完整的情况. 因此, 基于先前测量的单时间序列数据, 提出针对 PM2.5 的浓度预测模型, 是一种非常实用的方式.

针对单时间序列数据的预测已有多种算法, 例如指数平滑法 (Exponential smoothing, ES)、马尔科夫模型 (Markov model, MM)、灰色预测模型 (Gray model, GM)、自回归滑动平均模型、支持向量机、人工神经网络等. 为了预测巴西电能消费总额和消费类别, Macaira 等[7] 采用 Pegels 指数平滑法进行动态预测和优化. 为了制定更好的电力调度决策, Taylor 等[8] 对电力需求的变化进行了预测, 针对单序列数据的指数平滑法表现出了最好的效果. 为了预测 EEG 信号同步的状态转换, Jamal 等[9] 提出了基于马尔科夫链 (Markov chain, MC) 的概率模型. 为了构建道路交通模型, Lawlor 等[10] 提出了一个基于离散时间马尔科夫链的混合模型, 用于交通路线的预测并取得很好的效果. 为了改善机器人的人性化内部模式, Razin 等[11] 利用分层隐马尔科夫模型 (LHMM) 对人体手臂肌肉肌电图生理数据变化进行了预测, 预测效果明显优于朴素贝叶斯和支持向量机. Soualhi 等[12] 针对降解工业的故障发生进行预测, 借助隐马尔科夫模型 (Hidden Markov model, HMM) 成功提高了工业安全性与可靠性. 为了完善无功功率补偿装置性能, Samet 等[13] 针对电弧炉的无功功率变化进行预测, 采用滚动灰色模型结合马尔科夫模型的方法, 并与自回归滑动平均模型进行了比较. Chen 等[14] 通过灰色模型 GM(1, 1) 进行了锂离子电池的充放电性能的预测研究. Lima 等[15] 利用人工神经网络, 对赤道附近圣路易斯的电离层闪烁现象进行了预测, 并取得了很好的效果. 为了研究非浮式波浪能转换器的起伏位移, Nagulan 等[16] 利用人工神经网络开发了位置预测模型. 在核能研究中, Moshkbar-Bakhshayesh 等[17] 利用自回归滑动平均模型和 BP 神经网络相结合的模型对时间序列变量进行预测与诊断. Wei 等[18] 利用时间序列技术建模和分析流量数据, 设计了一个基于 ARMA 的数据流量预测模型. 吴奇等[19] 设计了一种新的鲁棒小波 v-SVM, 针对汽车销售时序进行预测, 得到了有效的预测结果. Liu 等[20] 采用正交测试结合支持向量机的混合模型对风电爬坡事件进行预测, 有效减小了爬坡事件对风力发电的影响. 在智能电网领域, Gupta 等[21] 基于概率框架, 用历史数据训练 SVM 模型, 进而预测停电等事件.

在上述方法中, 指数平滑是十分有效的单时间序列预测算法. 但指数平滑不能分析预测值与实际值间的误差并加以调整, 因此降低了预测的准确性. 马尔科夫模型也是常用的单时间序列预测算法, 通过分析预测误差的变化规律来提高预测精度. 因此, 针对单时间序列 PM2.5 预测, 本文提出了一个指数平滑法和马尔科夫模型组合的动态调整新算法, 即通过指数平滑法预测 PM2.5 的浓度, 然后用马尔科夫模型修正预测误差. 且相关参数都基于预测精度加以动态调整, 以得到精度更高的预测结果.

1　动态调整的预测算法

为了说明预测算法相关参数的动态调整, 从湖北省武汉市沌口新区监测站点 2015 年 3 月份以小时为间隔的 PM2.5 浓度实时监测数据 (单位: $\mu g/m^3$) 中随机选取以下 5 组 PM2.5 序列数据进行分析, 每组单时间序列含 N 个测量值.

$$
\begin{aligned}
X_1 &= \{28, 26, 22, 29, 15, 17, 21, 26, \cdots\}\\
X_2 &= \{90, 77, 58, 39, 35, 53, 63, 70, \cdots\}\\
X_3 &= \{33, 49, 68, 74, 84, 86, 87, 101, \cdots\}\\
X_4 &= \{60, 67, 68, 99, 106, 109, 102, 97, \cdots\}\\
X_5 &= \{70, 84, 97, 135, 145, 152, 86, 125, \cdots\}
\end{aligned}
$$

以上 5 组数据只用于说明算法参数动态调整的过程, 不包含在本文实验结果的评估数据中. 算法预测方式为: 以每 N_P 个连续数据为输入来计算第 $N_P + 1$ 个 PM2.5 的预测值, 以其预测值和实际值之差作为第 N_P+1 个样本的误差. 同理, 第 N_P+2 到第 N 个样本也如上计算, 每组得到 $N-N_P$ 个误差, 以这些误差统计出的均方根误差 (Root mean square error, RMSE) 作为每组 PM2.5 序列的误差. 本文实验中 $N = 118$, N_P 则因预测模型的需要而不同.

1.1　基于指数平滑法的动态预测

指数平滑法的预测结果由过去的实际值和其预测值加权平均得到, 指数平滑分为三种: 一次指数平滑、二次指数平滑、三次指数平滑. 一次指数平滑适用于趋势几乎没有变化的时间序列数据, 二次指数平滑适用于呈线性趋势的时间序列数据, 三次指数平滑适用于有明显变化且非线性的时间序列数据. 指数平滑的基本预测公式为

$$S_t = \alpha \times y_t + (1-\alpha) \times S_{t-1} \tag{1}$$

其中, S_t 是 t 时刻的指数平滑数值, S_{t-1} 是 $(t-1)$ 时刻的指数平滑值, y_t 是 t 时刻的样本实际测量值. α 是指数平滑参数, 取值范围在 0～1 之间, α 值越接近于 1, 当前时间数据所占比重越大.

以 X_1 序列为例分析参数 α 的预测准确性, 绘制不同 α 值对应的预测误差如图 1 所示. 其中, 参数 α 取值以 0.01 步长递增, 共 100 个值, 图中三

条数据线分别代表一次、二次、三次指数平滑所得 RMSE 的变化情况. 根据实验数据, 一次指数平滑最小 RMSE 为 4.4456, 对应 α 值为 0.90; 二次指数平滑最小 RMSE 为 4.4040, 对应 α 值为 0.69; 三次指数平滑最小 RMSE 为 4.0643, 对应 α 值为 0.73. 当然, 最小 RMSE 和对应的 α 只是 100 个采样值中表现最好的, 还不能直接作为最优解.

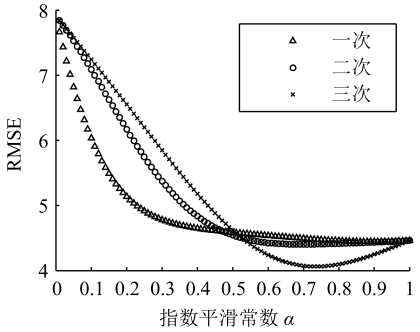


图 1 序列 X_1 指数平滑不同 α 值的预测误差

Fig. 1 ES prediction errors with different α values for sequence X_1

显然, 图 1 中每种指数平滑对应的误差分布都具有单峰特性. 因此, 我们基于二分查找的思路, 设计了一种迭代算法来搜索最优的 α 值. 迭代搜索步骤如下:

步骤 1. 从图 1 中找到某种指数平滑的最小 RMSE 与对应的 α_0;

步骤 2. 以 α_0 左邻 α 采样值为查找区间起点 α_s, 以右邻 α 采样值为查找区间终点 α_e;

步骤 3. 求区间中点 $\alpha_m = (alpha_s + alpha_e)/2$, 若 α_s 对应的 RMSE 大于 α_e 对应的 RMSE, 则以 α_m 作为新的区间起点取代 α_s, 反之, 以 α_m 作为新的区间终点取代 α_e;

步骤 4. 重复步骤 3, 直到 α_s 与 α_e 对应 RMSE 之差小于预定阈值如 0.0001;

步骤 5. 以迭代终止时的 α_s 或 $alpha_e$ 为最优 α, 同时求得对应的最小 RMSE.

根据上述二分查找算法, 我们得到了每种指数平滑法的最小 RMSE 值和相应的最优 α 值, 如表 1 所示.

表 1 二分查找得到的 X_1 序列最优 α 与 RMSE

Table 1 The optimal parameter α and related RMSE from binary search for sequence X_1

指数平滑法	最小 RMSE	最优 α
一次	4.4455	0.9050
二次	4.4040	0.6900
三次	4.0641	0.7350

同理, 对 5 组 PM2.5 序列数据进行了计算. 如表 2 所示, 每组序列数据对应着一个最小 RMSE 值及其相应的最优指数平滑次数. 其中, 三次指数平滑得到了 3 组数据的最小 RMSE, 说明三次指数平滑在大多数情况下更适合于 PM2.5 单时间序列的预测.

表 2 三种指数平滑法对 5 组序列数据的预测效果

Table 2 Performances of 3 ES methods for 5 sequences

序列	一次	二次	三次	最优
X_1	4.4455	4.4040	4.0641	三次
X_2	8.5289	9.4706	9.1253	一次
X_3	11.7953	11.2577	11.7502	二次
X_4	5.6960	4.7106	4.1102	三次
X_5	36.2899	36.2919	34.4010	三次

在 PM2.5 实际预测中, 基于上述方法动态调整最优参数 α 及对应的指数平滑次数. 计算过程需要用到 $N_{\rm ES}$ 个连续 PM2.5 数据为输入, 因为时间较早数据对当前时刻预测的影响不大, $N_{\rm ES}$ 应取较小值. 以序列数据 X_4 为例分析参数 $N_{\rm ES}$ 对预测效果的影响, 不同 $N_{\rm ES}$ 值与对应的指数平滑法预测误差的关系曲线如图 2 所示. 可知, 当 $N_{\rm ES}$ 取值为 6 时 RMSE 最小. 因此在本文实验中, 指数平滑的 $N_{\rm ES} = 6$, 即基于每 6 个连续数据计算第 7 个 PM2.5 预测值.

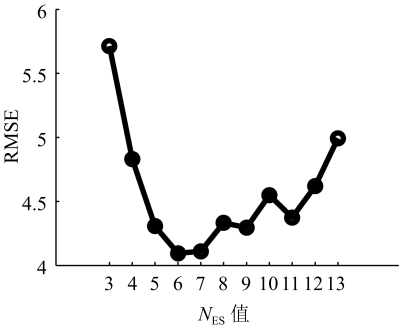


图 2 不同 $N_{\rm ES}$ 值的序列 X_4 指数平滑法预测误差

Fig. 2 ES prediction errors with different $N_{\rm ES}$ values for sequence X_4

1.2 基于马尔科夫模型的动态预测

马尔科夫模型分为两种: 马尔科夫链和隐马尔科夫模型. 马尔科夫链的数据状态是可观察的, 可直接预测. 隐马尔科夫模型的数据状态是不可观察的, 但这些数据的变化能反映在其他的可观察因素上, 因此将可观察因素用作预测介质, 通过介质来预

测隐状态的变化情况. 马尔科夫模型预测包括 3 个基本步骤: 状态划分、计算状态转移概率矩阵、得到预测值.

设 N_{MM} 个样本的时间序列实际值为

$$X=\{x(1),x(2),\cdots,x(N_{\mathrm{MM}})\} \tag{2}$$

设 N_{MM} 个样本的时间序列预测值为

$$\hat{X}=\{\hat{x}(1),\hat{x}(2),\cdots,\hat{x}(N_{\mathrm{MM}})\} \tag{3}$$

通过 $\varepsilon=\hat{x}(k)-x(k)$, 得到预测值与实际值的残差序列如下:

$$\xi=\{\varepsilon(1),\varepsilon(2),\cdots,\varepsilon(N_{\mathrm{MM}})\} \tag{4}$$

1.2.1　马尔科夫链预测的动态调整

基于马尔科夫链的预测算法中, 将残差序列分为 n_{MC} 个状态. 设第 i 个状态为

$$\otimes_i=[\widetilde{\otimes}_{1i},\widetilde{\otimes}_{2i}] \tag{5}$$

$$\widetilde{\otimes}_{1i}=\hat{x}(k)+A_i \tag{6}$$

$$\widetilde{\otimes}_{2i}=\hat{x}(k)+B_i \tag{7}$$

其中, A_i 是状态区间 i 的下限, 即 $A_i=\min(\varepsilon(k))$, B_i 是状态区间 i 的上限, 即 $B_i=\max(\varepsilon(k))$.

状态转移概率矩阵的计算公式为

$$P_i=P(\varepsilon(N_{\mathrm{MM}}+1)\in\otimes_j\mid\varepsilon(N_{\mathrm{MM}})\in\otimes_i) \tag{8}$$

$$P_{ij}=\frac{M_{ij}(m)}{M_i},\quad i=1,2,\cdots,n \tag{9}$$

其中, n 为状态的数量, 在马尔科夫链中 n 是 n_{MC}, 在隐马尔科夫模型中则是观察状态的数量. M_i 是序列 ξ 中状态为 $\otimes_i$ 的样本数量, $M_{ij}(m)$ 是从状态 $\otimes_i$ 经 m 步转换为状态 $\otimes_j$ 的样本数量. 在本文实验中, 选择了 $m=1$ 进行预测, 因为 $P(1)$ 是最常用情况, 具有高准确率和低计算量的特点.

根据序列 ξ 中的 N_{MM} 个样本和计算出的状态转移概率矩阵, 可以得到第 $N_{\mathrm{MM}}+1$ 个样本的预测结果, 即残差值, 然后再用这个预测残差来矫正第 $N_{\mathrm{MM}}+1$ 个预测值. 对于马尔科夫链, 预测结果的校正方式为

$$\hat{y}(N_{\mathrm{MM}}+1)=\hat{x}(N_{\mathrm{MM}}+1)+\frac{A_{N_{\mathrm{MM}}+1}+B_{N_{\mathrm{MM}}+1}}{2} \tag{10}$$

其中, $\hat{x}(N_{\mathrm{MM}}+1)$ 是第 $N_{\mathrm{MM}}+1$ 个样本预测值, 而 $\hat{y}(N_{\mathrm{MM}}+1)$ 是校正后的预测结果.

残差序列 ξ 可构成一个随时间变化的函数, N_{MM} 个残差被均分为 n_{MC} 个状态, 每个状态的累计概率是相同的, 示意图如图 3 所示. 以序列数据 X_2 为例分析参数 n_{MC} 对预测效果的影响, 不同 n_{MC} 值与对应的马尔科夫链预测误差的关系曲线如图 4 所示. 可知, 当 n_{MC} 取值为 7 时 RMSE 最小.

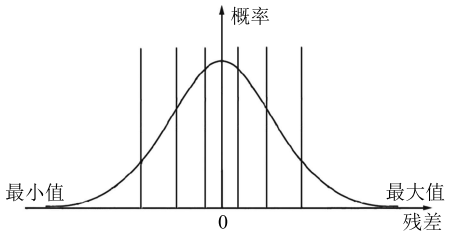


图 3　马尔科夫链的 n_{MC} 个状态

Fig. 3　The n_{MC} states of Markov chain

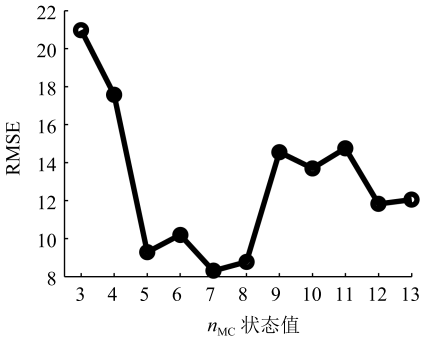


图 4　不同 n_{MC} 值的序列 X_2 马尔科夫链预测误差

Fig. 4　MC prediction errors with different n_{MC} values for sequence X_2

用相同方法测试 5 组 PM2.5 数据, 每组序列对应的最小误差和相应 n_{MC} 值如表 3 所示, 显然 7 是 n_{MC} 的最优值. 在 PM2.5 实际预测中, 基于该方法动态调整最优参数 n_{MC}. 马尔科夫链状态转移概率矩阵有 $n_{\mathrm{MC}}\times n_{\mathrm{MC}}$ 个元素, 例如, 当最优 $n_{\mathrm{MC}}=7$ 时有 49 个元素. 为使矩阵中的元素都有意义, 样本数量 N_{MM} 应取较大值. 在本文实验中, 马尔科夫模型的 $N_{\mathrm{MM}}=100$, 即基于每 100 个连续数据计算第 101 个 PM2.5 预测值.

表 3　基于 5 组序列数据的马尔科夫链 n_{MC} 最优值

Table 3　The optimal n_{MC} values of MC for 5 sequences

序列	最小 RMSE	最优 n_{MC}
X_1	7.2398	7
X_2	8.2994	7
X_3	8.2055	7
X_4	2.4731	7
X_5	20.4794	7

1.2.2 隐马尔科夫模型预测的动态调整

残差序列 ξ 中, 每 2 个相邻样本之间存在 3 种变化状态: 增加、不变、减少. 以变化趋势为预测的介质因素, 则隐马尔科夫模型的隐状态数量为 $n_{\mathrm{HMMH}}=3$. 因此在序列 ξ 中, 连续 3 个样本对应着 9 个变化状态, 如图 5 所示, 即观察状态数量为 $n_{\mathrm{HMMO}}=9$. 当然, 其他数目的隐状态和观察状态也可以用同样方法来定义. 当隐状态为 3, 连续样本有 4 个的时候, 对应着 27 个观察状态. 此外, 2 个相邻样本之间的变化也可分为 4 种情况: 显著增加, 略微增加, 略微减少, 显著减少. 当隐状态为 4, 连续样本有 3 个的时候, 对应着 16 个观察状态. 在隐马尔科夫模型中, 式 (9) 的 n 对应为观察状态数量, 状态转移概率矩阵有 $n_{\mathrm{HMMH}}\times n_{\mathrm{HMMO}}$ 个元素. 例如, 当 $n_{\mathrm{HMMH}}=3$ 且 $n_{\mathrm{HMMO}}=9$ 时, 状态转移概率矩阵有 27 个元素.

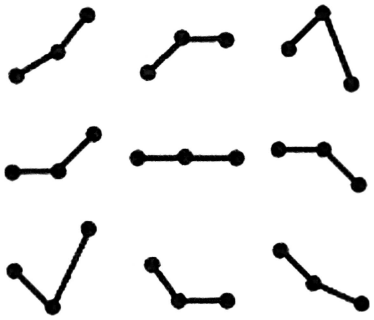


图 5 HMM 的 9 种观察状态

Fig. 5 The 9 observable states of HMM

为了描述隐马尔科夫模型的隐状态, 设两个相邻的残差值为 $\varepsilon(k)$ 和 $\varepsilon(k+1)$, 它们之间的变化为 $(\varepsilon(k+1)-\varepsilon(k))/\varepsilon(k)$. 针对此变化, 我们预设了一个阈值参数 C_t. 当变化大于 C_t 时为增加状态, 当变化小于 $-C_t$ 时为减少状态, 其他情况则为不变状态.

我们还设置了另一个参数 I_t, 用来调整预测值. 对于隐马尔科夫模型, 预测结果的校正方式为

$$\hat{y}(N_{\mathrm{MM}}+1)=\begin{cases}\hat{x}(N_{\mathrm{MM}}+1)(1+I_t), & \text{增加}\\ \hat{x}(N_{\mathrm{MM}}+1), & \text{不变}\\ \hat{x}(N_{\mathrm{MM}}+1)(1-I_t), & \text{减少}\end{cases}\tag{11}$$

隐马尔科夫模型的预测误差与参数 I_t 和 C_t 的对应关系如图 6 所示, 其中误差值 RMSE 的最大值表示为浅色, 最小值表示为深色. C_t 和 I_t 的取值范围均为 [0.1,1], 步长均为 0.01. 因此图 6 中的曲面由 $91\times 91=8\,281$ 个样本点生成, 从中可以找到最小预测误差及其对应的 C_t 值和 I_t 值. 用隐马尔科夫模型预测 PM2.5 时, 参数 C_t 和 I_t 可通过这种方式进行动态调整.

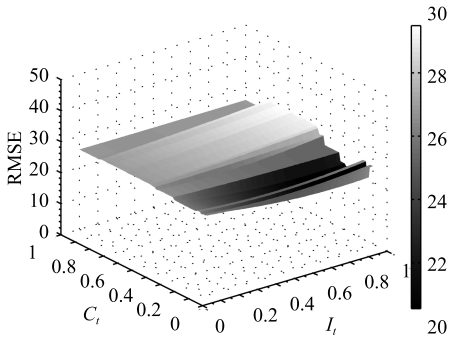


图 6 不同 I_t 值和 C_t 值的序列 X_3 预测误差

Fig. 6 Prediction errors of different I_t and C_t values for sequence X_3

针对 5 组序列数据的隐马尔科夫模型预测, 我们计算得到了每组的最小 RMSE 和相应的最优 C_t 值和 I_t 值, 如表 4 所示. 结果表明, 最优 C_t 值的取值范围为 [0.1, 0.3], 而最优 I_t 值的取值范围为 [0.7, 1.0]. 这些经验范围, 有助于参数动态调整的快速实现.

表 4 隐马尔科夫模型预测的 5 组序列数据的最优 C_t 和 I_t 值

Table 4 The optimal C_t and I_t values of HMM prediction for 5 sequences

序列	C_t	I_t	RMSE
X_1	0.13	0.93	4.4000
X_2	0.26	0.91	19.9012
X_3	0.13	0.82	20.9012
X_4	0.13	1.00	6.0537
X_5	0.26	0.75	28.7373

2 针对 PM2.5 的组合预测

指数平滑法可以提供一组预测值, 但预测值与真实值之间存在误差. 马尔科夫模型可以将误差划分为不同状态, 并预测之后的误差变化. 因此, 我们提出了一个针对 PM2.5 预测的组合算法, 算法结构如图 7 所示. 整个流程由上至下进行, 每个箭头交汇处表示此处需要对两个输入进行操作才能得到下一个输出. 算法中, 指数平滑法得到的预测值由马尔科夫模型 (马尔科夫链或隐马尔科夫模型) 进行校正, 从而提高预测准确度.

在组合算法中, 基于样本数据动态调整相关参数, 包括: 指数平滑法的次数、参数 α、马尔科夫链的

残差状态数 n_{MC}、隐马尔科夫模型的参数 C_t 和 I_t、隐状态数和连续样本数. 为了节省时间, 也可以基于样本学习事先得到针对 PM2.5 的上述最优参数, 然后基于这些参数直接进行预测.

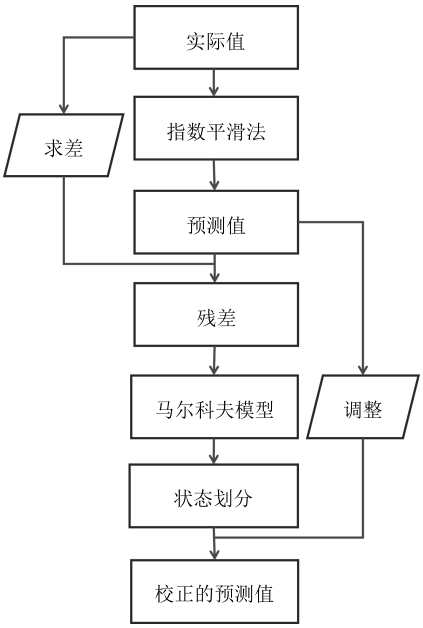


图 7　指数平滑法和马尔科夫模型组合的预测算法

Fig. 7　The combined prediction algorithm from exponential smoothing and Markov model

为确定隐马尔科夫模型的最优隐状态数量和连续样本数量, 我们尝试了 3 个隐状态、3 个连续样本即 9 个观察状态 (3H3S9O) 的情况、3 个隐状态和 4 个连续样本即 27 个观察状态 (3H4S27O) 的情况以及 4 个隐状态和 3 个连续样本即 16 个观察状态 (4H3S16O) 的情况. 基于最优指数平滑和这 3 种参数的隐马尔科夫模型结合得到的算法, 分别对 5 组序列数据进行预测, 预测误差 (RMSE) 如表 5 所示. 由结果可知, 3H3S9O 能在绝大多数情况下得到最好的预测效果.

表 5　三种隐马尔科夫模型对 5 组数据的预测效果

Table 5　Performances of 3 kinds of HMM methods for 5 sequences

序列	3H3S9O	3H4S27O	4H3S16O	最优算法
X_1	3.9309	3.8443	5.5246	3H4S27O
X_2	8.2849	26.8030	27.2460	3H3S9O
X_3	17.0960	17.4780	21.7550	3H3S9O
X_4	5.6534	7.5703	8.8407	3H3S9O
X_5	42.5348	43.9865	54.6437	3H3S9O

3　实验结果与分析

3.1　实验数据和评估标准

为了评估提出的组合算法, 我们采集了 100 组 PM2.5 序列数据进行实验, 这些数据是湖北省武汉市 2014 年 10 月 ∼ 2016 年 10 月的实测数据, 数据采集的时间点是随机选择的, 间距以小时为单位. 此外, 为了评估组合算法在多个城市中的预测效果, 我们采集了 2017 年 3 月 ∼ 4 月武汉、北京、天津、郑州四个城市的 PM2.5 数据进行对比检测, 四个城市的 PM2.5 数据监测点分布图如图 8 所示.

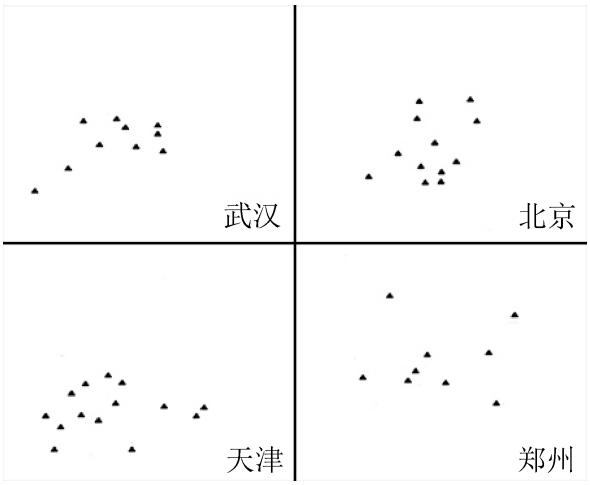


图 8　若干城市 PM2.5 监测点分布示意图

Fig. 8　Distribution of PM2.5 stations in several cities

每组数据预测值与实际值的总偏差, 用以下 3 个评估标准来衡量.

1) 均方根误差 RMSE:

$$RMSE=\sqrt{\frac{(\varepsilon(1))^2+(\varepsilon(2))^2+\cdots+(\varepsilon(n_p))^2}{n_p}} \tag{12}$$

2) 绝对平均误差 AME:

$$AME=\frac{\sum\limits_{i=1}^{n_p}|\varepsilon(i)|}{n_p} \tag{13}$$

3) 百分比平均估计误差 PAEE:

$$PAEE=\frac{\sum\limits_{i=1}^{n_p}(\varepsilon(i))^2}{n_p\times\overline{Y}} \tag{14}$$

其中, $\overline{Y}=(\hat{y}(1)+\hat{y}(2)+\cdots+\hat{y}(n_p))/n_p$ 为校正后预测的平均值, n_p 为每组的个体误差数.

上述实验数据和评价标准用于测试本文相关算法: 指数平滑法 ES、马尔科夫链 MC、隐马尔科夫模型 HMM、指数平滑法结合马尔科夫链 ESMC、指数平滑法结合隐马尔科夫模型 ESHMM. 比较后得到的最优算法, 又与人工神经网络 ANN、支持向量机 SVM、自回归滑动平均模型 ARMA、灰色模型 GM 四种常用预测方法进行对比. 其中, ANN 采用的是 BP-ANN 模型, 中间层数为 3, 中间层节点数为 27, 每层的权系数和阈值均为自学习, 网络隐含层神经元传递函数和输出层神经元函数分别为 tansig 和 purelin, 训练函数和学习函数分别选取 traingcgf, learndm; SVM 采用的是 libsvm 工具, 模型参数 C 和 γ 均基于训练数据自主调整; ARMA 采用的是已知预测效果最好的 ARMA(5,3) 模型, 自回归阶数为 5, 滑动平均阶数为 3, 自回归系数和滑动平均系数等均为自动调整; GM 采用的是 GM(1,1), 其中累加生成序列的权重系数 μ 为常用值 0.5, 基于最小二乘法自动求解参数向量 $\boldsymbol{a}$, $\boldsymbol{b}$.

3.2 结果与分析

从实验中我们发现, 事先学习参数的方法和动态调整参数的方法在时间上有所区别. 例如, 在 Intel i5 CPU 2.50 GHz, 4.00 GB RAM 与 MATLAB R2010b 配置电脑上, 事先学习参数的算法在单次预测中耗时 1.35 秒, 而临时学习动态调整参数的算法在单次预测中耗时 156.40 秒. 对于同一监测点, 其地理位置、污染源分布、所属气候特征等因素都相对稳定, 因此空气污染指标的变化有其规律性. 我们以武汉市 20 组 PM2.5 预测数据为例, 统计出事先学习参数和动态调整参数两种方法所得预测结果的 RMSE 的平均偏差 $(|RMSE_{\text{事先学习}} - RMSE_{\text{动态调整}}|)/(RMSE_{\text{动态调整}})$ 为 6.70 %, 也证明了两种方法预测效果差别不大. 与本文提出算法一致, 在接下来的实验中, 使用的是参数动态调整的方法.

基于 3 种评价指标, 针对本文算法在参数动态调整情况下的实验结果如表 6 所示. 可见平均 RMSE 和平均 AME 按照降序排列均为: ESMC, MC, ES, HMM, ESHMM. 平均 PAEE 按降序排列为: ESMC, MC, HMM, ES, ESHMM. 可知, 本文的动态 ESHMM 的平均预测误差最小.

动态 ESHMM 算法与 4 种常用预测方法的比较如表 7 所示. 可知, ESHMM 的平均 RMSE、平均 AME 和平均 PAEE 均为最小. 基于平均 RMSE 值, 算法按照降序排列为: ANN, SVM, ARMA, GM, ESHMM.

从每组 PM2.5 数据的预测序列中各取 6 个测量点, 将 600 个数据的实际值和预测值绘制成散点图, 如图 9 所示. 散点沿中轴线的聚集程度反映了预测算法的准确性与稳定性, 可见, 动态 ESHMM 结果的聚集程度最高, 预测值和实际值间的一致性最强, 提供了最为可靠的 PM2.5 浓度预测.

表 6 针对 100 组序列数据的平均评估值

Table 6 Averaged evaluation criteria of 100 sequences

序列	ES	MC	HMM	ESMC	ESHMM
平均 RMSE	12.6745	17.0434	12.4850	17.8595	10.1843
平均 AME	10.3838	13.1197	9.9381	14.0072	8.8336
平均 PAEE	3.2151	5.6512	3.4993	6.6281	2.5090

表 7 与现有 4 种算法的预测误差比较

Table 7 The comparison of prediction errors with 4 existing algorithms

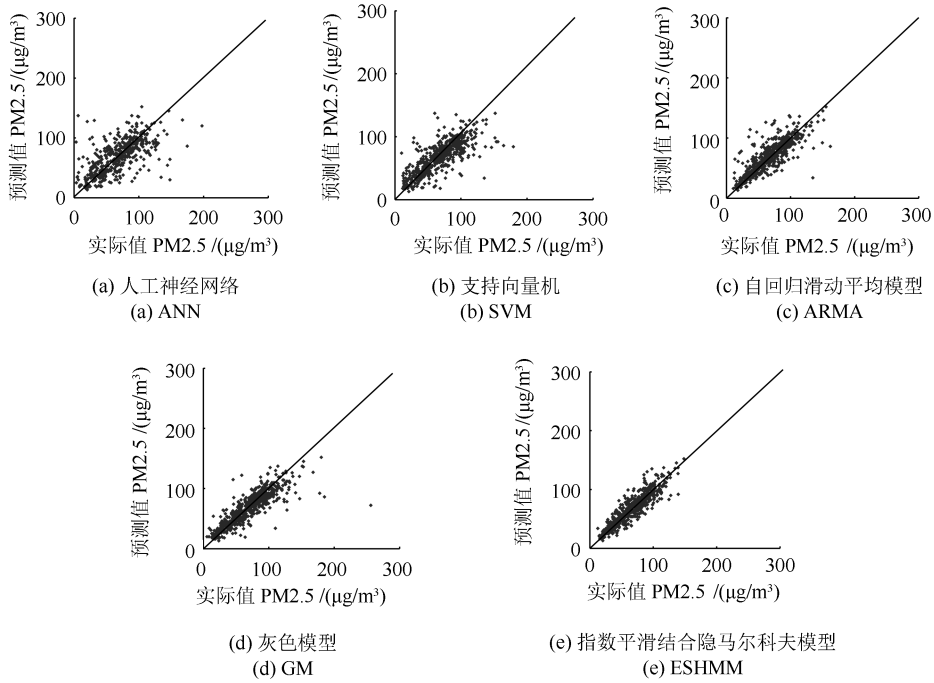
序列	ANN	SVM	ARMA	GM	ESHMM
平均 RMSE	23.7594	18.4532	15.7469	13.9438	10.1843
平均 AME	17.7772	14.3728	10.3086	10.9673	8.8336
平均 PAEE	5.6802	3.7900	4.0373	3.8927	2.5090

此外, 我们针对若干城市 PM2.5 数据进行了预测. 实验数据采集于武汉、北京、天津、郑州四个城市, 每个城市各 22 组, 数据以小时为间距, 采集时间为 2017 年 3 月和 4 月. 预测效果如图 10 所示, 在 RMSE、AME、PAEE 的评价指标均值柱状图中, 对应于每个城市, 从左到右对应的算法模型依次为 ES、MC、HMM、ESMC、ESHMM. 可见本文提出的 ESHMM 组合算法均得到了最优的预测效果. 实验结果表明, ESHMM 模型适用于不同城市的 PM2.5 数值预测.

4 结论

细颗粒物 PM2.5 是重要的空气污染测量数据, PM2.5 的浓度预测对于环境保护有重要意义. 考虑到预测的效率、实用性和准确性, 本文提出了一个基于单时间序列的, 将动态指数平滑法和动态马尔科夫模型相结合的组合算法. 算法的最优参数可以通过历史数据的学习预先确定, 或者通过当前样本的学习动态调整.

本文通过 PM2.5 序列样本说明了算法的实现过程与参数的动态调整, 并基于大量实测 PM2.5 数据与多种评估指标验证了算法的有效性, 而且与现有的常用预测算法做了对比, 证明了动态 ESHMM 算法的优点. 我们提出的动态调整预测模型不局限于 PM2.5 数据, 通过样本的学习, 也可以应用到其他类型的单时间序列数据的预测.

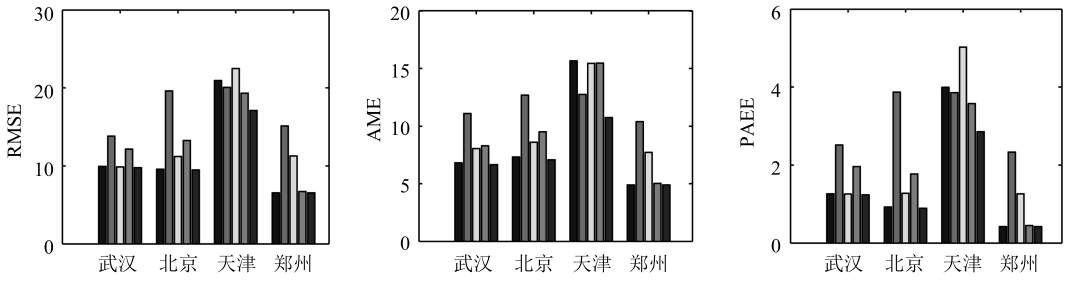


(a) 人工神经网络 (a) ANN　(b) 支持向量机 (b) SVM　(c) 自回归滑动平均模型 (c) ARMA

(d) 灰色模型 (d) GM　(e) 指数平滑结合隐马尔科夫模型 (e) ESHMM

图 9　预测值与实际值的 PM2.5 散点图

Fig. 9　Scatter plot of predicted versus observed PM2.5



(a) 评估标准 RMSE (a) Evaluation RMSE　(b) 评估标准 AME (b) Evaluation AME　(c) 评估标准 PAEE (c) Evaluation PAEE

图 10　基于 5 种算法的武汉、北京、天津、郑州 PM2.5 预测误差

Fig. 10　Prediction errors of PM2.5 in Wuhan, Beijing, Tianjin, and Zhengzhou from 5 algorithms

References

1 Chan Y, Xia L, Ren Y, Chen Y T. Multi-scale modelling on PM2.5 encapsulation inside doubly-layered graphene. *IET Micro and Nano Letters*, 2015, **10**(12): 696−699

2 Zhan H L, Li Q, Zhao K, Zhang L W, Zhang Z W, Zhang C L, Xiao L Z. Evaluating PM2.5 at a construction site using terahertz radiation. *IEEE Transactions on Terahertz Science and Technology*, 2015, **5**(6): 1028−1034

3 Rajasegarar S, Havens T C, Karunasekera S, Leckie C, Bezdek J C, Jamriska M, Gunatilaka A, Skvortsov A, Palaniswami M. High-resolution monitoring of atmospheric pollutants using a system of low-cost sensors. *IEEE Transactions on Geoscience and Remote Sensing*, 2014, **52**(7): 3823−3832

4 Shaban K B, Kadri A, Rezk E. Urban air pollution monitoring system with forecasting models. *IEEE Sensors Journal*, 2016, **16**(8): 2598−2606

5 Díaz-Robles L, Ortega J C, Fu J S, Reed G D, Chow J C, Watson J G, Moncada-Herrera J A. A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: the case of Temuco, Chile. *Atmospheric Environment*, 2008, **42**(35): 8331−8340

6 Stadlober E, Hörmann S, Pfeiler B. Quality and performance of a PM10 daily forecasting model. *Atmospheric Environment*, 2008, **42**(6): 1098−1109

7 Macaira P M, Sousa R C, Oliveira F L C. Forecasting brazil′s electricity consumption with pegels exponential smoothing techniques. *IEEE Latin America Transactions*, 2016, **14**(3): 1252−1258

8 Taylor J W, Roberts M B. Forecasting frequency-corrected electricity demand to support frequency control. *IEEE Transactions on Power Systems*, 2016, **31**(3): 1925−1932

9 Jamal W, Das S, Oprescu I A, Maharatna K. Prediction of synchrostate transitions in EEG signals using Markov chain models. *IEEE Signal Processing Letters*, 2015, **22**(2): 149−152

10 Lawlor S, Rabbat M G. Time-varying mixtures of Markov chains: an application to road traffic modeling. *IEEE Transactions on Signal Processing*, 2017, **65**(12): 3152−3167

11 Razin Y S, Pluckter K, Ueda J, Feigh K. Predicting task intent from surface electromyography using layered hidden Markov models. *IEEE Robotics and Automation Letters*, 2017, **2**(2): 1180−1185

12 Soualhi A, Clerc G, Razik H, El Badaoui M, Guillet F. Hidden Markov models for the prediction of impending faults. *IEEE Transactions on Industrial Electronics*, 2016, **63**(5): 3271−3281

13 Samet H, Mojallal A. Enhancement of electric arc furnace reactive power compensation using Grey-Markov prediction method. *IET Generation, Transmission and Distribution*, 2014, **8**(9): 1626−1636

14 Chen L, Tian B B, Lin W L, Ji B, Li J Z, Pan H H. Analysis and prediction of the discharge characteristics of the lithium-ion battery based on the Grey system theory. *IET Power Electronics*, 2015, **8**(12): 2361−2369

15 de Lima G R T, Stephany S, de Paula E R, Batista I S, Abdu M A. Prediction of the level of ionospheric scintillation at equatorial latitudes in Brazil using a neural network. *Space Weather*, 2015, **13**(8): 446−457

16 Nagulan S, Selvaraj J, Arunachalam A, Sivanandam K. Performance of artificial neural network in prediction of heave displacement for non-buoyant type wave energy converter. *IET Renewable Power Generation*, 2017, **11**(1): 81−84

17 Moshkbar-Bakhshayesh K, Ghofrani M B. Development of a robust identifier for NPPs transients combining ARIMA model and EBP algorithm. *IEEE Transactions on Nuclear Science*, 2014, **61**(4): 2383−2391

18 Wei M, Kim K. Intrusion detection scheme using traffic prediction for wireless industrial networks. *Journal of Communications and Networks*, 2012, **14**(3): 310−318

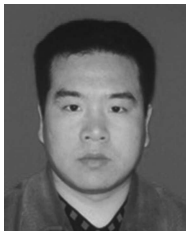
19 Wu Qi, Yan Hong-Sen, Wang Bin. Product sales forecasting model based on robust wavelet ν-support vector machine. *Acta Automatica Sinica*, 2009, **35**(7): 1227−1232
(吴奇, 严洪森, 王斌. 基于鲁棒小波 $\nu-$支持向量机的产品销售预测模型. 自动化学报, 2009, **35**(7): 1227−1232)

20 Liu Y Q, Sun Y, Infield D, Zhao Y, Han S, Yan J. A hybrid forecasting method for wind power ramp based on orthogonal test and support vector machine (OT-SVM). *IEEE Transactions on Sustainable Energy*, 2017, **8**(2): 451−457

21 Gupta S, Kambli R, Wagh S, Kazi F. Support-vector-machine-based proactive cascade prediction in smart grid using probabilistic framework. *IEEE Transactions on Industrial Electronics*, 2015, **62**(4): 2478−2486

张熙来 武汉大学计算机学院硕士研究生. 2016 年获武汉大学计算机学院学士学位. 主要研究方向为模式识别, 自然语言处理.
E-mail: runningman_hamei@163.com
(**ZHANG Xi-Lai** Master student at the School of Computer, Wuhan University. She received her bachelor degree from Wuhan University in 2016. Her research interest covers pattern recognition and natural language processing.)

赵俭辉 武汉大学计算机学院副教授. 2004 年获新加坡南洋理工大学博士学位. 主要研究方向为模式识别, 计算机图形图像, 图像处理. 本文通信作者.
E-mail: jianhuizhao@whu.edu.cn
(**ZHAO Jian-Hui** Associate professor at the School of Computer, Wuhan University. He received his Ph. D. degree from Nanyang Technological University, Singapore in 2004. His research interest covers pattern recognition, computer graphics, and image processing. Corresponding author of this paper.)



蔡 波 武汉大学计算机学院副教授. 2003 年获武汉大学博士学位. 主要研究方向为模式识别, 计算机图形学, 虚拟现实技术. E-mail: bo_cai@yeah.net
(**CAI Bo** Associate professor at the School of Computer, Wuhan University. He received his Ph. D. degree from Wuhan University in 2003. His research interest covers pattern recognition, computer graphics, and virtual reality technology.)