DOI:10.13878/j.cnki.jnuist.2021.04.016

殷利平[1,2]　刘宵瑜[1,2]　盛绍学[3]　温华洋[3]　邱康俊[3]

基于 SVM-BP 神经网络的气象能见度数据缺失值预估

摘要

自动气象站能见度检测仪多采用光学装置采样,雨雪、粉尘等天气因素会对部分仪器镜头造成污染,导致能见度要素数据缺测.针对能见度数据缺失问题,本文选用安徽部分气象站的历年数据,首先运用灰色关联分析方法筛选出与能见度密切相关的其他气象要素,通过支持向量机和 BP 神经网络单一预估方法预估不同地形的能见度缺失值,然后采用最优权重组合将两种方法预估的能见度值进行组合,并与单一预估方法进行对比.结果表明组合方法的预估结果误差均值小、整体准确度高,可以保证台站观测资料的完备性,为短时天气预报、实况分析和气象公共服务工作提供有效依据.

关键词

组合模型;缺失值预估;关联分析;BP 神经网络;能见度;支持向量机

中图分类号 P457.7

文献标志码 A

收稿日期 2020-10-12

资助项目 国家自然科学基金(61573190,61571014);安徽省气象局科研项目(KM201907);安徽省创新团队建设计划

作者简介

殷利平,女,博士,副教授,研究方向为随机控制.lpyin@ nuist.edu.cn

1 南京信息工程大学 自动化学院,南京,210044
2 南京信息工程大学 大气环境与装备技术协同创新中心,南京,210044
3 安徽省气象信息中心,合肥,230031

0 引言

为了将气象研究对社会的积极作用融入到公共服务中,中国气象局于 2002 年投资建设“三站四网”的大气监测工程,在全国各地陆续建立自动气象站.这些自动气象站引进许多高精度的气象观测设备,大大提高了对气象要素进行实时探测的能力[1-2].安徽省大多数气象站采用散射式能见度仪采样,但是在日常工作中能见度仪会出现采样数据缺失的情况,一般由以下几种情况造成:1)能见度仪的镜头前或两个镜头之间有异物堵塞,如蜘蛛结网、小鸟做窝等;2)在一些施工区,或省道县道等公路旁,灰沙和扬尘可能导致能见度采样区内颗粒物变化不定;3)恶劣天气下,局部地区的风速、风向变化大且快,导致树叶、杂物被吹起恰好位于能见度仪的采样区内,雨雪天气和天气寒冷凝结的冰霜也可能使能见度仪镜头表面受污染严重,导致能见度数据不准确;4)传感器各接线端出现接触不良、松动,以及传感器的某一单元模块发生故障、仪器年久失修得不到有效的维护等情况[3-4].有些自动气象站建立在高山丘陵地带,人工维护难度大、成本高,迫切需要一种既可以及时得到气象站所测的完备气象信息,又可以减轻工作人员对问题气象站进行维护的工作量的方法.

目前,处理能见度仪数据缺失的方法主要可以分为基于统计的修补算法、基于邻近性的修补算法、基于机器学习的修补算法三大类.基于统计的修补算法包括均值插补[5]、回归插补[6]、多重插补[7]等,其中均值插补以数据序列的平均值作为填充缺失值;回归插补是把缺失属性作为因变量,其他相关属性作为自变量,利用它们之间的关系建立回归模型来预测缺失值的;多重插补是用一组近似值替换每个缺失值,再用标准的统计分析过程对多次替换后产生的若干数据进行分析、比较,从而得到缺失值的估计值.基于统计的插补方法虽然简单易操作,但容易扭曲数据分布,且该类算法需要预先知道数据分布特征,但很多实际应用场景中却无法得到.基于邻近性的修补算法中最具有代表性的是 K 近邻算法(K-Nearest Neighbor,KNN)[8-9].K 近邻算法首先要找出数据集中与缺失数据的欧式距离最小的 K 个点,然后用这 K 个点的平均值替换缺失值,其修补效果易受到邻近阈值的影响,且容易受到噪声数据的干扰,若对数据集未做初步预处理,修补精度容易产生较大的偏差.基于机器学习的修补算法能够直接处理缺失数据,并对缺失数据集进行训练,该类方法的优点是可以直接

处理完全随机缺失模式下的数据集.该类算法主要包括:集成方法(以神经网络集成方法为主)[10]、多层感知机插补[11]、决策树、贝叶斯[12-13]、支持向量机(Support Vector Machine,SVM)[14-15]等,其中集成方法修补缺失数据以 BP 神经网络应用最为广泛.BP 神经网络是指利用误差逆传播(Error Back Propagation)算法训练的多层网络,BP 算法是将误差反向传播使神经元各层权值不断调整,直到网络输出的误差减少到可接受的程度,其优点是在处理不完整大规模数据时速度快、泛化能力强.SVM 也是一种通用的机器学习算法,它以统计学习理论为基础,广泛应用在函数回归、时间序列预测等领域[16-17].SVM 算法首先是通过非线性映射函数把样本向量映射到高维特征空间,使得在特征空间中,原空间数据的像具有线性关系,然后在特征空间中构造线性最优决策函数,从而解决分类与回归问题.在处理缺失值回归分析时,SVM 算法可以修补任意缺失模式的数据,减少计算复杂度[18].

本文利用机器学习的相关算法在缺失值插补方面的优异性,综合运用 SVM 和 BP 神经网络预估能见度缺失值.首先选用安徽气象局历年来不同地区气象站的历史数据进行分析,然后建立数据填充模型,再运用权重优化不同模型对缺失值的预估值.实验结果表明,运用组合模型对不同地形的能见度缺失数据进行预估,预估结果可以有效地代替真实值,实现了对自动气象站的缺失数据的高精度填补.

1 数据的来源和模型构建

1.1 数据处理

本文中气象数据全部来自安徽气象局历年来汇总的气象站观测资料.如图 1 所示,安徽地形复杂多样,不同地形气候不一,因此所得到的观测数据差值较大.考虑到地形因素对模型处理缺测数据的影响,本文以高山、山谷、平原、水源地地形作为特征,分别选取黄山站(高山地形)、山南溪谷站(山谷地形)、灵璧站(平原地形)和白泽湖站(水源地地形)的历史数据进行试验(图 2).早期的自动气象站由于设备质量参差不齐,传感器检测精度低,得到的数据不完整.为了保证数据的有效性,本文选取从 2015—2019 年安徽省气象局记录较为完整的小时时序数据资料作为总样本.对于每种地形,将相应的样本分为 10 个样本集,其中 7 个样本作为训练集,3 个样本集作为验证集.测试集选取各站点 2019 年春季 3—4 月小时观测资料,一些时间区间内如果能见度数值变化较小,选取的数据量也相应减少.

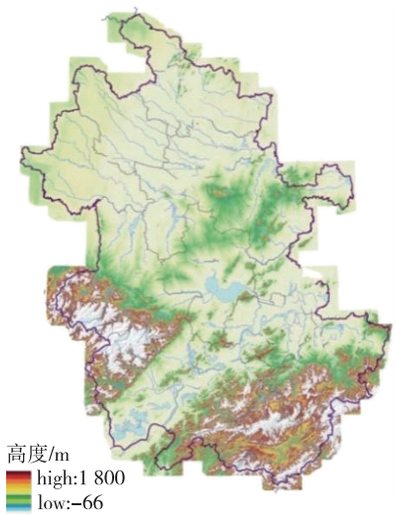

图 1 安徽省地形图

Fig. 1 Topographic map of Anhui Province

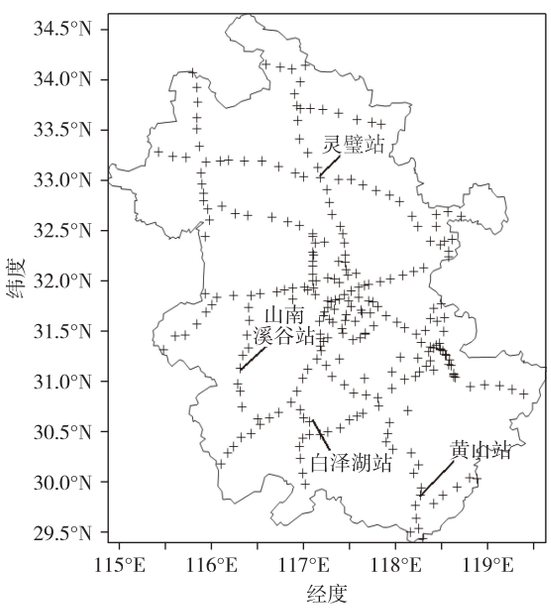

图 2 安徽省各区域小型气象站分布

Fig. 2 Distribution map of meteorological stations in Anhui Province

1.2 输入要素的选择

气象观测要素很多,有些气象观测要素对能见度的影响很小,如果将一个时序的全部观测要素作为输入不仅计算量大,而且会影响预估结果的准确性.人为筛选输入要素的方法具有很大的主观性,缺乏理论依据,因此本文选用灰色关联分析法进行输入变量的选择[19-20].该方法是根据各因素之间数值

变化趋势的程度来确定关联大小,这种方法对数据要求较低,步骤清晰且计算量小.灰色关联法中一个重要指标是灰色关联度,灰色关联度以数值的形式表征各变量间关系的强弱.本文对气象各要素之间的灰色关联分析步骤如下:

1)为研究能见度要素与其他气象要素之间的关系,先对气象要素进行编号,记为 $A_i(i=1,2,3,4,\cdots,11)$,并将各气象要素数据换算成标准单位制数值,A_i 与各气象要素之间的对应关系如表 1 所示.

表 1　各气象观测要素的编号序列

Table 1　Number sequence of meteorological observation elements

编号	气象要素	单位	编号	气象要素	单位
A_0	能见度	m	A_6	风速	m/s
A_1	气温	℃	A_7	风向	(°)
A_2	地表气温	℃	A_8	日照时数	h
A_3	5 cm 地温	℃	A_9	降水量	mm/h
A_4	气压	Pa	A_{10}	蒸发量	mm/h
A_5	相对湿度	%	A_{11}	$\rho(CO_2)$	mg/m^3

2)求表征关联度的关联系数.以能见度数据序列为参考数据,其他观测要素的数据序列作为比较数据.参考数列为 $A_0=\{A_0(1),\cdots,A_0(d),\cdots,A_0(N)\}$,比较数列为 $A_i=\{A_i(1),\cdots,A_i(d),\cdots,A_i(N)\}$,其中 d 代表各要素序列中的元素个数,N 是选取数据序列的总数,$1\leqslant d\leqslant N$.

数据序列 A_i 与 A_0 在第 d 点的关联系数 $\varepsilon_i(d)$ 为

$$\varepsilon_i(d)=\frac{\min\limits_i\min\limits_d|A_0(d)-A_i(d)|+\rho\max\limits_i\max\limits_d|A_i(d)-A_i(d)|}{|A_0(d)-A_i(d)|+\rho\max\limits_i\max\limits_d|A_0(d)-A_i(d)|},\tag{1}$$

式中:$\rho\in(0,+\infty)$ 称为分辨系数,通常在0到1之间选取,一般取 $\rho=0.5$;i 代表气象要素序号,$1\leqslant i\leqslant 11$.

3)求各气象要素之间的关联度 $\gamma(A_0,A_i)$:

$$\gamma(A_0,A_i)=\frac{1}{N}\sum_{d=1}^{N}\varepsilon_i(d),\tag{2}$$

其中关联度 $\gamma\in(0,1)$,数值越大表明该气象要素与能见度的关联度越高.本文按照关联度数值从大到小的顺序选择输入要素,不妨设选择的输入要素为 M 个.

1.3　能见度数据预估模型

1.3.1　SVM 能见度缺失值预估模型

SVM 是把线性不可分的样本通过核函数映射到特征空间,进而在特征空间中构造最优分类平面,使样本到平面的总距离最小,由此实现拟合的[21].对于模型给定的训练数据总样本 $D=\{(x_i(j),y(j)),i=1,2,\cdots,M,j=1,2,\cdots,N_1\}$,其中 $x_i(j)$ 为第 i 个气象输入要素的第 j 个样本,$y(j)$ 为对应的能见度实测值,N_1 为总样本容量.记 $\boldsymbol{x}=[x_1,x_2,\cdots,x_M]^{\mathrm{T}}$.首先利用一个非线性映射函数 $\varphi(\boldsymbol{x})$ 将样本 $\boldsymbol{x}$ 从原空间 $\mathbf{R}^M$ 映射到特征空间,然后在高维特征空间中构造最优决策函数:

$$y(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\cdot\varphi(\boldsymbol{x})+b,\tag{3}$$

式中:$\varphi(\boldsymbol{x})$ 为映射函数;$\boldsymbol{w}$ 为权向量;b 为偏置量.权向量 $\boldsymbol{w}$ 与 b 通过优化下式得到:

$$\min_{w,b,\xi}\phi(\boldsymbol{w},\xi)=\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+\frac{1}{2}C\sum_{j=1}^{N_1}\xi^2(j).\tag{4}$$

其约束条件为

$$y(j)-\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}(j))+b+\xi(j)=0,\tag{5}$$

式(4)中:C 为惩罚因子,为给定值,其数值越大表示对训练误差大于设定误差的样本惩罚越大;$\xi(j)$ 为松弛变量,定义为 $\xi(j)=1-y(j)$,$\xi(j)$ 数值越大表示对样本训练误差的容忍程度越大.

在求解最小化问题(4)和(5)之前,首先要找到合适的非线性函数 $\varphi(\boldsymbol{x})$,为此引入径向基核函数:

$$K(\boldsymbol{x},\boldsymbol{x}(j))=\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}(j)\|^2}{2\delta^2}\right),\delta>0,\quad j=1,2,\cdots,N_1,\tag{6}$$

并令 $K(\boldsymbol{x}(\beta),\boldsymbol{x}(j))=\varphi^{\mathrm{T}}(\boldsymbol{x}(\beta))\cdot\varphi(\boldsymbol{x}(j))$,$\beta=1,2,\cdots,N_1$.进一步引入 Lagrange 方程,从而可以求解(4)和(5),得出 SVM 最优决策函数的估计函数为

$$\hat{y}(\boldsymbol{x})=\sum_{j=1}^{N_1}\alpha(j)K(\boldsymbol{x},\boldsymbol{x}(j))+b,\tag{7}$$

式中 $K(\boldsymbol{x},\boldsymbol{x}(j))$ 为核函数,拉格朗日乘子 $\alpha(j)\in\mathbf{R}$,$\mathbf{R}$ 为实数集.

本文根据以上原理,通过能见度与其他气象观测要素之间复杂的非线性关系进行能见度缺失值预估.具体步骤如下:

1)对各要素的样本数据进行归一化处理.本文采用线性函数转换法:$x_i'(j)=\dfrac{x_i(j)-\min(x_i)}{\max(x_i)-\min(x_i)}$,其中 $x_i(j)$,$x_i'(j)$ 分别为样本各要素序列转换前、后的值,$\max(x_i)$,$\min(x_i)$ 分别为样本序列中各观测要素的最大值和最小值,这样就可以将数据集归一化到[0,1].

2)运用网络搜索法来分别对式(4)和式(6)中

的 C,δ 两个参数寻优,其中惩罚因子 C 的搜寻范围在 0.1 ~ 100,核参数 δ 的搜索范围在 0.001 ~ 1,利用交叉验证法可获得最优参数[22].

3) 利用建立的 SVM 能见度数据预估模型,对预处理后的样本数据进行训练,并对模型的预估结果进行评价.选用平均相对误差(MAPE,其量值记为 η_{MAPE})和均方根误差(RMSE,其量值记为 η_{RMSE})来评价:

$$\eta_{\mathrm{MAPE}} = \frac{1}{N_1}\sum_{j=1}^{N_1}\left|\frac{\hat{y}(j) - y(j)}{y(j)}\right| \times 100\%, \tag{8}$$

$$\eta_{\mathrm{RMSE}} = \sqrt{\frac{1}{N_1}\sum_{j=1}^{N_1}(\hat{y}(j) - y(j))^2}, \tag{9}$$

式中:N_1 为总样本容量;$\hat{y}(j)$ 为第 j 个时间点的预估能见度值;$y(j)$ 为第 j 个时间点实测的能见度值.

1.3.2 BP 神经网络的能见度缺失值预估模型

BP 神经网络能见度预估的基本结构如图 3 所示,其中输入层有 m 个节点,隐含层有 p 个节点,输出层有 1 个节点,$W_{ig}(i = 1,2,\cdots,m;g = 1,2,\cdots,p)$ 为输入层到隐含层的权值,$W_{gk}(g = 1,2,\cdots,p;k = 1)$ 为隐含层到输出层的权值,$\theta_g(g = 1,2,\cdots,p)$ 为隐含层的阈值,σ_1 为输出层阈值,$(X_1,X_2,\cdots,X_m)$ 为神经网络输入向量,Y_1 为神经网络输出量,Y_h 为期望输出,e 为神经网络期望输出与实际输出的误差.

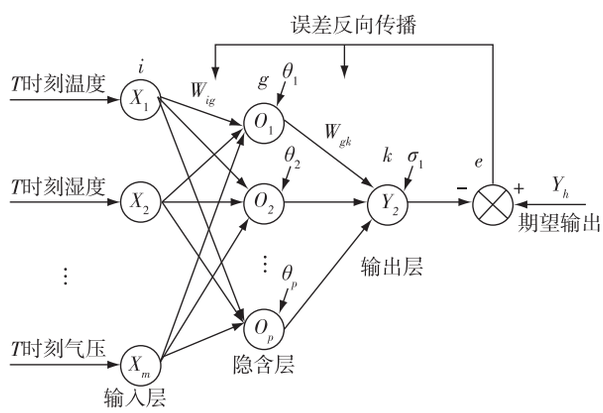

图 3 BP 神经网络能见度预估的结构

Fig. 3 Structure of BP neural network for visibility estimation

三层结构 BP 神经网络可用于预估气象站能见度缺失值,其中输入层对应与能见度相关性大的气象要素序列,输出层是能见度预估值.隐含层的神经元数量对模型预估结果的好坏产生直接的影响,但是目前没有能直接确认最优隐含层个数的方法,只有根据以下经验公式来计算:

$$K = \sqrt{n + m} + a. \tag{10}$$

设定不同隐含层神经元个数,然后采用“试凑法”逐步增大和减少隐含层神经元数目使网络误差最小.式(10)中:m 为输入层节点个数;n 为输出层节点个数;a 为常数,取值范围一般为 3 至 10.K 为隐含层神经元估算个数.

用于能见度缺失值预估的 BP 算法各步骤如下:

1) 权值初始化:$(w_{ig} \cup w_{gk}) = \mathrm{random}(\cdot)$,其中 $\mathrm{random}(\cdot)$ 表示权值在[0,1]之间的均匀分布.

2) 依次输入训练集中的样本,设当前输入第 q 个样本.

3) 依次计算各层的输出:X'_g,X''_k 及 Y_1.

4) 求各层的反传误差,并记下各个 $X''_k(q)$,$X'_g(q)$,$X_i(q)$ 的值.

5) 记录已学习过的样本个数 q.如果 $q < N_1$(N_1 为训练样本总量),继续步骤 2);如果 $q = N_1$,按权值修正公式修正各层的权值或阈值.

6) 按新的权值再计算 X'_g,X''_k 及 Y_1 和学习样本数的总误差 E,若 $E < \varepsilon$(ε 为预估给定误差),或达到最大学习次数,则终止学习.否则,转步骤 2) 继续新一轮学习.

1.3.3 组合模型

用不同的机器学习算法得出的能见度预估值与实测值都有误差.为了减小预估值与实测值之间的误差,可以整合不同模型的优点,对不同方法的预估结果进行加权组合,以提高预估精度.在组合模型预估中最关键的步骤是确定不同预估方法的权重.目前,针对多模型组合权重确定,常用的方法主要有以下几种:算术平均法、方差倒数法、均方倒数法以及最小二乘法.本文采用方差倒数法判断单项模型系数,即对误差平方和小的模型赋予较高的权重,误差平方和大的赋予较小的权重,使组合模型的误差和尽可能小.具体方法如下:

设 F 为观测对象,其实际观测值向量为 $(F_1,F_2,\cdots,F_n)$,$U_1,U_2,\cdots,U_r$ 为 r 种不同预估方法得出的预估值,向量 $\boldsymbol{S} = (S_1,S_2,\cdots,S_r)^{\mathrm{T}}$ 中元素分别是它们在组合模型中的权重,第 l 个预估方法 U_l 的预估值为 $(U_{1l},U_{2l},\cdots,U_{nl})$.则组合模型的估计值为

$$\bar{F} = \sum_{l=1}^{r} S_l U_l = S_1U_1 + S_2U_2 + \cdots + S_rU_r, \tag{11}$$

其中

$$S_l = D_l^{-1} / \sum_{l=1}^{n} D_l^{-1}, \quad \sum_{l=1}^{r} S_l = 1, \tag{12}$$

式(12)中,D_l 为第 l 个模型预估误差的平方和,D_l =

$\sum_{i=1}^{n}(F_i - U_{il})^2$.

2 实验与分析

2.1 能见度影响因子的选择

一般关联度大于等于 0.8 时,子序列与母序列关联度很好.根据 1.2 节的理论,可以计算得到其他观测要素与能见度之间的关联度,本次实验选择与能见度的关联度在 0.8 以上的气象观测要素,如表 2 所示.

表 2 部分气象要素与能见度之间的关联度

Table 2 Correlation degree between meteorological elements and visibility

	能见度
相对湿度	0.956
平均气温	0.942
气压	0.921
地表地温	0.915
风速	0.851
降水量	0.842
5 cm 地温	0.833
日照时数	0.818

表 3 SVM 最优参数值设置

Table 3 Optimal values for SVM parameters

SVM 参数值	参数性质	设定值
s	最大迭代次数	10
C	惩罚因子	2.828
δ	核函数参数	0.088
ε	目标误差	0.001

2.2 SVM 与 BP 方法组合模型预估

本文采用的 SVM 模型预估能见度实验,借助的是 Pycharm 软件的 libsvm 工具箱,其实验精度主要取决于参数选取是否合适,本文各参数设定值如表 3 所示.

在 BP 预估实验中,BP 神经网络模型以与能见度要素关联度高的 8 个气象要素作为输入,隐含层选用单层结构,依据估算最优隐含层神经元个数的经验公式,推算出 K 值在[6,13]之间.为了保证隐含层神经元个数对模型预估结果的准确性,设定隐含层神经元个数 K 在[5,20]区间.将平原组训练集归一化处理后输入 BP 模型中,取不同隐含层神经元个数,用一组验证集记录相对误差均值.由于初始权值随机分配,相同个数的隐含层神经元运行的结果也有不同,所以 BP 网络中每个 K 值的设定都运行 10 次,误差结果算平均值,寻优过程如图 4 所示.

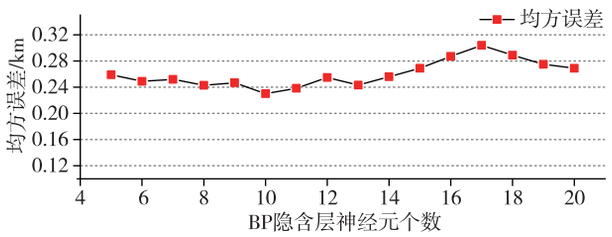

图 4 隐含层神经元个数寻优

Fig. 4 Optimization of the number of neurons in hidden layer

由隐含层神经元寻优结果可知,BP 神经网络预估模型选用 10 个隐含层神经元最佳.其中 BP 神经网络的训练最大迭代次数设定为 50 000 次,学习率取 0.1,迭代循环次数上限值取 20,训练最终误差设定为 0.001,激活函数选择双曲正切函数,训练函数及学习函数均采用 Levenberg-Marquardt 算法.

对不同地形代表站的测试集数据分别采用已训练好的 SVM 和 BP 模型进行预估,测试集预估结果和实测结果进行对比分析,求出各地形预估值与实测值的误差平方和,运用方差倒数法,得各自的权重系数如表 4 所示.

表 4 两种方法单一预估结果

Table 4 Performance of SVM and BP neural network for visibility estimation

预估方法	地形特征	RMSE/km	MAPE/%	最大误差/km	权重系数
SVM	高山	0.274	14.05	3.159	0.576 4
	平原	0.219	11.29	2.706	0.491 3
	山谷	0.235	13.34	2.846	0.531 2
	水源地	0.328	17.08	2.954	0.475 8
BP 神经网络	高山	0.346	13.70	3.412	0.423 6
	平原	0.197	10.02	2.898	0.508 7
	山谷	0.321	13.21	3.243	0.468 8
	水源地	0.289	12.94	3.365	0.524 2

2.3 组合模型预估实验结果

将两种方法的训练集预估结果进行对比分析,由表 4 可知,在这四种地形中,SVM 缺失值预估模型要比 BP 神经网络的更加稳定,误差也更小,但是在水源地和平原地形中 BP 神经网络的预估结果准确度相对更高,结合两种模型预估的结果,可以提高能见度预估的精度.实验输入测试集数据得到两种模型的预估结果,运用上文所述的方差倒数法,加权组合求出组合模型的预估值,并计算组合模型预估值

和实测值的平均相对误差、误差均值和最大误差.测试集能见度组合模型预估性能指标结果如表 5 所示,预估效果如图 5 所示.

表 5　组合方法的性能参数对比

Table 5　Performance of the combined SVM-BP neural network method for visibility estimation

	地形特征	RMSE/km	MAPE/%	最大误差/km
组合模型	高山	0.151	9.156	2.164
	平原	0.116	9.645	1.921
	山谷	0.128	9.423	2.098
	水源地	0.132	9.531	2.011

从表 5 中的实验结果数据可以看出,无论是哪种地形,组合模型预估的平均相对误差更低,整体误差均值小,效果要明显好于单一模型.由此可知,组合方法可以保留单一模型的预估优势,增加对缺失数据预估的可靠性.

3　结论

为解决自动气象站能见度要素缺测的问题,本文利用组合模型对缺测数据进行精确的预估,并以预估值代替实测值来保证数据的完备性.首先通过灰色关联分析方法精简预估模型输入,筛选出与能见度相关度较大的气象要素,再从气象信息的多种要素中建立能见度数据预估模型.在实验部分,两种模型对能见度都具有良好的预估能力,SVM 模型在四种地形中对能见度数据的预估结果比较好且稳定,而 BP 神经网络则对平原和水源地的数据预估能力突出.对不同模型预估的结果加权组合,结合测试集的 RMSE 和 MAPE 的数据,将这些数据与单一预估的方法对比,结果表明组合模型预估的方法更加接近实测值,更能充分地利用数据信息,从而提高对缺失数据预估的准确性.

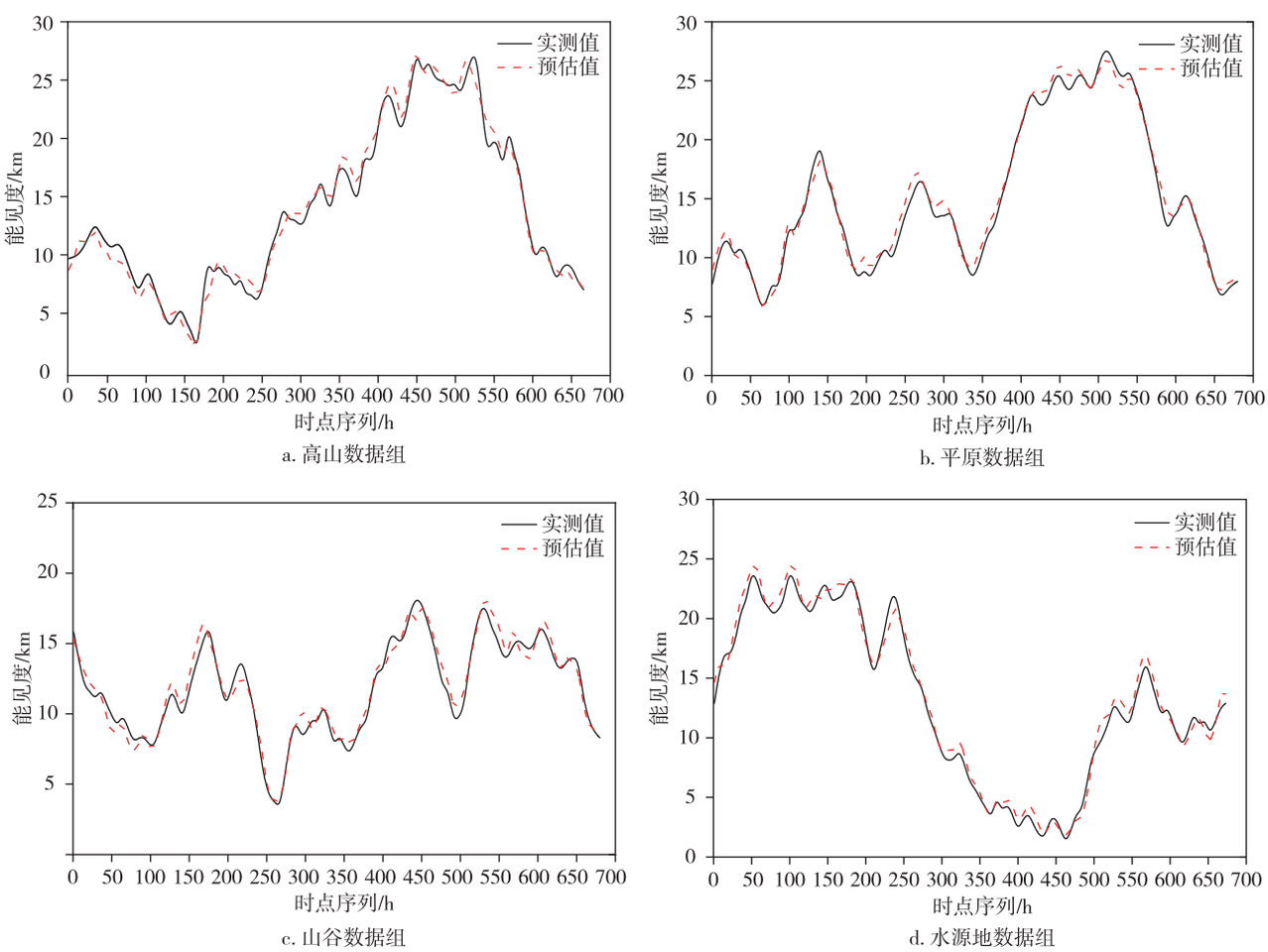

图 5　各地形能见度数据组预估效果

Fig. 5　Comparison of observed visibility and estimation by the combined SVM-BP neural network method for mountainous (a),plain (b),valley (c),and water source (d) areas

参考文献
References

[1] Feng S,Hu Q,Qian W H.Quality control of daily meteorological data in China,1951-2000:a new dataset[J].International Journal of Climatology,2004,24(7):853-870
[2] 李良富,王汉杰,刘金玉,等.基于黑板模型的地面气象数据质量控制[J].气象科技,2006,34(2):199-204
LI Liangfu,WANG Hanjie,LIU Jinyu,et al.Surface meteorological data quality control based on blackboard model [J].Meteorological Science and Technology,2006,34(2):199-204
[3] 窦以文,屈玉贵,陶士伟,等.北京自动气象站实时数据质量控制应用[J].气象,2008,34(8):77-81
DOU Yiwen, QU Yugui, TAO Shiwei, et al. The application of quality control procedures for real-time data from automatic weather stations[J].Meteorological Monthly,2008,34(8):77-81
[4] 熊安元,朱燕君,任芝花,等.观测仪器和百叶箱的变化对地面气温观测值的影响及其原因分析[J].气象学报,2006,64(3):377-384
XIONG Anyuan,ZHU Yanjun,REN Zhihua,et al.Differences of surface temperature observations recorded by different sensors in different screens and its causes[J].Acta Meteorologica Sinica,2006,64(3):377-384
[5] Jung K H,Yoo K Y.Data hiding method using image interpolation[J].Computer Standards & Interfaces,2009,31(2):465-470
[6] Hubbard K G,You J S.Sensitivity analysis of quality assurance using the spatial regression approach:a case study of the maximum/minimum air temperature[J].Journal of Atmospheric and Oceanic Technology,2005,22(10):1520-1530
[7] Lorenc A C.A global three-dimensional multivariate statistical interpolation scheme [J]. Monthly Weather Review,1981,109(4):701-721
[8] Hwang W J,Wen K W.Fast KNN classification algorithm based on partial distance search[J].Electronics Letters,1998,34(21):2062
[9] 冷泳林,陈志奎,张清辰,等.不完整大数据的分布式聚类填充算法[J].计算机工程,2015,41(5):19-25
LENG Yonglin,CHEN Zhikui,ZHANG Qingchen,et al.Distributed clustering and filling algorithm of incomplete big data[J].Computer Engineering,2015,41(5):19-25
[10] 杨毅,卢诚波.一种基于极限学习机的缺失数据填充方法[J].计算机应用与软件,2016,33(10):243-246
YANG Yi,LU Chengbo.A method for missing data imputation based on extreme learning machine[J].Computer Applications and Software,2016,33(10):243-246
[11] 郑斌.基于改进遗传算法的不完整大数据填充挖掘算法[J].微电子学与计算机,2016,33(2):96-99
ZHENG Bin. Incomplete data filling mining algorithm based on the improved genetic algorithm[J].Microelectronics & Computer,2016,33(2):96-99
[12] Lorenc A C,Hammon O.Objective quality control of observations using Bayesian methods. Theory, and a practical implementation [J]. Quarterly Journal of the Royal Meteorological Society,1988,114(480):515-543
[13] Poulos J,Valle R.Missing data imputation for supervised learning [J]. Applied Artificial Intelligence, 2018, 32(2):186-196
[14] Burges C J C.A tutorial on support vector machines for pattern recognition[J].Data Mining and Knowledge Discovery,1998,2(2):121-167
[15] Zhang J N,Song S J,Zhang X N.Sparse Bayesian ELM handling with missing data for multi-class classification [C]//Proceedings of ELM-2014,2015(1):1-13
[16] Vapnik V,Izmailov R.Knowledge transfer in SVM and neural networks[J].Annals of Mathematics and Artificial Intelligence,2017,81(1/2):3-19
[17] Chen R J,Lin J H.Identification of feature risk pathways of smoking-induced lung cancer based on SVM[J].Plos One,2020,15(6):e0233445
[18] 符欲梅,朱芳,咎昕武.基于支持向量机的桥梁健康监测系统残缺数据填补[J].传感技术学报,2012,25(12):1706-1710
FU Yumei, ZHU Fang, ZAN Xinwu. Missing data imputation in bridge health monitoring system based on the support vector machine [J]. Chinese Journal of Sensors and Actuators,2012,25(12):1706-1710
[19] Shardell M,Hicks G E.Statistical analysis with missing exposure data measured by proxy respondents:a misclassification problem within a missing-data problem[J].Statistics in Medicine,2014,33(25):4437-4452
[20] 丁小欧,王宏志,张笑影,等.数据质量多种性质的关联关系研究[J].软件学报,2016,27(7):1626-1644
DING Xiao'ou,WANG Hongzhi,ZHANG Xiaoying,et al.Association relationships study of multi-dimensional data quality[J].Journal of Software,2016,27(7):1626-1644
[21] Pelckmans K,De Brabanter J,Suykens J A K,et al.Handling missing values in support vector machine classifiers [J].Neural Networks,2005,18(5/6):684-692
[22] 张军华,任雄风,赵杰,等.基于交叉验证支持向量机储层预测方法及应用[J].科学技术与工程,2020,20(13):5052-5057
ZHANG Junhua,REN Xiongfeng,ZHAO Jie,et al.Reservoir prediction method and its application of support vector machine based on cross validation [J]. Science Technology and Engineering,2020,20(13):5052-5057

SVM-BP neural network based meteorological visibility data filling

YIN Liping[1,2] LIU Xiaoyu[1,2] SHENG Shaoxue[3] WEN Huayang[3] QIU Kangjun[3]
1 School of Automation,Nanjing University of Information Science & Technology,Nanjing 210044

2 Collaborative Innovation Center of Atmospheric Environment and Equipment Technology,
Nanjing University of Information Science & Technology,Nanjing 210044
3 Anhui Meteorological Information Center,Hefei 230031

Abstract Most automatic weather stations sample visibility with optical devices,which are vulnerable to interference from rain,snow and dust,resulting in the inaccuracy or missing of visibility data.To address this and provide complete data for meteorological prediction,this paper proposes a Support Vector Machine-Back Propagation (SVM-BP) neural network based data quality control method for visibility data correction and filling.First,the grey correlation analysis is used to select meteorological elements closely related with visibility.Second,the visibility data of different terrains are estimated by SVM and the BP neural network independently,which are then combined by optimal combination weights.Historical weather visibility data from Anhui Meteorological Bureau are used to verify the proposed method.The results show that compared with the independent SVM or the BP neural network,the combined estimation has smaller mean error and higher overall accuracy.The proposed SVM-BP neural network method provides an effective tool for visibility data filling,thus lays theoretical basis for weather forecasting,weather analysis,meteorological research and public service.

Key words combinatorial model;missing value estimation;correlation analysis;BP neural network;visibility;support vector machines (SVM)