2015年全国大学生统计建模大赛

（第四届）获奖论文选

全国大学生统计建模大赛执行委员会　编

中国统计出版社
China Statistics Press

2015年全国大学生统计建模大赛

（第四届）获奖论文选

全国大学生统计建模大赛执行委员会　编

中国统计出版社
China Statistics Press

图书在版编目(CIP)数据

2015年全国大学生统计建模大赛(第四届)获奖论文选 / 全国大学生统计建模大赛执行委员会编. —— 北京 : 中国统计出版社, 2016.9

ISBN 978-7-5037-7999-2

Ⅰ. ①2… Ⅱ. ①全… Ⅲ. ①统计模型—文集 Ⅳ. ①C8-53

中国版本图书馆CIP数据核字(2016)第225890号

2015年全国大学生统计建模大赛(第四届)获奖论文选

作　　者/全国大学生统计建模大赛执行委员会
责任编辑/张　赏
特约编辑/孙　慧　李　锐
封面设计/李雪燕　黄　晨
出版发行/中国统计出版社
通信地址/北京市丰台区西三环南路甲6号　邮政编码/100073
电　　话/邮购(010)63376909　书店(010)68783171
网　　址/http://www.zgtjcbs.com
印　　刷/三河双峰印刷装订有限公司
经　　销/新华书店
开　　本/710×1000mm　1/16
字　　数/295千字
印　　张/15
版　　别/2016年9月第1版
版　　次/2016年9月第1次印刷
定　　价/38.00元

目录

大数据统计建模类　本科生组

基于 PSO-BP 神经网络预测 广州市日均 PM_{10} 浓度

南方医科大学　林愿仪、林伟俊、尹安琪

摘　要

背景：可吸入颗粒物 PM_{10} 是指悬浮在空气中，空气动力学当量直径小于或等于 10μm 的颗粒物，其浓度的升高会给人群健康造成很大的危害。对 PM_{10} 浓度进行预测预报可以为环境管理决策提供依据，同时有助于市民及时采取相应的防控措施降低污染的影响。

目的：拟用 2008—2011 年广州市前一天的日均 PM_{10} 浓度结合同期的气象等因素，建立 PSO-BP 神经网络模型，对 PM_{10} 浓度进行预测。

方法：将广州市 2008—2011 年的 PM_{10} 浓度和气象资料数据分为训练样本和测试样本。在训练样本中运用基于 BP 的 MIV 神经网络方法筛选出影响 PM_{10} 浓度预测的主要因素，建立随机森林模型和 PSO-BP 神经网络模型对 PM_{10} 浓度进行预测。利用测试样本检验两模型的预测效果，并用 RMSE、MAE、MAPE、PMAD、R^2 等指标对两模型的预测效果进行比较，以说明 PSO-BP 神经网络模型的预测效果。本研究主要采用 SPSS 20.0 统计软件对数据进行描述性分析；采用 R 3.2.0 软件 randomForest 包进行随机森林的回归预测；采用 Matlab 2014a 统计软件进行 MIV 值特征筛选变量和建立 PSO-BP 神经网络模型。

结果：运用基于 BP 的 MIV 神经网络法筛选出前一天的 PM_{10}、平均气压、平均气温、最低气压、水汽压、平均相对湿度、最高气压、最低气温、极大风速、最大风速共 10 个主要影响变量。对 PSO-BP 神经网络模型与随机森林模型测试样本的预测效果进行比较，随机森林模型的 RMSE、MAE、MAPE、PMAD、R^2 分别为 30.022、23.630、0.246、0.246 和 0.635；PSO-BP 神经网络模型的 RMSE、MAE、MAPE、PMAD、R^2 分别为 25.482、21.120、0.225、0.220 和 0.760。可见 PSO-BP 神经网络模型预测效果更优，模型的拟合效果与实际数据的误差更小。

结论：利用 PSO-BP 神经网络模型预测广州市未来一天的日均 PM_{10} 浓度效果较好，与随机森林模型相比其误差更小，预测效果更优，可为环境管理决策提供依据。

关键词：PM_{10} MIV 特征值筛选　随机森林　PSO-BP 神经网络　气象因素　广州市

引　言

(一)研究背景和目的

1. 研究背景

随着工业的发展和城市进程的加剧,我国乃至全球的大气环境污染形势愈趋严峻,人类健康因而受到巨大威胁。世界卫生组织(WHO)最新估计数据显示:每年有700万例的过早死亡与空气污染有关[1],而流行病学研究表明,随着大气中的悬浮颗粒物(Particulate Matter,PM)浓度的升高,人体的呼吸道症状会加剧,因上呼吸道疾病就诊或住院的人数也会增加[2],同时也会引起人体肺功能的降低及心肺疾病死亡率的增加[3,4]。而随着科研工作的深入,人们逐渐认识到直径小于或等于10μm的颗粒物(PM_{10})是导致城市人群患病率和死亡率增加的主要因素[5]。因此,如何及时、准确的预测PM_{10}的浓度,为环境管理决策提供信息成为大家十分关注的问题。

研究表明,大气污染物与特定的气象因素有着密切的关系,气象因素往往制约着大气污染物的稀释、扩散、输送和转化,进而影响大气污染的浓度和分布[6,7]。在不同的气象条件下,同一污染源排放所造成的空气污染物浓度可相差几十倍甚至几百倍[8]。因此,运用气象因素等对PM_{10}浓度进行预测有着重要的意义。

广州市是广东省重要的政治和文化中心,其迅速发展的同时也带来了严重的环境污染问题。2008年广州市人民政府发布了空气污染综合整治实施方案[9],有效的改善了空气质量,但日均PM_{10}浓度还保持在较高的水平,2008－2011年期间有3.765%的日子PM_{10}浓度超标(按我国标准日均浓度150μg/m3计算),其防治问题依然值得探索。

2. 研究目的

近年来,随着计算机和信息技术的快速发展,使许多以前难以预测的空气质量预测逐渐成为可能。我国目前已有空气质量形势预报[10],能够预测未来2－3日的空气质量情况,但其只将我国划分为京津冀、长三角、珠三角等三大区域,范围过大不够精准,且其对空气质量情况只能给出简单评价,未能提供具体大气颗粒物浓度指标。因此,建立一个对特定城市的PM_{10}浓度的预测模型,为环境管理决策提供准确、全面、及时的环境污染水平信息,对环境污染问题的治理和改善提供有力的工具显得尤为重要。广州市是珠江三角洲的重点经济发展城市,环境污染问题尤为严重,如何建立适合广州市的PM_{10}浓度预测模型是我们需要解决的问题。

(二) 研究现况

目前,国内有许多学者致力于研究城市空气污染浓度预测模式。吴嘉荣[11]用线性回归法建立了城市环境空气质量预报模式,表明了前一日PM_{10}浓度和气象

因素与第二日的 PM_{10} 存在相关关系，对 PM_{10} 浓度进行了简单预测，但未进行预测效果评价。李祚泳等率先将神经网络应用于空气污染预测的探索性研究，预测了 SO_2 的浓度，并指出 BP 网络的预测精度优于模糊识别模型的预测精度[12]。周国亮等[13]利用 BP 神经网络对空气质量级别等计数资料作出了预测，准确率较高。石灵芝等[14]对长沙市 PM_{10} 每小时浓度进行预测，尽管检验预测时间较短（2008－01－05 至 2008－01－09 共 5 天），但预测效果较好，整体 R^2 达到 0.62。于宗艳等[15]利用免疫粒子群优化算法得到了空气质量评价模型，但未对空气质量作出预测。

在国外的研究中，Misiti M，等[16]对每日 PM_{10} 浓度建立了混合线性回归进行预测，Thomas S 等[17]考虑了气象因素的滞后效应建立了多元线性模型，并用神经网络较好的预测了 $PM_{2.5}$ 的浓度（$R^2=0.79$）。同样，Ul－Saufie AZ 等[18]利用其他污染物变量和气象变量也分别建立了多元线性回归模型和人工神经网络模型对 PM_{10} 浓度进行较为准确的预测。Jef H 等[19]在建立神经网络模型预测未来一天的 PM_{10} 浓度时加入了新的自变量边界层高度，但发现神经网络未能从中提取到有用的信息，因而并没有提高预测精度。W. Z. LU 等[20]提出了 PSO-BP 模型预测空气质量的可行性，但未进行预测效果评价。

综上所述，现有文献通过对 PM_{10} 浓度与气象因素等的分析，建立多元线性回归模型和神经网络模型等模型对 PM_{10} 浓度进行了不同程度的预测。但存在以下不足之处：一是部分国内研究只限于用当天的气象数据预测当天的颗粒物浓度，而在现实生活中当天的气象数据往往未能提前获取，因而其预测应用的意义不大；二是部分研究只根据经验理论对气象因素进行选择，没有运用科学的方法对气象因素等进行筛选并运用适合当地气象影响因素建立预测模型；另外，不同模型的预测效果如何尚有待比较，尤其是何种机器学习方法更优有必要予以探讨，为该方法的推广应用提供参考。

（三）本文研究思路与创新之处

1. 研究思路

基于以上研究目的与研究现况，本研究提出如下（图 1）研究思路：以 2008－2011 年广州市地区内 9 个空气质量监测站点（天河职幼，市检测站，市 86 中，市 5 中，麓湖，花都师范，广雅中学，广东商学院，番禺中学等）的 PM_{10} 日平均浓度和气象因素等为研究对象，应用 BP 神经网络与粒子群优化算法（PSO）结合的 PSO-BP 神经网络模型预测未来一天的 PM_{10} 日平均浓度，同时与随机森林模型预测结果进行比较，从而说明 PSO-BP 神经网络模型对 PM_{10} 日平均浓度的预测效果，并为预测和控制空气污染提供一系列科学的依据。

2. 创新之处

（1）本研究以 2008－2011 年长时间段的日均 PM_{10} 浓度和气象因素等历史数

图1 本文研究思路

据为基础,运用前一日的PM_{10}浓度和气象数据建立适合广州市的PM_{10}预测模型,对日均PM_{10}浓度进行提前一天的预测。

(2)在BP神经网络的基础上加上PSO算法构建PSO-BP神经网络模型对日均PM_{10}浓度进行预测,更好的减小了预测误差。

(3)气象因素等变量主要运用了科学的基于BP的MIV神经网络变量筛选方法,可以应用于多重共线性的数据,使得筛选出影响PM_{10}浓度预测的主要气象因素等,进而建立随机森林模型和PSO-BP神经网络模型,并进行两模型的比较,更有力的说明重点研究模型PSO-BP神经网络模型的预测效果。

(4)对两种机器学习方法进行比较,从而选出对于广州市PM_{10}浓度预测更优的模型。

一、方法

(一)研究方法的基本原理

1. 基于BP的MIV神经网络变量筛选

与传统的统计模型不一样,神经网络模型本身可以应用于多重共线性的数据,所谓多重共线性是指回归模型中的自变量之间由于存在精确相关关系或高度相关关系而使模型估计失真或难以估计准确。因此,为了网络的训练效果更佳,神经网络的变量筛选方法不应该跟一般的回归模型变量筛选方法一样。

选择神经网络输入的方法有多种,其基本思路是:尽可能将作用效果显著的自变量选入神经网络中,将作用不显著的自变量排除在外。本文将结合BP神经网络应用平均影响值(Mean Impact Value,MIV)算法来筛选变量。MIV被认为是在神经网络中评价变量相关性最好的指标之一。[21]

MIV是用于确定输入神经元对输出神经元影响大小的一个指标,其符号代表相关的方向,绝对值大小代表影响的相对重要性。

具体计算过程:

(1)网络训练终止后,将训练样本P中每一自变量特征在其原值的基础上分别加和减10%构成两个新的训练样本P_1和P_2。

(2)将P_1和P_2分别作为仿真样本利用已建成的网络进行仿真,得到两个仿真结果A_1和A_2。

(3)求出 A_1 和 A_2 的差值,即为变动该自变量后对输出产生的影响变化值(Impact Value,IV)。

(4)最后将 IV 按观测例数平均得出该自变量对应于应变量——网络输出的 MIV。

(5)按上面步骤依次算出各个自变量的 MIV 值,最后根据 MIV 绝对值的大小为各自变量排序,得到各自变量对网络输出影响相对重要性的位次表,从而判断输入特征值对于网络结果的影响程度,即实现了变量筛选。[21]

2. 随机森林

随机森林(Random Forest,RF)算法是一种基于分类和回归树(Classification and Regression Trees,CART)的数据挖掘方法,由 Breiman 和 Cutler 在 2001 年提出的一种较新的机器学习算法。目前随机森林算法主要应用于生态学领域,并且表现出较高的预测精度。[22]

随机森林算法的实质是基于决策树的分类器集成算法,其中每一棵树都依赖于一个随机向量,随机森林的所有向量都是独立同分布的。随机森林就是对数据集的列向量和行观测进行随机化,生成多个分类树,最终将分类树结果进行汇总。

随机森林相对于一般神经网络,降低了运算量的同时也提高了预测精度,而且该算法对多元共线性不敏感以及对缺失数据和非平衡数据比较稳健。与传统的回归模型相比,随机森林算法不需要预先设定函数的具体形式,可以克服自变量之间的交互作用,而且不容易出现过度拟合的现象。[22]

随机森林算法通过自助法(bootstrap) 重抽样技术,由随机向量 θ(即回归树)构成组合模型 $\{h(X,\theta k),k=1,\cdots,p\}$。预测变量为数值型变量,生成的随机森林为多元非线性回归分析模型。随机森林预测的形成是通过求 k 棵树 $\{h(X,\theta k)\}$ 的平均值,形成随机森林的训练集各自独立,选自随机向量 Y,X。数值型预测向量 h(X)的推广误差均方为

$$E_{X,Y}(Y-h(X)) \tag{1}$$

随机森林回归有以下特性:

(一)当森林中树的个数趋于无穷大时,有:

$$E_{X,Y}(Y-\alpha v_k h(X,\theta k))^2 \to E_{X,Y}(Y-E_\theta h(X,\theta))^2 \tag{2}$$

(二)如果对于所有的 θ,$E(Y)=E_X h(X,\theta)$,则:

$$PE^*(forest) \leqslant \bar{\rho} PE^*(tree) \tag{3}$$

其中 $PE^*(tree)=E_\theta E_{X,Y}(Y-h(X,\theta))^2$,$\bar{\rho}$ 为剩余 $Y-h(X,\theta)$ 和 $Y-h(X,\theta)$ 间的权重关系,θ 是独立的。

随机森林回归算法实现流程为:

(1)原始数据样本含量为 n,应用 bootstrap 有放回地随机抽取 b 个自助样本集,并由此构建 b 棵回归树,每次 bootstrap 抽样未被抽到的样本组成了 b 个袋外

数据(out-of-bag,OOB),作为随机森林的测试样本;

(2)设原始数据的变量个数为 p,则在每一棵回归树的每个节点处随机抽取 m_{try} 个变量($m_{try} \ll$p)作为备选分枝变量,然后在其中根据分枝优度准则选取最优分枝。在随机森林回归中,参数 m_{try}=p/3;

(3)每棵回归树开始自顶向下的递归分枝,设定叶节点的最小尺寸 nodesize = 5,以此作为回归树生长的终止条件;

(4)将生成的 b 棵回归树组成随机森林回归模型,回归的效果评价采用袋外数据(OOB)预测的残差均方,见公式(4)和(5)。

$$MSE_{OOB}=n^{-1}\sum_{1}^{n}\{y_i-\hat{y}_i^{OOB}\}^2 \tag{4}$$

$$R_{RF}^2=1-\frac{MSE^{OOB}}{\hat{\sigma}_y^2} \tag{5}$$

其中,y_i 为袋外数据中因变量的实际值,$\hat{y}_i$ 为随机森林对袋外数据的预测值,$\hat{\sigma}_y^2$ 为随机森林对袋外数据预测值的方差。[23]

3. BP 神经网络

线性模型只能解决线性可分问题,而 BP 神经网络属于多层感知器(Multi-layer Perceptrons,MLP)的一种,能够解决预测中的线性不可分问题。多层感知器除了输入层和输出层外,还具有若干隐含层。上下层之间实现全连接,而每层单元之间无连接。大部分情况下多层感知器采用误差反向传播(Back Propagation)的算法进行权值调整,即当一学习样本提供给网络之后,神经元的激活值从输入层经中间层向输出层传播,在输出层的各个神经元获得网络的输入响应。随后,按照减小目标输出与实际误差的方向,从输出层经过中间层逐层修正各层的连接权值,最后回到输入层。具体的方法如下:

(1)输入信息的顺向传播

隐含层中第 i 个神经元的输出为:

$$a_{1i}=f_1\left[\sum_{j=1}^{r}\omega_{1ij}P_j+\theta_{1i}\right],\ \mathrm{i}=1,2,\cdots,\mathrm{S1} \tag{6}$$

输出层中第 k 个神经元的输出为:

$$a_{2k}=f_2\left[\sum_{j=1}^{S1}\omega_{2ki}a_{1i}+\theta_{2k}\right],\ \mathrm{k}=1,2,\cdots,\mathrm{S2} \tag{7}$$

误差函数为:

$$E(\omega,B)=\frac{1}{2}\sum_{j=1}^{S1}(t_k-a_{2k})^2 \tag{8}$$

式中 S1、S2 分别为隐含层、输出层的神经元个数,k 为迭代次数[24]。

(2) 误差函数的反向传播

输出层的权值变化,对从第 i 个输入到第 k 个输出的权值有:

$$\sum \omega_{2ki} = -\eta \frac{\partial E}{\partial \Delta \omega_{2ki}} = -\eta \frac{\partial E}{\partial a_{2k}} \frac{\partial a_{2k}}{\partial \Delta \omega_{2ki}} = \eta (t_k - a_{2k}) f'_2 a_{1i} = \eta \delta_{ki} a_{1i} \quad (9)$$

式中：$\delta_{ki} = (t_k - a_{2k}) f'_2 = e_k f'_2$，$e_k = t_k - a_{2k}$；

同理可得：

$$\Delta b_{2ki} = -\frac{\partial E}{\partial \Delta \theta_{2ki}} = -\eta \frac{\partial E}{\partial a_{2k}} \frac{\partial a_{2k}}{\partial \Delta \theta_{2ki}} = \eta (t_k - a_{2k}) f'_2 = \eta \delta_{ki} \quad (10)$$

隐含层的权值变化，对从第 j 个输入到第 i 个输出的权值有：

$$\begin{aligned} \Delta \omega_{1ij} &= -\eta \frac{\partial E}{\partial \Delta \omega_{1ij}} = -\eta \frac{\partial E}{\partial a_{2k}} \frac{\partial a_{2k}}{\partial a_{1i}} \frac{\partial a_{1i}}{\partial \Delta \omega_{1ij}} \\ &= -\eta \sum (t_k - a_{2k}) f'_2 \omega_{2ki} f'_1 p_j = \eta \delta_{ki} P_j \end{aligned} \quad (11)$$

同理可得：

$$\Delta \theta_{1i} = \eta \delta_{ij} \quad (12)$$

式中负号表示梯度下降，常数 $\eta (0 < \eta < 1)$ 表示比例系数，即学习率[25,26]。

(3)模型表达式

基于本研究，单个输出节点(PM_{10})的 BP 神经网络模型的构建可简化为下式表示：

$$\hat{y}(I) = A_2 \left(\sum_{i=1}^{N_i} \omega_m^2 \cdot \left(A_1 \left(\sum_{m=1}^{N_m} \omega_{im}^1 \cdot x_i(I) + b_m^1 \right) \right) + b_o^2 \right) \quad (13)$$

其中 A_1 和 A_2 分别为隐含层和输出层的传递函数；ω_{im}^1 和 ω_m^2 分别表示为输入层 i 个节点到隐含层 m 个节点的权重和隐含层 m 个节点到单个输出节点的权重；b_{1m} 和 b_o^2 分别表示第 m 个隐含层节点偏倚和输出层的偏倚；N_m 和 N_i 分别表示输入层和隐含层节点数。

建立 BP 神经网络的参数选择具体分为 6 个：

(1)网络层数

BP 网络可以包含一到多个隐含层。不过，理论上证明单个隐含层网络可以通过适当增加神经元节点的个数实现任意非线性映射。

(2)隐含层节点数

目前并没有一个理想的解析式可以用来确定合理的神经元个数。通常做法是采用经验公式给出估计值：

$$\sum_{i=0}^{n} C_M^i > k \quad (14)$$

$$M = \sqrt{n+m} + a \quad (15)$$

$$M = \log_2 n \quad (16)$$

其中，k 为样本数，M 为隐含层神经元个数，n 为输入层神经元个数，m 为输出层神经元个数，a 是[0,10]之间的常数。若 i>M，规定 $C_M^i = 0$。

(3)BP学习时权值的初始值确定

初始值过大过小都会影响学习速度,经验值为$(-2.4/F,2.4/F)$或$(-3/\sqrt{F},3/\sqrt{F})$之间,其中F为权值输入端神经元个数。另外,为避免每一步权值的调整方向是同向的,应将初始权值设为随机数,本文也取初始权值和阈值为[0,1]之间的随机数[27]。

(4)传递函数的选择

传递函数必须可微。一般隐含层使用Sigmoid函数,而输出层为线性函数。如果输出层也采用Sigmoid函数,则输出值将会被限制在(0,1)或(-1,1)之间。Sigmoid函数又可分为Log-Sigmoid函数和Tan-Sigmoid函数。

(5)学习速度的选定

学习速度不能选太大,否则算法不收敛。也不能太小,会使训练时间太长。一般选择0.01~0.1之间的值。

(6)训练方法的选择

BP修正权值的方式有两种:串行方式和批量方式。在串行方式中,每一个输入被作用于网络后,权重和阈值被更新一次。在批量方式中,所有的输入被应用于网络后,权重和阈值才被更新一次。使用批量方式不需要为每一层的权重和阈值设定训练函数,而只需为整个网络指定一个训练函数,使用起来相对方便,而且许多改进的快速训练算法只能采用批量方式,在这里我们只用批量方式。

4. 粒子群算法(PSO)

粒子群算法,也称粒子群优化算法(PSO),是近年来发展起来的一种新的进化算法[28]。其源于生物社会学家对鸟群、鱼群或者昆虫捕食行为的研究,是一种实现简单、全局搜索能力强且性能优越的启发式搜索技术。鸟类捕食时,每只鸟找到食物最简单有效的方法就是搜索当前距离最近食物的鸟的周围区域,可视鸟群为粒子群,将食物视为全局最优解,将鸟群捕获食物的过程等价于粒子群寻找全局最优解的过程[29]。

在PSO算法中,每粒子都代表极值优化问题的一个潜在最优解,用位置、速度和适应度值三项指标表示该粒子的特征,适应度值由适应度函数计算得到,其值的好坏表示粒子的优劣。粒子在解空间中运动,通过跟踪个体极值Pbest和群体极值Gbest更新个体位置,个体极值Pbest是指个体所经历位置中计算得到的适应度值最优位置,群体极值是指种群中的所有粒子搜索到的适应度最优位置。粒子每更新一次位置,就计算一次适应度值,并且通过比较新粒子的适应度值和个体极值,群体极值的适应度值更新个体极值Pbest和群体极值Gbest。

在D维搜索空间中,有m个粒子,其中第i个粒子的位置是$\vec{x}_i=(x_{i1},x_{i2},\cdots,x_{iD})$,也代表问题的一个潜在解$i=1,2,\cdots\cdots,m$,其速度为$\vec{v}_i=(v_{i1},v_{i2},\cdots,v_{iD})$。将$\vec{x}_i$带入目标函数可计算出其适应值。记第$i$个例子搜索到的最优位置为

$\vec{p}_i=(p_{i1},p_{i2},\cdots,p_{iD})$，整个粒子群搜索到的最优位置为 $\vec{p}_g=(p_{g1},p_{g2},\cdots,p_{gD})$。粒子状态更新操作如下：

$$v_{id}^{k+1}=wv_{id}^{k}+c_1r_1(p_{id}^{k}-x_{id}^{k})+c_2r_2(p_{gd}^{k}-x_{id}^{k}) \tag{17}$$

$$x_{id}^{k+1}=x_{id}^{k}+v_{id}^{k+1} \tag{18}$$

其中，i=1，……，m，d=1，……，D；w 是非负常数，称为惯性因子。w 也可以随着迭代线性地减小；学习因子 c_1 和 c_2 是非负常数；r_1 和 r_2 是介于[0，1]之间的随机数；$v_{id}\in[-v_{max},v_{max}]$，$v_{max}$ 是常数。

迭代中止条件一般选为最大迭代次数和粒子群，迄今为止搜索到的最优位置满足适应阈值[30]。

5. PSO-BP 神经网络模型

由于 BP 算法收敛速度慢而且极易陷入局部最优，在应用中网络结构的确定基本依赖经验，主要是采用递增或递减的试探方法来确定的网络隐节点，这些缺陷使得神经网络的训练样本和测试样本的输出具有不一致性和不可预测性，极大的限制了神经网络在实际预报中的应用[31]。

为了避免 BP 神经网络陷入局部极小值和增加其泛化性能，提供预测精度，采用 PSO 算法优化 BP 神经网络的权值和阈值。PSO 的适应度函数为神经网络的输出误差，公式为：

$$f_i=\frac{1}{n_i}\sum_{q=1}^{n_i}(O_{iq}-T_{iq})^2 \tag{19}$$

其中，n_i 为训练样本的个数，Q_{iq}、T_{iq} 分别为训练样本 q 在第 i 粒子的位置所确定的网络权值和阈值下的网络实际输出和期望输出[32]。

PSO-BP 神经网络的具体流程如图 2 所示：

由图 2 可知，PSO-BP 神经网络算法的具体步骤为：

(1)初始化 BP 神经网络和粒子群

根据样本数据设计 BP 网络的输入、输出和隐含层神经元数目、学习函数及训练函数；根据粒子群的规模，按照个体结构产生一定数目的粒子群，其中不同的个体代表神经网络的 1 组不同的权值。同时，初始化粒子的速度、位置、个体历史最优 p_i、全局最优 p_g、迭代误差精度和最大迭代次数等[28]。

(2)迭代与更新

利用式(17)(18)更新粒子的速度和位置，并用式(19)计算粒子的适应值。判断当前迭代次数是否大于最大迭代次数或当前最优适应值小于设定精度，若是满足条件，则输出全局最优粒子位置及 BP 网络的权值和阈值。

(3)训练 BP 网络

根据输出的 BP 网络权值和阈值训练 BP 神经网络，并运用测试样本对其进行检验，PSO-BP 神经网络完成。

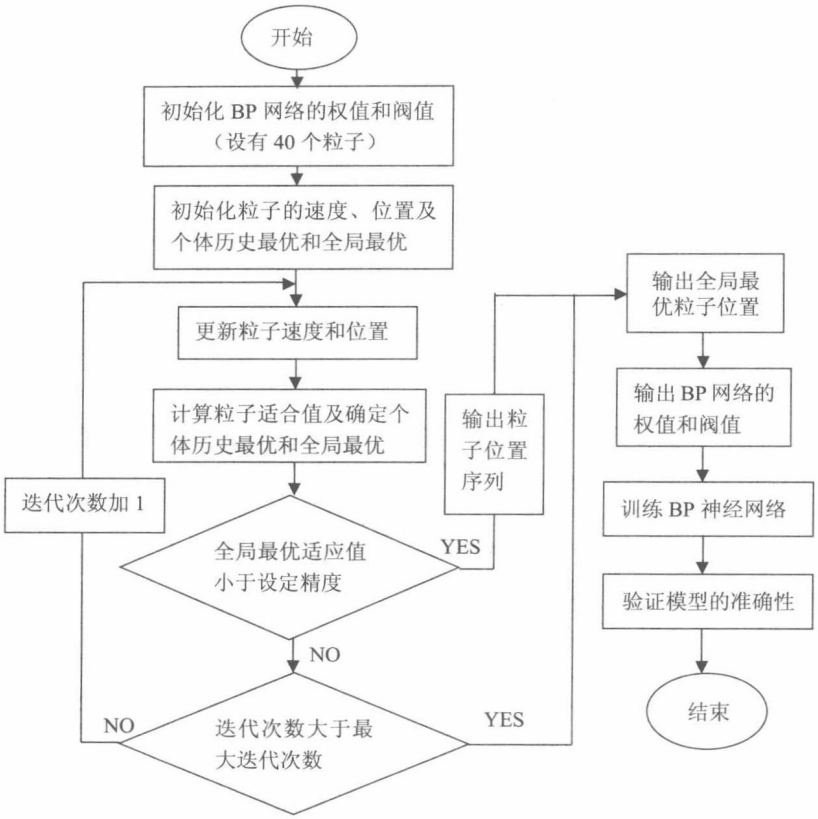

图 2　PSO-BP 神经网络流程图

(二) 统计分析软件

本研究主要采用 SPSS 20.0 统计软件对数据进行描述性分析；采用 R 3.2.0 软件 RandomForest 包进行随机森林的回归预测；采用 Matlab 2014a 软件进行 MIV 值特征筛选变量和建立 PSO-BP 神经网络模型。

(三) 数据的来源

1. 地面 PM_{10} 数据：从广州市环境保护局官网[33]获得 2008－2011 年广州市 9 个监测站点(天河职幼，市检测站，市 86 中，市 5 中，麓湖，花都师范，广雅中学，广东商学院，番禺中学等)的日均 PM_{10} 浓度数据。9 个站点的具体分布如图 3 和表 1 所示。

2. 气象数据和能见度数据：从中国气象科学数据共享服务网获得广州市 2008－2011 年日均降水量、风速、风向、气压、气温、水汽压、相对湿度、日照时数等气象因素数据；从 Weather underground 网站获得 2008－2011 年能见度数据[34,35]。

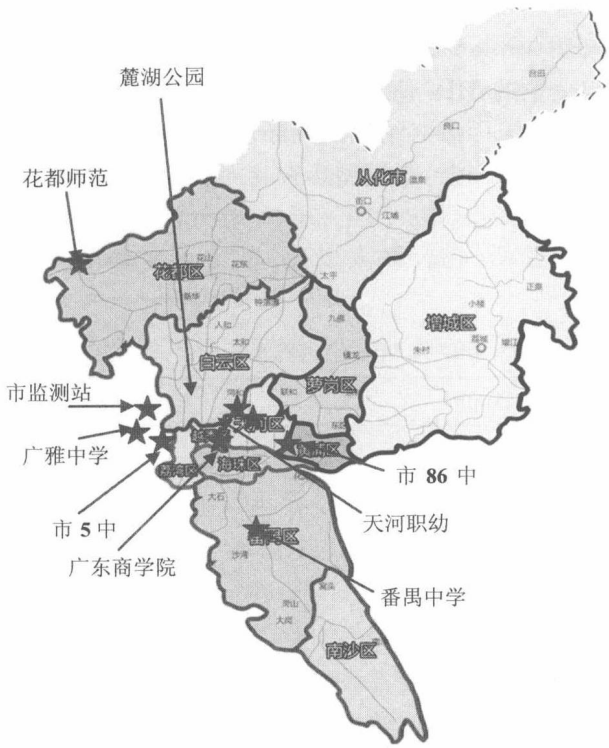

图 3　广州市空气监测点位图

表 1　广州市空气监测点位

测点名称	经纬度	所属行政区
广雅中学	E：113° 14'01" N：23° 08'31"	荔湾区
市 5 中	E：113° 15'35" N：23° 06'15"	海珠区
市监测站	E：113° 15'35" N：23° 07'59"	越秀区
天河职幼	E：113° 19'02" N：23° 08'09"	天河区
麓湖	E：113° 16'50" N：23° 09'25"	天河区
广东商学院	E：113° 21'12" N：23° 05'31"	海珠区
市 86 中	E：113° 25'54" N：23° 06'18"	黄埔区
番禺中学	E：113° 21'14" N：22° 57'05"	番禺区
花都师范	E：113° 12'40" N：23° 23'30"	花都区

（四）数据的处理与变量的选择

1. 数据集的划分

在本研究中，我们建立了一个训练样本（包含 1430 行数据）用来建立模型以及一个测试样本（包含 31 行数据）用来检验模型的预测效果，具体分配如下：

（1）训练样本：训练样本的数据从 2008 年 1 月 1 日－2011 年 11 月 30 日，共 1430 行数据，其中自变量数据从 2008 年 1 月 1 日－2011 年 11 月 29 日，预测的 PM_{10} 数据从 2008 年 1 月 2 日－2011 年 11 月 30 日。

（2）测试样本：测试样本的数据从 2011 年 12 月 1 日－2011 年 12 月 31 日，共 31 行数据，其中自变量数据从 2011 年 11 月 30 日－2011 年 12 月 30 日，预测的 PM_{10} 数据从 2011 年 12 月 1 日－2011 年 12 月 31 日。

2. 数据预处理

本次研究意在建立适合广州市的 PM_{10} 预测模型，因而我们对 9 个站点的日均 PM_{10} 浓度数据求平均值来代表广州市中心城区的 PM_{10} 污染水平。同时，本文主要考虑前一天的 PM_{10} 浓度和气象因素与预测的 PM_{10} 浓度的相关关系。

3. 变量的选择

（1）因变量：预测的日均 PM_{10} 浓度；