计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(1):277−281 [doi: 10.15888/j.cnki.csa.007735]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

土壤属性数据 pH 缺失的插补方法①

张逸飞[1,2], 曹　佳[1,2]

[1](北京林业大学 信息学院, 北京 100083)
[2](国家林业草原林业智能信息处理工程技术研究中心, 北京 100083)
通讯作者: 曹　佳, E-mail: caojia@bjfu.edu.cn

摘　要: 土壤分析研究中属性数据缺失的现象时常发生, 为了提高研究结果的可靠性, 有必要对土壤属性数据的缺失值插补方法进行研究. 从数据挖掘的角度利用多种缺失值处理方法来对缺失值进行插补, 以中国主要农田生态系统土壤养分数据库的 pH 属性为研究对象, 并且从真实值和插补值的拟合优度和插补误差两个方面评估各个方法在不同缺失率的数据集上的表现. 结果表明, 对比其他方法, 如多元回归、SVM、神经网络, 采用最优参数的 KNN 和随机森林插补方法对土壤属性数据 pH 进行插补是有效可行的. KNN 和随机森林在不同缺失率的数据集上插补缺失数据 pH 的 *MAE*、*RMSE* 和 R^2 的均值分别为 0.132 和 0.131, 0.174 和 0.178, 0.775 和 0.765.

关键词: 土壤属性数据; pH; 缺失数据; K 最近邻居; 随机森林

引用格式: 张逸飞,曹佳.土壤属性数据 pH 缺失的插补方法.计算机系统应用,2021,30(1):277–281. http://www.c-s-a.org.cn/1003-3254/7735.html

Imputation Method to Predict Missing pH Data of Soil Attribute

ZHANG Yi-Fei[1,2], CAO Jia[1,2]

[1](School of Information Science and Technology, Beijing Forestry University, Beijing 100083, China)
[2](Engineering Research Center for Forestry-Oriented Intelligent Information Processing of National Forestry and Grassland Administration, Beijing 100083, China)

Abstract: The problem of the absence of attribute data often occurs in soil analysis and research. To improve the reliability of the research results, it is necessary to study the imputation methods for soil attribute missing data. In this study, a variety of imputation methods have been evaluated to interpolate the soil attribute missing data from the perspective of data mining. Using soil attribute pH as an interpolation object, the Soil Nutrient Database of China's Major Ecosystems is used as the source of physical and chemical soil attribute data. We evaluate the performance of each method on the dataset of different missing rates in terms of model fitting and imputation error. The result shows that it is feasible to impute soil attribute pH missing data using the optimal parameter K-Nearest Neighbor (KNN) and random forest than other methods, such as multivariable regression, support vector machine, and neural network. The mean value of *MAE*、*RMSE* and R^2 of the imputed missing data pH of KNN and random forest on the dataset with different missing rates are 0.132 and 0.131, 0. 174 and 0.178, 0.775 and 0.765, respectively.

Key words: soil attribute data; pH; missing data; K-Nearest Neighbor (KNN); random forest

1　引言

土壤是农业生产和人类活动中最重要的物质基础, 土壤属性数据是分析土壤理化性质和指导农作物种植的重要参考[1]. 土壤 pH 是土壤属性数据中的重要部分, 土壤酸碱化会影响土壤性质及微量元素的有效性, 直接或间接改变土壤肥力, 对植物生长发育造成影响[2].

① 基金项目: 国家自然科学基金 (61602042)
Foundation item: National Natural Science Foundation of China (61602042)
收稿时间: 2020-05-23; 修改时间: 2020-06-19; 采用时间: 2020-06-28; csa 在线出版时间: 2020-12-31

然而由于各种原因, 在土壤普查的过程中存在土壤 pH 缺失的情况. 本文将基于数据分析的方法, 研究土壤数据集的 pH 缺失值的填充方法.

对于土壤属性数据缺失的处理, 国际应用系统分析协会 (IIASA) 的和谐世界土壤数据库 (HWSD) 中采用拥有相同土壤类型的最适合的邻居单元的土壤属性数据来替代缺失值[3]. 韩光中等人运用了传统的土壤属性推绎模型, 通过逐步回归方法对土壤属性建立土壤传递函数, 插补了容重、速效养分、CEC 和氧化铁的缺失值[4]. 沈汉灵运用灰色关联系数法, 挖掘土壤属性之间的关联关系, 构建经验公式来插补土壤盐基饱和度[5]. Gargiulo 等人使用基于条件分布模型的多元回归方法, 归纳土壤属性数据之间的经验公式, 预测土壤属性数据的缺失值. 该方法考虑变量之间的相关性问题, 在预测土壤质地、容重等属性时表现很好, 但不能很好插补 pH 数据[6]. 专门针对土壤属性数据 pH 缺失值插补的具体研究较少.

数据缺失问题是一个常见的计算问题, 常用的缺失数据处理方法是插补法,即采用一个替代值填补样本中的缺失数据, 使填补后的数据与已有数据集的分布一致. 多元回归插补法运用数据自变量与因变量之间的关系进行插补, 线性插补法比均值填补法在环境数据集上填补缺失值有更优的表现[7]. Schafer 在的 EM (Expectation Maximization) 算法的基础上, 研究了多重插补法的应用[8]. 随着机器学习技术的发展, 运用机器学习方法处理缺失数据近几年引起了研究者们的广泛关注. Jerez 等运用乳腺癌的真实数据,比较了机器学习插补法和统计学插补法,认为机器学习插补法在处理高维数据时有显著的优势[9]. KNN 及其改进算法运用本身的 K 个具有完整值的最近邻居实现对缺失数据的插补, 由于操作简单被广泛运用[10]. 徐凯等将随机森林回归预测算法运用在地震插值中, 结果表明随机森林插补方法能够很好补全缺失信息, 而且数据差异性较小[11]. 吴郁等比较了 Logistic 回归、Probit 回归、朴素贝叶斯和随机森林方法在船舶交通事故数据集上的应用, 并证明了随机森林方法插补缺失值的精度更优[12]. 朱梦成等将 SVM 算法应用于处理医疗数据和社会调查数据中, 处理分类数据和连续型数据的缺失值[13]. 谢晓凯等运用 BP 神经网络建立空间结构中测点应力间、温度与应力间的相关关系模型, 并对其进行了适用性分析[14].

本文针对土壤属性数据 pH 的缺失问题, 将对比多元回归、KNN、随机森林、SVM 和神经网络 5 个插补方法, 从而选取插补正确率最高的方法.

2 研究方法

由于土壤属性数据中全是数值型连续变量, 以下介绍多元回归、K 最近邻、随机森林、支持向量机和神经网络共五种方法插补数值型连续变量的原理, 以及缺失值插补方法的评价方法.

2.1 多元回归插补法

多元回归插补法 (Multiple Regression, MR) 考虑到变量之间的线性相关性, 运用回归模型预测缺失值. 插补缺失数据时, 引入随机残差项与插补值相加, 作为最终插补结果, 使多元回归插补法插补的缺失数据不会扭曲样本的分布.

2.2 K 最近邻插补法

K 最近邻插补法 (K-Nearest Neighbor, KNN) 运用数据集中每条样本的完整属性, 计算缺失数据样本与完整数据样本之间的距离. 在所有完整数据样本中, 选择与目标缺失数据样本最小的 K 个数据样本作为目标缺失样本的最近邻. 最后利用这 K 个数据样本的缺失属性的平均值来插补目标缺失样本中的缺失值.

2.3 随机森林插补法

随机森林的思想是通过 Bootstrap 抽样技术, 有放回的在原始训练集上获得 N 个子训练集, 然后在这 N 个子训练集的基础上分别构建回归树, 组合得到随机森林模型. 当输出是连续型变量时, 其基础是 CART 回归树算法. CART 算法使用 Gini 指数来度量随机变量的不确定度的大小, 以此选择划分属性.

2.4 支持向量机插补法

采用 SVM 回归模型插补连续型数据. SVM 回归模型的优化问题是构造精度高、复杂性低的模拟函数来拟合真实样本数据. 模型引入损失函数来量化模型的预测值和样本的真实值的差距, 惩罚参数判断预测模型拟合的好坏.

2.5 神经网络插补法

神经网络 (neural networks) 在系统建模与非线性映射方面具有很强的适用性, 因此可以认为是缺失数据插补的有效手段. 在众多的神经网络当中, 反向传播 (Back Propagation, BP) 神经网络由于其较高的稳定性和精度被广泛运用. BP 神经网络通过误差反馈传播算

法, 建立输入与相应输出之间的映射关系,从而建立缺失数据的预测模型

2.6 缺失数据插补方法的评价方法

在进行插补方法的评价时, 为了避免计算出的插补数据没有参照, 选取土壤完整属性数据样本中的数据, 按照一定缺失比例使属性数据 pH 缺失, 生成对应的缺失属性数据样本. 其中, 生成的缺失属性数据样本数量为N, y_i表示第i条属性数据 pH 的真实值, $\hat{y}_i$表示其插补值.

(1) 拟合优度

采用决定系数R^2(coefficient of determination) 用于判断真实值和插补值的拟合优度, 其取值范围是[0,1]. 其值越接近于 1 代表变量之间有更好的拟合. 决定系数的计算公式如下:

$$R^2=\frac{SSR}{SST} \tag{1}$$

其中, *SSR* (Regression Sum of Squares) 称为回归平方和, *SST* (Total Sum of Squares) 称为总平方和. 对于简单线性回归而言, 决定系数为样本相关系数的平方[15], 即有:

$$R^2=\left[\frac{1}{N}\frac{\sum_{i=1}^{N}\left[(y_i-\bar{y})\left(\hat{y}-\bar{\hat{y}}\right)\right]}{\sigma_y\sigma_{\hat{y}}}\right]^2 \tag{2}$$

其中, σ_y和$\sigma_{\hat{y}}$分别是 pH 真实值和插补值的标准差.

(2) 插补误差

本文采用平均绝对误差 (Mean Absolute Error, *MAE*) 与均方根误差 (Root Mean Square Error, *RMSE*) 来反映了真实值与插补值之间的误差, 它们的值越小, 代表插补值与真实值越接近, 插补方法对数据集的插补效果越好.

$$MAE=\frac{1}{N}\sum_{i=1}^{N}|y_i-\hat{y}_i| \tag{3}$$

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2} \tag{4}$$

3 结果与分析

3.1 数据说明

在一条土壤属性数据样本中, 如果一条样本中包含所有监测的土壤属性, 称为完整属性数据样本, 否则称为缺失属性数据样本. 本文数据来自于中国科学院南京土壤研究所“中国主要农田生态系统土壤养分数据库 (1990–2006)”[16]. 本文采用该数据库中砂粒含量、粉粒含量、容重平均值、容重标准差、有机质、全氮、全磷、全钾和 pH 共 9 个属性为分析对象, 选取 458 条土壤数据样本, 其中完整属性数据样本 148 条, 仅缺失 pH 属性数据样本 310 条.

本文采用交叉验证法对土壤数据缺失数据的插补结果进行评估. 在 148 条土壤的完整属性数据样本中, 通过随机剔除属性数据 pH 的方法, 以不同的比例分别构造训练集和验证集. 例如, 在完整属性数据样本中随机剔除 10% 的属性数据 pH 产生缺失属性数据样本作为验证集, 其余完整属性数据样本为训练集; 以此再以 20%, 30%, 40%, 50%, 60% 和 70% 的比例分别构造 6 对训练集和验证集. 插补方法运用各训练集来建立对应的缺失数据的插补模型, 验证集用来调节各方法的参数, 选择具有最小泛化误差的模型作为最终模型. 我们将 310 条 pH 有缺失的数据构成测试集, 用来最终评估模型对缺失数据的插补效果.

3.2 各方法的最佳参数的设置

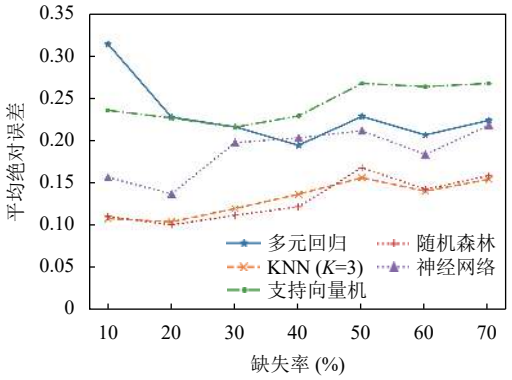
采用不同的插补方法, 在训练集样本分别建立不同的缺失值插补模型, 运用对应的验证集对方法的参数进行最优化调参. 本文分别选取各方法的部分主要参数进行调节, 采用网格搜索算法选出各方法的最优参数. 经过调研, 在 KNN 方法中, 调节待插补样本的最近邻居数量 K 值[17]; 在随机森林方法中, 调节控制生成一棵决策树所随机选取的属性特征数量和最终生成的决策树数[18]; 在 SVM 方法中, 采用 RBF 核函数, 调节核参数和误差惩罚因子[19]; 在 BP 神经网络方法中, 调节网络的隐含层节点数量、学习速率、优化算法、最大训练次数、dropout 比例、期望误差和各层神经元的激活函数[20]. 通过在上述验证集上训练进行调参, 得到不同方法在中国主要农田生态系统土壤养分数据库上建立插补模型插补土壤属性数据 pH 时的最佳参数如表 1.

3.3 插补方法对比

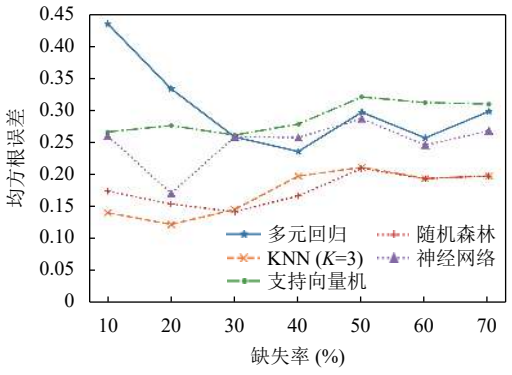
针对中国主要农田生态系统土壤养分数据库中属性属性 pH 缺失的问题, 在对应缺失率的验证集上, 5 个具有最优化参数的方法构造的模型所得的插补结果的平均绝对误差 *MAE*、均方根误差 *RMSE* 和决定系数 R^2 如图 1 所示.

表 1 不同插补方法的最优参数选择

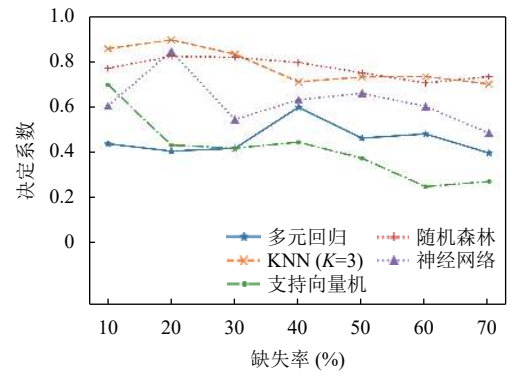
插补方法	最优参数选择
*K*最近邻居	*K*=3
随机森林	构成决策树随机选取的属性特征数量=8, 最终生成的决策树数=10
支持向量机	核函数为RBF, 核参数=0.125, 误差惩罚因子=10
神经网络	隐含层节点数量=64, 学习速率=0.01, 优化算法=adam, 最大训练次数=200, dropout=0.2, 期望误差=0.01, 隐含层神经元激活函数=ReLU, 输出层神经元激活函数=Linear



(a) *MAE*



(b) R*MSE*



(c) R^2

图 1 不同插补方法的实验结果对比

由图 1 可知,随着缺失率的增加, KNN、随机森林和支持向量机的插补效果均呈下降趋势. 其中, KNN 和随机森林的插补效果波动性较小. 在缺失率 10%~20% 时, KNN 方法表现更好, 在缺失率 40% 时, 随机森林方法表现更好, 其他情况下两方法的评价指标均较为接近. SVM 方法插补效果受缺失率影响较大, 随着缺失率的增加, 该方法的插补效果越来越差. 多元回归方法插补缺失数据在缺失为 40% 时插补效果最优. 该方法插补缺失值的表现随着缺失率的增加, 先增加后下降, 此结论与文献 [7] 一致. 神经网络插补效果的波动性较大, 在缺失率为 20% 时插补效果较好.

由评价指标可知, 在任何缺失率下, 多元回归、SVM 和神经网络插补属性数据 pH 时, 插补能力均较弱. KNN 和随机森林方法的 *MAE* 和 *RMSE* 值都是最小, R^2值都更接近于 1, 因此二者的插补效果都是最好的. 为了进一步对比 KNN 和随机森林方法, 我们对中国主要农田生态系统土壤养分数据库中 310 条 pH 有缺失的测试集进行插补操作, 并且将插补后的数据特征与 148 条完整数据样本进行对比, 结果如表 2 所示. 从表可见, KNN 所得插补后的均值、最大值和最小值更接近完整数据样本更接近完整数据样本, 因此 KNN 可以更灵活地插补 pH 数据的最值.

表 2 测试集和完整数据样本的 pH 数据特征

样本	插补方法	平均值	标准差	最小值	下四分位点	中位数	上四分位点	最大值
完整数据样本	--	8.782	0.363	8.100	8.470	8.730	9.100	9.630
测试集	KNN	8.700	0.306	8.233	8.366	8.753	8.983	9.232
	随机森林	8.632	0.169	8.244	8.569	8.692	8.737	8.998

4 结束语

针对土壤属性数据 pH 缺失这个在土壤调查研究中的常见问题, 本文从真实值和插补值的拟合优度和插补误差两个方面比较了 5 种缺失数据插补方法在不同 pH 缺失率情况下插补效果. 实验结果表明, 多元回归、支持向量机和神经网络方法不适合用于插补 pH 数据. KNN 算法和随机森林方法所受数据集和缺失率的影响较小, 建立的模型表现稳定, 两者均适合用于土壤属性数据 pH 值的插补.

参考文献

1 龚子同, 张甘霖, 陈志诚, 等. 土壤发生与系统分类. 北京: 科学出版社, 2007.

2 唐琨, 朱伟文, 周文新, 等. 土壤 pH 对植物生长发育影响的研究进展. 作物研究, 2013, 27(2): 207–212. [doi: 10.3969/j.issn.1001-5280.2013.02.25]
3 FAO/IIASA/ISRIC/ISS-CAS/JRC. Harmonized world soil database (version 1.1). Rome: FAO, 2009.
4 韩光中, 杨银华, 吴彬, 等. 基于传递函数的土壤数据库缺失数据的填补研究. 土壤, 2019, 51(5): 1036–1041.
5 沈汉灵. 基于数据挖掘技术的土壤属性数据处理研究 [博士学位论文]. 广州: 华南农业大学, 2016.
6 Gargiulo O, Morgan KT. Procedures to simulate missing soil parameters in the Florida soils characteristics database. Soil Science Society of America Journal, 2015, 79(1): 165–174. [doi: 10.2136/sssaj2014.05.0194]
7 Noor NM, Abdullah MMAB, Yahaya AS, *et al*. Comparison of linear interpolation method and mean method to replace the missing values in environmental data set. Materials Science Forum, 2014, 803: 278–281. [doi: 10.4028/www.scientific.net/MSF.803.278]
8 Schafer JL. Multiple imputation: A primer. Statistical Methods in Medical Research, 1999, 8(1): 3–15. [doi: 10.1177/096228029900800102]
9 Jerez JM, Molina I, García-Laencina PJ, *et al*. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial Intelligence in Medicine, 2010, 50(2): 105–115. [doi: 10.1016/j.artmed.2010.05.002]
10 郝胜轩, 宋宏, 周晓锋. 基于近邻噪声处理的 KNN 缺失数据填补算法. 计算机仿真, 2014, 31(7): 264–268. [doi: 10.3969/j.issn.1006-9348.2014.07.060]
11 徐凯, 孙赞东. 基于随机森林方法的地震插值方法研究. 石油科学通报, 2018, 3(1): 22–31.
12 吴郁, 张金奋, 范存龙, 等. 基于随机森林的船舶碰撞事故缺失数据插补. 武汉理工大学学报 (交通科学与工程版), 2019, 43(6): 1120–1124.
13 朱梦成. 面向缺失数据处理的 SVM 算法研究 [硕士学位论文]. 天津: 天津大学, 2017.
14 谢晓凯, 罗尧治, 张楠, 等. 基于神经网络的大跨度空间钢结构应力实测缺失数据修复方法研究. 空间结构, 2019, 25(3): 38–44.
15 Junninen H, Niska H, Tuppurainen K, *et al*. Methods for imputation of missing values in air quality data sets. Atmospheric Environment, 2004, 38(18): 2895–2907. [doi: 10.1016/j.atmosenv.2004.02.026]
16 潘贤章, 施建平, 宋歌, 等. 中国主要农田生态系统土壤养分数据 (1990–2006). 国家科技资源共享服务平台-国家地球系统科学数据中心-土壤分中心.
17 毋雪雁, 王水花, 张煜东. *K* 最近邻算法理论与应用综述. 计算机工程与应用, 2017, 53(21): 1–7. [doi: 10.3778/j.issn.1002-8331.1707-0202]
18 温博文, 董文瀚, 解武杰, 等. 基于改进网格搜索算法的随机森林参数优化. 计算机工程与应用, 2018, 54(10): 154–157. [doi: 10.3778/j.issn.1002-8331.1612-0328]
19 林升梁, 刘志. 基于 RBF 核函数的支持向量机参数选择. 浙江工业大学学报, 2007, 35(2): 163–167. [doi: 10.3969/j.issn.1006-4303.2007.02.010]
20 王巧利, 林剑辉, 许彦峰. 基于 BP 神经网络的土壤容重预测模型. 中国农学通报, 2014, 30(24): 237–245. [doi: 10.11924/j.issn.1000-6850.2014-0307]