基于 SSA 优化 BP 神经网络的污染物浓度二次预测模型

黄邦菊，张炜亮

（中国民用航空飞行学院 空中交通管理学院，四川 广汉 618307）

摘要：为解决一次预报模型模拟结果不理想的情况，使用主成分分析（PCA）对 14 项气象影响因素进行降维处理并提取 4 项综合评价指标，使用麻雀搜索算法（SSA）优化 BP 神经网络的二次预测模型和将一次预报的结果与真实数据的差值作为输出，对预测模型进行训练并做出预测的方法。将模型应用于国内某个地区，用相应的数据对模型进行验证。结果表明，基于 SSA 优化 BP 神经网络的预测模型和将误差引入的新模型均较 BP 神经网络模型有更高的精确度和更强的泛化能力。

关键词：污染物浓度预测；主成分分析（PCA）；BP 神经网络；麻雀搜索算法（SSA）

中图分类号：X831 **文献标志码**：A **文章编号**：1671－1807(2023)05－0172－06

随着科学技术的进步和社会工业化的不断发展，工业排放和人类活动所造成的大气污染越来越严重，各种污染物随着排放进入大气中，在阳光、温度、湿度等条件下发生不同的化学反应，不仅对大气造成二次污染，同样也会对生活在地球上的人类和其他生物的健康造成严重威胁[1]。污染防治实践表明，精确的空气质量预报模型能够有效地检测、监控大气中的污染物质，提前采取的合理、有效的防治手段，能够大大降低污染物质对环境以及人体健康的危害。

早期对污染物浓度预测使用的主要是一些比较原始的数学统计模型，但污染物浓度变化会受到多种外界因素影响，其复杂性较大，使用自回归移动平均（autoregressive moving integrated average，ARIMA）等统计模型处理相关数据时，其结果并不能达到预期目标[2]。目前常用 WRF-CMAQ 模拟体系对空气质量进行预报，但受制于种种限制因素，该预报模型的结果也并不理想[3]。随着机器学习在预测领域的广泛应用，人工神经网络（artificial neural network，ANN）开始在污染物浓度预测方面使用[4]，随机森林算法因其高灵活性、训练速度快、泛化能力强等优点也广泛应用在污染物浓度预测领域[5]。混合模型也在此基础上开始广泛应用[6]。BP 神经网络具有容易陷入局部最小值而无法获得最优解以及学习能力有限等缺点[7]。本文使用主成分分析（principal component analysis，PCA）方法对数据进行降维处理[8]，在此基础上提出了麻雀搜索算法（SSA）优化 BP 神经网络的污染物浓度二次预测模型以及将一次预报的结果与真实数据的差值作为输出，对预测模型进行训练并做出预测的方法。通过小时维度预测可以获得未来 72 h 即 3 d 的污染物浓度变化情况，且结果显示模型具有良好的预测能力。在环境保护越来越重要的今天，该模型能够为城市环境监测及管理部门提供相应的帮助。

1 方法

1.1 主成分分析

主成分分析方法是一种多变量分析方法，其使用前提是原始数据需要具有高度相关性。该方法是将多个具有相关性的变量重新组合，通过降维将其组合成几个新的综合指标，并将原始数据中冗余的信息剔除，使得数据更加简洁[9]。可以分为以下几个步骤进行计算[10]：

1）对整个数据进行标准化处理。

收稿日期：2022-10-09

基金项目：民航局空管局横向项目（0052119）。

作者简介：黄邦菊（1966—），女，四川什邡人，中国民用航空飞行学院空中交通管理学院，副教授，硕士，研究方向为航空情报；通信作者张炜亮（1998—），男，新疆昌吉人，中国民用航空飞行学院空中交通管理学院，硕士研究生，研究方向为航空气象。

$$x_{ij}^{*} = \frac{x_{ij} - \overline{x}_j}{\sigma_j} \tag{1}$$

式中:x_{ij} 为样本数据集中观测的第 i 个样本第 j 个变量的原始值;$\overline{x}_j$ 为第 j 个变量的均值;σ_j 为第 j 个变量的标准差;x_{ij}^{*} 为第 i 个样本第 j 个变量的标准化值。

2)计算相关系数矩阵 $\boldsymbol{R} = (r_{ij})_{m\times m}$。

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)^2 \sum_{k=1}^{n}(x_{kj} - \overline{x}_j)^2}} \tag{2}$$

式中:x_{ki} 为第 k 个样本第 i 个变量的原始值;x_{kj} 为第 k 个样本第 j 个变量的原始值;$\overline{x}_i$ 和 $\overline{x}_j$ 为第 i 和第 j 个变量的均值;r_{ij} 为第 i 个和第 j 个变量的相关系数。

3)求解相关系数矩阵的特征值与特征向量。

4)提取主成分。

$$w_t = \frac{\lambda_t}{\sum_{u=1}^{m}\lambda_u} \tag{3}$$

式中:w_t 为主成分 p_t 的方差贡献率,w_t 越大,说明主成分 P_t 对原始变量的解释能力越强;λ_t 为第 t 个主成分 p_t 的方差;λ_u 为第 u 个变量的方差;$\sum_{u=1}^{m}\lambda_u$ 为 m 个原始变量的总方差。前 n 个主成分的累计方差百分比 W_n 的计算公式为

$$W_n = \sum_{i=1}^{n}\frac{\lambda_i}{\sum_{u=1}^{m}\lambda_u} \tag{4}$$

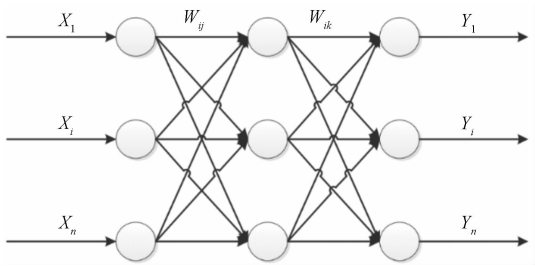
式中:W_n 为 n 个提取了的主成分对原始变量信息量解释的占比。

1.2 BP 神经网络

BP 神经网络是一个多层前馈神经网络。在传播过程中,输入的信号通过网络持续向前方传递,而误差朝后方传递。隐含层可以设置为一个或多个,在经过隐含层的处理后,输入信号便从输出层输出,此时在经过比较后,如果输出与期望的差值较大,系统则会反向传播。根据不断的循环,整个网络的运行结果会逐渐向着之前所设定的期望值靠近。BP 神经网络结构如图 1 所示。

由图 1 可知,X_1 等作为该函数的自变量,Y_1 等作为因变量。因此,整个系统中的所有自变量与所有因变量产生的函数映射关系便成为 BP 神经网络的基本构造。BP 神经网络的运行步骤如下[11]。

步骤 1:将各权数和阈数初始化,并赋值取值为[−1,+1]之间的随机数。



X_1、X_i、X_n 为输入值,通过输入层经过多层隐含层的处理后输出;ω_{ij}、ω_{ik} 为权值;Y_1、Y_i、Y_n 为通过预测得到的值

图 1 BP 神经网络结构

步骤 2:为网络提供初始数据 $A_k = (a_1^k, a_2^k, \cdots, a_n^k)$,$Y_k = (y_1^k, y_2^k, \cdots, y_n^k)$。

步骤 3:把 $A_k = (a_1^k, a_2^k, \cdots, a_n^k)$ 作为输入层,联合权值 $\{\omega_{ij}\}$ 和阈值 $\{\theta_{ij}\}$ 计算中间层各输出值 $\{b_j\}$。

步骤 4:再用 $\{b_j\}$、$\{\omega_{ij}\}$ 和 $\{\theta_{ij}\}$ 计算输出层各个单元的响应 $\{C_j\}$。

步骤 5:使用希望输出 $Y_k = (y_1^k, y_2^k, \cdots, y_n^k)$,实际输出 $\{C_j\}$ 来计算各单元误差 $\{d_j^k\}$。

步骤 6:用权值 $\{\omega_{ij}\}$,中间层的输出值 $\{b_j\}$ 以及整个输出层的误差 $\{d_j^k\}$ 来计算中间层误差 $\{e_j^k\}$。

步骤 7:根据已有结果修正权值 $\{\omega_{ij}\}$ 和阈值 $\{\theta_{ij}\}$。

步骤 8:直至误差符合要求。

1.3 麻雀搜索算法(SSA)

麻雀搜索算法(SSA)是一种新型的智能优化算法,其灵感是在观察到鸟类的觅食活动后产生。在鸟群中有负责搜索工作的,有追随搜索者的,还有负责警戒侦察的。在 SSA 中,群体内的最优个体在搜索过程中会优先获取食物。作为探索者,其有着更为宽广的活动范围,相比于追随者更大,并且每迭代一次,探索者的位置就会有相应的更新[12]。

$$\boldsymbol{X}_{i,j}^{t+1} = \begin{cases} \boldsymbol{X}_{i,j}^{t}\exp\left(\dfrac{-i}{\alpha \mathrm{iter}_{\max}}\right), & R_2 < \mathrm{ST} \\ \boldsymbol{X}_{i,j}^{t} + \boldsymbol{Q}\boldsymbol{L}, & R_2 \geqslant \mathrm{ST} \end{cases} \tag{5}$$

式中:$\boldsymbol{X}_{i,j}$ 为每个麻雀所存在的位置;i 表示目前该算法已经迭代运行了的次数;$\mathrm{iter}_{\max}$ 为整个算法运行过程中最大的迭代次数;α 为[0,1]区间内一个随机分配的数;R_2($R_2 \subseteq [0,1]$) 为该算法的预警值;ST(ST⊆[0.5,1])为该算法的安全值;$\boldsymbol{Q}$ 为一个随机分配且服从正态分布的数;$\boldsymbol{L}$ 为一个矩阵,其大小为 $1\times d$,其构成元素全为 1。

当 $R_2 < \mathrm{ST}$ 时,数值域内无捕食者,此时负责

探索任务的鸟开始在区域内进行搜索；当 $R_2 \geqslant ST$ 时，探索者侦测到该区域内存在捕食者，会有一定的危险，该群体迅速向安全区域移动。追随者通过如下公式来更新位置：

$$\boldsymbol{X}_{i,j}^{t+1} = \begin{cases} Q\exp\left(\dfrac{\boldsymbol{X}_{\text{worst}}^{t} - \boldsymbol{X}_{i,j}^{t}}{i^2}\right), & i > n/2 \\ \boldsymbol{X}_{P}^{t+1} + \left|\boldsymbol{X}_{i,j}^{t} - \boldsymbol{X}_{P}^{t+1}\right| \boldsymbol{A}^{+}\boldsymbol{L}, & \text{其他} \end{cases} \tag{6}$$

式中：$\boldsymbol{X}_P$ 为全局搜索后搜索者发现的最优的位置；$\boldsymbol{X}_{\text{worst}}$ 为全局搜索后搜索者所处的最差位置；n 为种群规模；$\boldsymbol{A}$ 为一个 $1 \times d$ 的矩阵，且矩阵由随机的 1 或 -1 构成；$\boldsymbol{A}^+$ 定义如下：

$$\boldsymbol{A}^{+} = \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1} \tag{7}$$

当 $i > n/2$ 表明第 i 个状态较差的追随者其适应度值较低，需要寻找其他能够觅食的地方。在算法中，假设种群内的 10%～20%的个体（本文中设定 20%）会意识到危险，并且感到危险的个体其初始位置将会随机产生[13]。

$$\boldsymbol{X}_{i,j}^{t+1} = \begin{cases} \boldsymbol{X}_{\text{best}}^{t} + \beta\left|\boldsymbol{X}_{i,j}^{t} - \boldsymbol{X}_{\text{best}}^{t}\right|, & f_i > f_{\text{g}} \\ \boldsymbol{X}_{i,j}^{t} + K\left[\dfrac{\left|\boldsymbol{X}_{i,j}^{t} - \boldsymbol{X}_{\text{worst}}^{t}\right|}{(f_i - f_{\text{w}}) + \varepsilon}\right], & f_i = f_{\text{g}} \end{cases} \tag{8}$$

式中：$\boldsymbol{X}_{\text{best}}$ 为目前整个算法运行完后所找到的最好的位置；β 为正态分布随机数，其方差大小为 1 且服从于均值 0，设置其的目的是用来控制步长；K 为一个随机数，其范围在区间[-1,1]内；ε 的存在目的是为了防止分母为 0，其大小为最小常数；f_i 为算法运行中第 i 个麻雀的适应度值；f_{g} 为在当前情况下的最佳适应度值；f_{w} 为当前情况下的最差适应度值。当 $f_i > f_{\text{g}}$ 时，量值处于数值域的边缘，易遭遇捕食者；当 $f_i = f_{\text{g}}$ 时，收到异常的干扰，所有项向中心移动，处于安全保护状态。

2　实验数据及结果比较

2.1　数据来源

本文所使用的数据为国内某地区监测点长期测量得到的空气质量预报基础数据，包括从 2020 年 7 月 23 日至 2021 年 7 月 13 日的监测点的空气质量预报基础数据，内容为二氧化硫（SO_2）、二氧化氮（NO_2）、粒径小于 10 μm 的颗粒物（PM_{10}）、粒径小于 2.5 μm 的颗粒物（$PM_{2.5}$）、臭氧（O_3）、一氧化碳（CO）这 6 种污染物浓度的一次预报数据以及从 2019 年 4 月 16 日至 2021 年 7 月 13 日的近地 2 m 温度、地表温度、比湿、湿度、近地 10 m 风速、近地 10 m 风向、雨量、云量、边界层高度、大气压、感热通量、潜热通量、长波辐射、短波辐射、地面太阳辐射和 6 种污染物浓度的实测数据。

2.2　数据预处理

在收集好数据后，经过检查发现数据集中还存在一些异常数据，在使用之前还需进行处理。数据异常出现的原因在于监测站点设备调式、维护等导致部分时间段内的监测数据出现缺失或者受监测站点及其附近某些偶然因素的影响，实测数据在某个小时（某天）的数值偏离数据正常分布。因此，采用局部拉以达法则（3σ 准则）和线性拟合这两种方法对异常数据进行处理。局部拉以达法则用于周围数据波动较小的情况，整个大气系统处于一个较稳定的状态。线性拟合用于局部浓度值异常的情况，因为在这种情况下单纯使用插值法或是均值法得到的数据精确性较差，所以可以使用几个月或者一年的数据做出趋势图，观察其季度性变化规律，从而使插值更加精确。

2.3　影响因素分析

由于数据中的影响因素多达 14 项，且各个因素对 AQI（空气质量指数）的影响程度不同，所以需要对它们进行主成分分析处理，方便后面进行预测分析。

首先将 14 项气象影响因素使用 PCA 进行降维处理，并通过 KMO（Kaiser-Meyer-Olkin）与 Bartlett 球形度检验结果（表 1）来判断该数据进行 PCA 的合理性。

表 1　KMO 与 Bartlett 球形度检验结果

KMO 取样适切性量数		0.760
巴特利特球形度检验	近似卡方	20 510.834
	自由度	91
	显著性	0.000

从表 1 可以看出，KMO 取样适切性量数的值为 0.760，大于 0.500 并且巴特利特球形度检验的显著性要小于 0.001，说明这 14 项气象影响因素能够进行主成分分析。

再对数据进行了 PCA 降维处理后，可以提取出 4 个主成分的初始特征值、方差百分比和累计方差百分比，见表 2。

根据表 2 可知，提取了 4 个主成分后，其累计方差百分比达到了 81.493%，大于 80%，可以说明提取的这 4 个主成分具有代表性，是比较合适的。根据碎石图（图 2）可以看出，过了第 4 个主成分，整体的曲线斜率便减小了，变得平缓，这同样也能证明对 14 项气象影响因素进行 PCA 降维处理后提取出 4 个主成分是合适的。

表 2　主成分方差百分比和累计方差百分比

主成分	初始特征值	方差百分比/%	累计方差百分比/%
主成分 1	6.769	48.352	48.352
主成分 2	2.352	16.799	65.150
主成分 3	1.186	8.470	73.621
主成分 4	1.102	7.872	81.493

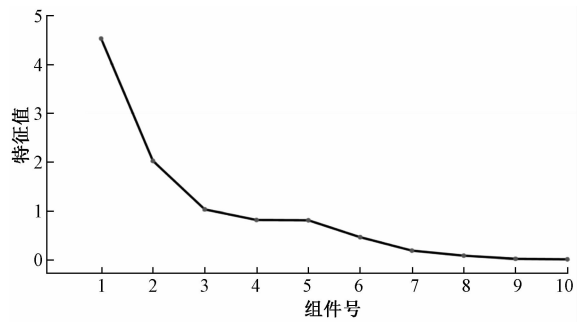


图 2　经过 PCA 降维处理后的碎石图

根据主成分因子矩阵(表 3)可以看出,主成分 1 中主要包括地面太阳能辐射、地表温度、潜热通量、感热通量以及边界层高度,主成分 2 中主要包括近地 2 m 温度、湿度、长波辐射和比湿,主成分 3 中主要包括雨量、云量以及大气压,主成分 4 则是近地 10 m 风速和近地 10 m 风向。从各个主成分的结构来看,主成分 1 与主成分 2 主要包含的是温度和湿度状况的气象影响因素,并且主成分 1 和主成分 2 的累计方差占比达到了 65.150%,这就能够表明与温度和湿度有关的指标对污染物浓度有着较大的影响。其次,主成分 3 主要包括的是雨量以及云量等影响因素,因此可以将其作为雨量云量的综合描述指标。主成分 4 则主要是与风有关的影响因素,故将将其作为风状态的综合描述指标。

表 3　主成分因子负荷矩阵

变量	主成分 1	主成分 2	主成分 3	主成分 4
近地 2 m 温度	−0.032	0.299	−0.024	−0.020
湿度	0.022	−0.260	0.084	0.046
近地 10 m 风速	−0.139	0.219	−0.079	−0.651
雨量	−0.030	0.110	0.337	0.272
云量	0.024	−0.030	0.333	−0.042
边界层高度	0.061	0.174	0.013	−0.004
大气压	−0.058	0.006	−0.304	0.092
地面太阳能辐射	0.326	−0.280	−0.012	0.060
长波辐射	−0.133	0.461	0.375	−0.006
地表温度	0.172	0.014	0.033	0.081
比湿	0.002	−0.161	0.133	0.062
近地 10 m 风向	−0.135	0.267	0.031	0.521
潜热通量	0.283	−0.185	0.031	0.000
感热通量	0.254	−0.134	0.039	−0.030

2.4　预测结果及评价

在完成了对影响因素的分析后,就可以开始进行 BP 神经网络的训练和预测。将经过 PCA 处理后的 3 316 h 的 4 种气象综合描述指标的数据、3 316 h的某一污染物浓度的数据作为输入数据,使用 MATLAB 进行预测。经过多次训练,得到一个表现效果最好的模型,但是其模型拟合度只有 66.17%,显然不能满足需求。为此使用麻雀搜索算法优化 BP 神经网络的初始权重与阈值,首先将训练集与测试集整合,用该整体的均方误差(mean square error,MSE)来表示适应度值。通过结果可以判断,如果适应度函数值越小,则表明训练的效果越好、准确度越高,并且还能表明该模型具有更好的预测精度。基于 SSA-BP 的污染物浓度预测流程如图 3 所示。

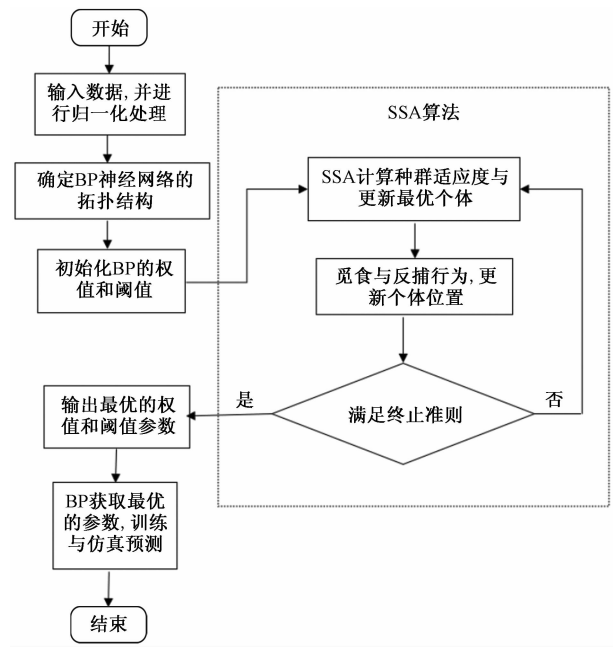


图 3　基于 SSA-BP 的污染物浓度预测流程

为了提高预测精确度,选择采用两种方法来对污染物浓度进行预测[14]。

方法 1:使用经过麻雀搜索算法优化后的 BP 神经网络预测模型直接对污染物浓度进行预测,并得到相应的预测结果。

方法 2:先计算出一次预报污染物浓度与相对应时间点真实浓度之间的差值,即误差,然后通过将预测数据的气象综合描述指标以及一次预报污染物浓度作为该模型的输入,将计算出的误差作为输出,对优化后的预测模型进行训练[15]。

通过方法 1 的输入与输出数据对模型的训练效

果由仿真预测误差(图 4)可以看出,仿真误差基本控制在－0.2～0.15,属于正常范围。并且整个模型拟合度达 89.73%,将精确度与优化前 BP 神经网络预测模型进行对比,发现其结果提高了 23.56%。可以明显看出通过麻雀搜索算法对 BP 神经网络进行优化后,预测精度有了明显的提升。

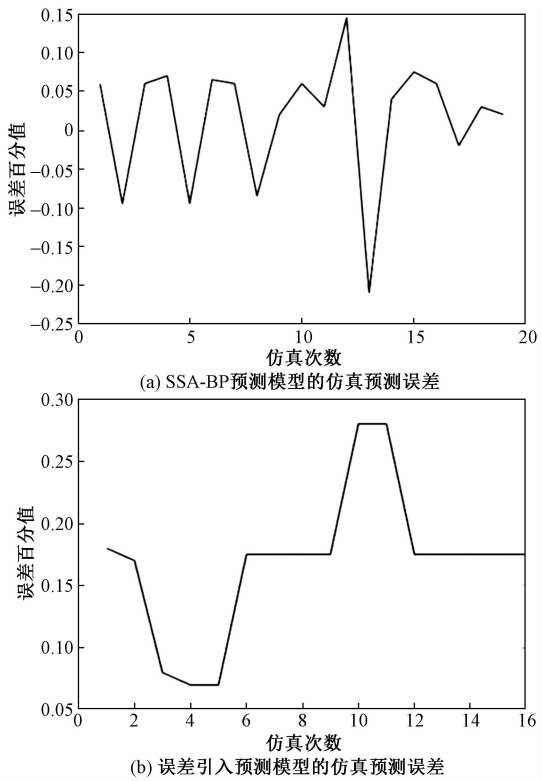


图 4　SSA-BP 预测模型与将误差引入的预测模型的仿真预测误差

然后通过将预测数据的气象条件以及一次预报污染物浓度作为训练模型的输入,将计算出的误差作为输出对模型进行训练,发现其效果依然较好,对误差的预测中仿真预测误差在 0.025～0.255,处于高水平范围。总体模型拟合度也达到了 93.71%(表 4),依然高于优化前的 BP 神经网络预测模型的预测精度。这证明通过数据对该模型训练用来实现对误差的预测是可行的。

表 4　3 种模型的精确度与 MSE 对比

模型	精确度/%	MSE
BP 神经网络	66.17	1.854 7
SSA-BP 神经网络	89.73	0.056 219
引入误差的 SSA-BP 神经网络	93.71	0.052 687

3　结论

1)使用主成分分析法对 14 种气象影响因素进行分析处理,提取出 4 个具有代表性主成分,将其作为预测模型的输入,使得预测模型的复杂性大大降低。

2)提出了一种基于 SSA 算法优化 BP 神经网络的污染物浓度二次预测模型。使用 SSA 算法对 BP 神经网络的权值和阈值进行优化,提升了模型的运行效率,改善了 BP 神经网络容易陷入局部最小值等缺陷。

3)本文提出的优化模型其预测精度较 BP 神经网络模型有了大幅提高。并且在预测过程中将一次预测数值与真实值的误差考虑在内,也大幅提升了预测精确度,该误差引入方法为污染物浓度二次预测提供了一种新的思路。这两种方法的模型结构简单,较易实现,且具有一定的实用性。可以用于单一污染物浓度的日均预测,也可以做小时或各种时间维度的污染物浓度预测。其结果不仅可以用来估算 AQI 以及预测时间段内的首要污染物,还可以为城市治理和环境监测提供数据支持和决策依据。

参考文献

[1] 郝吉明,马广大,王书肖. 大气污染控制工程[M]. 北京:高等教育出版社,2010.

[2] 周尧. 基于 ARIMA 和长短期记忆模型的南京市空气质量预测[D]. 南京:南京审计大学,2021.

[3] 伯鑫. 空气质量模型(SMOKE、WRF、CMAQ 等)操作指南及案例研究[M]. 北京:高等教育出版社,2010.

[4] 张勇,黎云祥,权秋梅. 基于属性简约和 BP 神经网络的 $PM_{2.5}$ 预测模型[J]. 环境科学与技术,2017,40(S1):341-346.

[5] 张志刚,徐莹,张锦秋,等. 基于随机森林的公路隧道 CO 气体浓度预测模型[J]. 科学技术与工程,2022,22(26):11729-11735.

[6] 刘炳春,来明昭,齐鑫,等. 基于 Wavelet-LSTM 模型的北京空气污染物浓度预测[J]. 环境科学与技术,2019,42(8):142-149.

[7] 熊兴隆,崔雅峰,马愈昭. 基于消光系数的机场 $PM_{2.5}$ 质量浓度神经网络预测模型[J]. 科学技术与工程,2017,17(32):274-279.

[8] 黄影平. 贝叶斯网络发展及其应用综述[J]. 北京理工大学学报,2013,33(12):1211-1219.

[9] 桑慧茹,王丽学,陈韶明,等. 基于主成分分析的 RBF 神经网络在需水预测中的应用[J]. 水电能源科学. 2017,35(7):58-61.

[10] 卢彬,马行,穆春阳,等. 基于 PCA-BN 的银川市空气质量预测[J]. 安全与环境工程,2020,27(5):70-76.

[11] 张旭. 基于神经网络的空气质量预测[D]. 南京:南京信息工程大学,2019.

[12] 刘可真,阮俊枭,赵现平,等. 基于麻雀搜索优化的 Attention-GRU 短期负荷预测方法[J]. 电力系统及其自动化学报,2022,34(4):99-106.

[13] 马飞燕,李向新. 基于改进麻雀搜索算法-核极限学习机耦合算法的滑坡位移预测模型[J]. 科学技术与工程,2022,22(5):1786-1793.

[14] 许亮,张紫叶,陈曦,等. 基于改进麻雀搜索算法优化 BP 神经网络的气动光学成像偏移预测[J]. 光电子·激光,2021,32(6):653-658.

[15] 常东峰,南新元. 基于混合麻雀算法改进反向传播神经网络的短期光伏功率预测[J]. 现代电力,2022,39(3):287-298.

Secondary Prediction Model of Pollutant Concentration Based on SSA Optimized BP Neural Network

HUANG Bangju, ZHANG Weiliang

(School of Air Traffic Control, Civil Aviation Flight University of China, Guanghan 618307, Sichuan, China)

Abstract: In order to solve the problem that the simulation results of the primary prediction model are not ideal, the principal component analysis (PCA) is used to reduce the dimensions of 14 meteorological factors and extract four comprehensive evaluation indicators. The sparrow search algorithm(SSA) is used to optimize the secondary prediction model of BP neural network and the difference between the results of the primary prediction and the real data is used as the output to train the prediction model and make a prediction method. Apply the model to some regions, and verify the model with corresponding data. The results show that the prediction model based on SSA optimized BP neural network and the new model introduced error have higher accuracy and stronger generalization ability than BP neural network model.

Keywords: pollutant concentration prediction; principal component analysis(PCA); BP neural network; sparrow search algorithm(SSA)