Statistics and Application 统计学与应用, 2019, 8(2), 218-226
Published Online April 2019 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2019.82024

Hans汉斯

Study on Heavy Metal Pollution in European Soil Based on Random Forest Algorithms

Shenhui Song[1], Shuyun Xie[2], Ruiyan Yang[1]

[1]School of Mathematics and Physics, China University of Geosciences, Wuhan Hubei
[2]Faculty of Earth Sciences, China University of Geosciences, Wuhan Hubei
Email: 2419291977@qq.com

Received: Mar. 6th, 2019; accepted: Mar. 21st, 2019; published: Mar. 29th, 2019

Abstract

Under the background of large data, in order to improve the efficiency of evaluating heavy metal pollution in soil, a random forest algorithm in machine learning is introduced. In this paper, Random forest model was established to analyze the pollution degree of As, Co, Cr, Cu, Ni, Pb and Zn in top soil of Europe. Then, the KPCA-Random forest model is established by adding the kernel principal component analysis to improve the model, and the classification accuracy and running time are compared. The results show that the classification accuracy of the improved model is improved from 93.41% to 94.67%, and the running time is reduced from 12.530601 s to 9.437811 s. Finally, the advantages and disadvantages of the Random forest model are evaluated, and the future research directions are also proposed.

Keywords

Random Forest, Node Splitting Algorithm, Kernel Principal Component Analysis, Heavy Metal Pollution

基于随机森林算法的欧洲土壤重金属污染研究

宋申辉[1]，谢淑云[2]，杨瑞琰[1]

[1]中国地质大学（武汉）数学与物理学院，湖北 武汉
[2]中国地质大学（武汉）地球科学学院，湖北 武汉
Email: 2419291977@qq.com

收稿日期：2019年3月6日；录用日期：2019年3月21日；发布日期：2019年3月29日

摘　要

在大数据背景下，为提高评价土壤中重金属污染的效率，引入机器学习中的随机森林算法。本文以欧洲

文章引用：宋申辉，谢淑云，杨瑞琰．基于随机森林算法的欧洲土壤重金属污染研究[J]．统计学与应用，2019，8(2): 218-226. DOI: 10.12677/sa.2019.82024

表层土壤为例，建立Random forest模型，对As、Co、Cr、Cu、Ni、Pb、Zn 7种重金属的污染程度进行分类；然后通过加入核主成分分析对模型进行改进，建立KPCA-Random forest模型，并从分类精度和运行时间两个维度上进行对比。结果显示：改进后模型的分类精确度由93.41%提高到94.67%，运行时间从12.530601 s缩减到9.437811 s。最后本文对建立的随机森林模型的优缺点进行了评价，并提出今后的研究方向。

关键词

随机森林，节点分裂算法，核主成分分析，重金属污染

1. 引言

随着技术发展的不断加快和人们对土壤质量的关注，目前土壤质量问题越来越受到人们的重视[1] [2]。传统的土壤质量评价方法是基于大量人力物力，从采样到室内分析，人们几乎要参与到各个方面，如果仅仅靠手工来做分析，那无疑是最耗时的。随着实验设备的不断改进，产生了大量的数据，这就需要快速的分析和产生结果。因此，引入机器学习并合理参考相关成熟算法来解决地学中的相关问题是非常重要的[3] [4] [5]。目前关于土壤重金属污染的研究主要还是基于传统的地质统计学方法，包括地累积系数法、富集系数与变异系数、多重分形、相关分析等。

随机森林是利用多棵决策树基于统计学对样本进行训练，提炼出某种规则，并将该规则用于预测的一种组合分类器，已成为在面临分类问题上，可以优先选择的算法之一。Scarpone 等人提出了一种使用随机森林分类器和遗留土地数据定位加拿大不列颠哥伦比亚省南部山地景观中 EB 区域的方法，与传统的土地覆盖图相比，使用射频模型后，电子地图的精确度从 48%提高到 88% [6]。到目前为止，随机森林算法已经被广泛用于生物信息学、生态学、遗传学、环境监控、金融学、遥感地理学等众多领域，却很少应用在重金属检测方面。本文引入随机森林算法，以欧洲地区为例，对该地区的表层土壤中重金属含量进行了分类预测，为了说明研究结果的合理性，最后采用 C-A 多重分形方法进行验证和对比。

2. 方法

随机森林属于一种集成方法，首先训练出多个模型，这些模型之间相互独立，然后把这些模型产生的结果放在一起，得票最多的为最终的预测结果[7]。“森林”之意来源于随机森林包含多个相互无关联的决策树。该方法优于传统统计学方法的地方在于不需要事先了解或假设数据的分布情况，而且克服了决策树这种单分类器过度拟合和局部最优解的问题，同时也是一种非参数的模式识别分类方法，可以应用于大部分的数据分类场景。如果数据是高维的，那么随机森林在学习速度、抗噪音能力、预测精度上的优势就越发明显。

2.1. 集成学习方法 Bagging

Bagging 获取训练集的方式是通过自助采样的方法(bootstrap sampling)，按照随机原则，每次在大小为 n 的原始集合中获取 m 个样本，而这 m 个样本中可能含有多次被取到的样本 i，同时也可能样本 j 永

远都没被取到[8]。样本在 m 次采样中始终不被采集到的概率为$\left(1-\frac{1}{m}\right)^m$，取极限可得：

$$\lim_{m\to\infty}\left(1-\frac{1}{m}\right)^m \to \frac{1}{e}\approx 0.368 \tag{1}$$

由公式(1)可知，通过 Bagging 方式得到的训练集与原始集合相比，有 63.2%的重复率。这种有重复性的选取样本的方法可以增加各个学习器之间的差异性，使每个学习器之间尽可能的相互独立，那么组合起来的学习器在泛化能力上有更好的优势。

Bagging 的算法步骤如下：

第一步：输入训练集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$，弱学习算法 L；训练次数 T。

第二步：输出最终分类器 $G(x)$。

对 $m=1,2,\cdots,M$：首先用 bootstrap sampling 对集合 T 选样，得到集合 T_m；然后再次使用该方法得出训练集 T_m、学习基本分类器 $G_m(x)$，最后得出最终分类器 $G(x)$。

$$G(x)=\arg\max_{y\in\gamma}\sum_{m=1}^{T} I\left[G_m(x)=y\right] \tag{2}$$

2.2. 构造决策树

本文在生成决策树时利用 ID3 算法。

ID3 算法选择熵作为寻找分裂特征的指标，每个节点会选择导致最大熵降幅的特征用于分裂，熵的减少量称为信息增益[9]。计算过程如下：

1) 计算每个节点在分裂前的熵：

$$\text{Entropy}(T)=-\sum_{i=1}^{c} P_i\log_2 P_i \tag{3}$$

其中，P_i 表示特定目标类别 i 在总样本量中的占比。

2) 根据特征 A 将数据集 T 分裂为 k 个子集后的熵：

$$\text{Entropy}_A(T)=\sum_{j=1}^{k}\left[\frac{T_i}{T}\bullet Entropy(T_j)\right] \tag{4}$$

其中，$\frac{|T_i|}{|T|}$ 代表第 j 个划分的权重，划分的纯度随着 $\text{Entropy}_A(T)$ 的减小而强。

3) 由公式(11)和公式(12)可得出信息增益：

$$\text{Gain}(T,A)=\text{Entropy}(T)-\text{Entropy}_A(T) \tag{5}$$

其中，$Gain(T,A)$ 表示信息增益。

2.3. 构造随机森林

详细步骤如下所示：

step 1：假设含 N 个样本的数据集 T，采用有放回随机抽样的方式再次组成一个大小为 N 的集合 T_i。

step 2：假设每个样本具有 M 个属性，随机选取 m 个属性($m<M$)，根据构造决策树的原理，通过比较选择最佳属性作为分裂节点在不进行剪枝的情况下递归生成一棵决策树。

step 3：重复步骤 1、2 中决策树构造过程，形成多棵相互独立的决策树，从而构成随机森林。

3. 实例研究

3.1. 数据来源

本文以欧洲表层土壤为研究对象，数据源于地球化学基准值填图计划[10]，采样按照全球参考网格(GRN)，每个网格大小是 160 km × 160 km，每有 5 个采样点，共计 773 个样本。样本点分布在整个欧洲(见图 1)，包括 25 个国家(见表 1)。为保证数据的一致性和避免实验室之间存在分析偏差，所有样本均在同一实验室测量，同一种元素在同一实验室用同一种方法分析。

Figure 1. Distribution of soil sampling points in Europe
图 1. 欧洲土壤采样点分布

Table 1. Statistics of sampling sites in European countries
表 1. 欧洲各国采样点统计

序号	国家	样本量
1	阿尔巴尼亚	2
2	奥地利	15
3	比利时	5
4	瑞士	9
5	捷克	10
6	德国	72
7	丹麦	5
8	西班牙	52
9	爱莎尼亚	11
10	法国	116
11	芬兰	65
12	希腊	41
13	克罗地亚	13
14	匈牙利	14
15	意大利	47
16	爱尔兰	7
17	立陶宛	15
18	拉脱维亚	8

Continued

19	荷兰	7
20	挪威	58
21	葡萄牙	19
22	波兰	55
23	斯洛伐克	15
24	瑞典	51
25	英国	60

3.2. 数据预处理

为了利用随机森林算法训练数据，根据国际土壤重金属含量标准，将原始数据分为四个层次(见表 2)。

Table 2. Classification criteria of heavy metals in top soils under international standards (mg/kg)
表 2. 表层土壤在国际标准下所含重金属分类标准(mg/kg)

重金属	级别			
	一级	二级	三级	四级
As	<29	29~55	>55	-
Cr	<100	100~380	>380	-
Cu	<36	36~190	>190	-
Pb	<85	85~300	300~530	>530
Ni	<35	35~60	60~210	>210
Zn	<140	140~250	250~720	>720
Co	<9	9~14	14~240	>240

3.3. Random forest 模型建立与结果分析

我们使用欧洲地区采样点数据，建立 Random forest 模型，利用 MATLAB 软件编程，对数据训练后，输出预测的分类结果，并计算出分类精确度(见表 3)。

Table 3. Precision of predicted results
表 3. 预测结果精确度

重金属	样本数量		测试集中错误分类数量	精确度(%)
	训练集	测试集		
Zn	579	193	2	98.96
As	579	193	6	96.89
Cr	579	193	0	100
Cu	579	193	5	97.41
Ni	579	193	21	89.12
Pb	579	193	5	97.41
Co	579	193	50	74.09
		平均精确度		93.41
		最高精确度		100

从表 3 可知，Random forest 模型的平均精确度为 93.41%，精确度最低为 74.08%，最高可达 100%。

3.4. 模型改进

为了进一步提高模型的精确度，降低树与树之间的相关性，本文考虑考虑引入加入核函数的主成分分析(KPCA)来达到降维的目的，从而提高随机森林的分类性能。

3.4.1. KPCA 实现步骤

KPCA 是在 PCA 基础上的一种非线性推广，在处理非线性问题上具有一定的优势[11]。

1) 对应给定的原始空间 $T=\{x_1,x_2,\cdots,x_n\}$，引入非线性映射 Φ，使原始空间 T 投影到特征空间 F，即：$x_i\to\Phi(x_i),(i=1,2,\cdots n)$，且满足 $\sum_{i=1}^{n}\Phi(x_i)=0$，则在 F 空间中的协方差阵为：

$$\bar{c}=\frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)\Phi(x_i)^T \tag{10}$$

2) 计算核矩阵 $K=(k_{ij})_{n\times n}$：

$$k_{ij}=K(x_i,x_j)=\Phi(x_i)\Phi(x_j) \tag{11}$$

3) 中心化后的核矩阵 $\hat{K}$ 为：

$$\begin{aligned}\hat{K}&=\hat{\Phi}(\mathrm{X})^T\hat{\Phi}(\mathrm{X})\\&=\left[\Phi(\mathrm{X})-\Phi(\mathrm{X})E_n\right]^T\left[\Phi(\mathrm{X})-\Phi(\mathrm{X})E_n\right]\\&=K-E_nK-KE_n+E_nKE_n\end{aligned} \tag{12}$$

其中，E_n 为 $n\times n$ 的矩阵，$E_{ij}=\frac{1}{n}$。

4) 计算出核矩阵 $\hat{K}/n$ 的特征根 $(\lambda_1,\lambda_2,\cdots,\lambda_n)$，以及对应的特征向量 $(v_1,v_2,\cdots,v_n)$。

5) 用斯密特正交化后的特征向量为：$(\alpha_1,\alpha_2,\cdots,\alpha_n)$。

6) 空间 T 的核主成分为：

$$T_j(x)=\left[\frac{1}{\sqrt{\lambda_1}}\sum_{i=1}^{n}\alpha_iK(x_i,x),\frac{1}{\sqrt{\lambda_2}}\sum_{i=1}^{n}\alpha_iK(x_i,x),\cdots,\frac{1}{\sqrt{\lambda_k}}\sum_{i=1}^{n}\alpha_iK(x_i,x)\right] \tag{13}$$

其中，$j=1,2,\cdots,n$，λ_k 为第 k 个主成分的特征根。

7) 输出数据：$\hat{T}=\{T_1,T_2,\cdots,T_n\}$。

3.4.2. KPCA-Random forest 模型建立与结果分析

本文在用 KPCA 提取特征之前，需要先估计数据集的基本维数，此时我们使用最大似然估计来估计基本维；通过 KPCA 处理后，将数据集的维度从八个维度减少到四个维度；然后用随机森林算法训练和测试。用 MATLAB 编程实现后输出此模型的分类精度，与改进前的模型进行对比，结果见表 4。

从表 4 可以看出，改进后的模型精度为 94.67%，比改进前提高了 1.26%。此外，运行时间也从 12.530601 s 缩减到 9.437811 s。结果表明，KPCA 处理的数据不仅可以用于支持向量机，还可以用于提高随机森林算法的分类精度，这为实现地理智能化处理提供了一种思路。

我们使用 MAPGIS 软件绘制了具体的空间分布情况，见图 2。

Table 4. Precision comparison of model prediction
表 4. 模型预测精确度对比

元素	样本量		改进前的精确度(%)	改进后的精确度(%)
	训练集	测试集		
Zn	579	193	98.96	96.89
As	579	193	96.89	96.89
Cr	579	193	100	82.38
Cu	579	193	97.41	100
Ni	579	193	89.12	97.41
Pb	579	193	97.41	95.85
Co	579	193	74.09	93.26
	平均精确度		93.41	94.67
	最大精确度		12.530601	9.437811

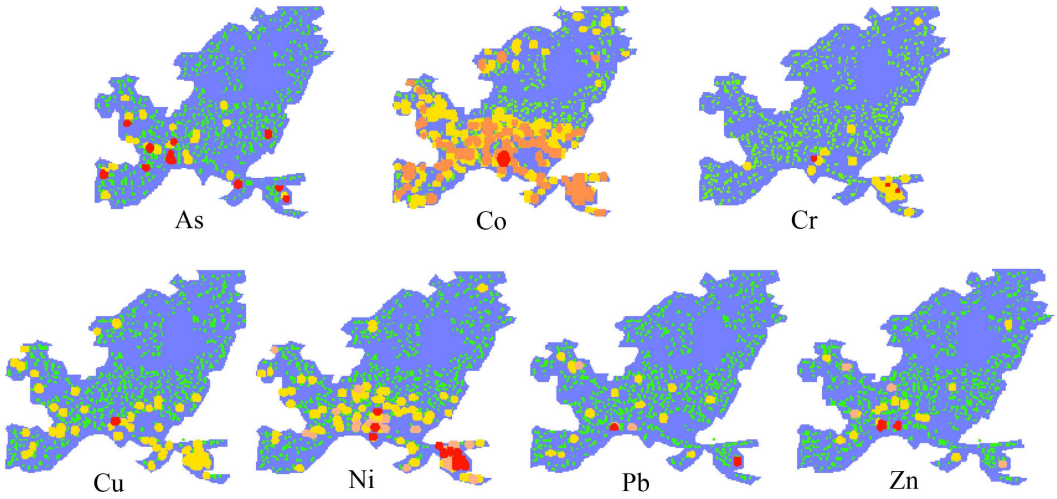

Figure 2. Distribution map of soil pollution under random forest model
图 2. 随机森林模型下土壤污染分布图

4. 模型评价

4.1. 模型的优点

本文对建立的随机森林模型分别从可靠性、精确度和重用性上进行评价。

1) 模型可靠性

为验证模型是否可靠，本文采用常用的研究土壤污染情况的 C-A 分形方法[12]，通过 GeoDAS 绘制出重金属元素空间分布图，见图 3。

对比可知，两种方法下提取的异常位置大体一致，从而可以看出随机森林模型对圈定化学异常位置上具有可靠性。

2) 模型精确度

对比分析表明，随机森林模型下缩小了异常范围，而且改进后模型的平均分类精度达到 94.67%，在预测精确度方面具有很好的效果。

3) 模型可重用性

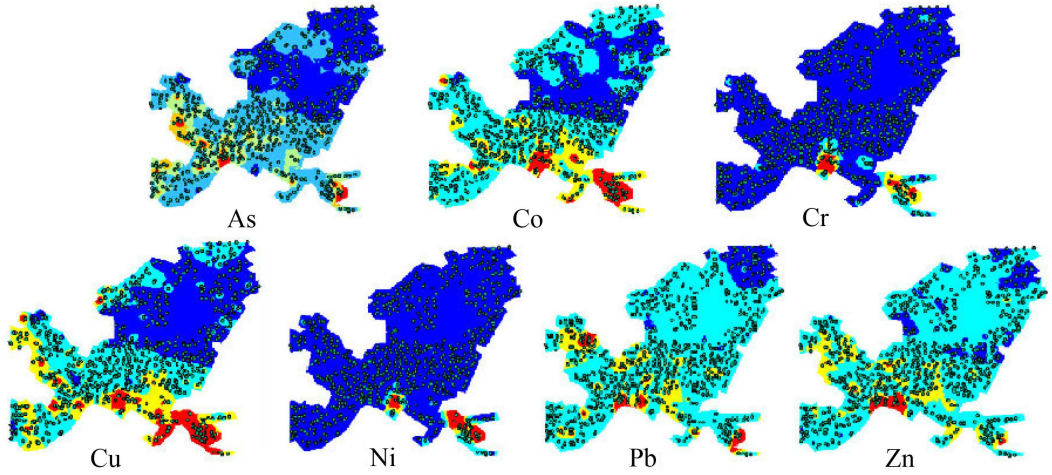

Figure 3. Distribution map of soil pollution based on c-a fractal method
图 3. C-A 分形方法下土壤污染分布图

相似数据可以直接用训练过的规则进行分类，不需要再训练，这比传统的地质分类方法具有更高的可重用性，同时效率也将大大提高。

4.2. 模型的缺点

本文建立的随机森林模型是在不考虑分布不平衡数据的情况下。该模型在面对类别分布不平衡的数据时，会出现分类效果不佳、误差率变高等一系列问题。

5. 结束语

本文使用了随机森林来研究土壤重金属污染问题，同时对建立的随机森林模型进行了改进。从实验结果来看，核主成分分析与随机森林结合建立的模型在分类效率和运行时间上均优于标准的随机森林模型，且评价结果与传统地质学中 C-A 多重分形得出的结果相符。在接下来的工作中，我们将考虑如何实现不平衡数据集下的随机森林模型，以及如何实现分布式并行计算，对于提高大数据时代的分类效率具有重要意义。

参考文献

[1] Rivera, M.B., Giráldez, M.I. and Fernández-Caliani, J.C. (2016) Assessing the Environmental Availability of Heavy Metals in Geogenically Contaminated Soils of the Sierra de Aracena Natural Park (SW Spain). Is There a Health Risk? *Science of the Total Environment*, **560-561**, 254-265. https://doi.org/10.1016/j.scitotenv.2016.04.029

[2] Batjargal, T., Otgonjargal, E., Baek, K. and Yang, J.S. (2012) Assessment of Metals Contamination of Soils in Ulaanbaatar, Mongolia. *Journal of Hazardous Materials*, **184**, 872-876. https://doi.org/10.1016/j.jhazmat.2010.08.106

[3] Alavi, A.H., Gandomi, A.H. and Lary, D.J. (2016) Progress of Machine Learning in Geosciences: Preface. *Geoscience Frontiers*, **7**, 1-2. https://doi.org/10.1016/j.gsf.2015.10.006

[4] Lary, D.J., Alavi, A.H., Gandomi, A.H. and Walker, A.L. (2016) Machine Learning in Geosciences and Remote Sensing. *Geoscience Frontiers*, **7**, 3-10. https://doi.org/10.1016/j.gsf.2015.07.003

[5] Anifowose, F.A., Labadin, J. and Abdulraheem, A. (2017) Ensemble Machine Learning: An Untapped Modeling Paradigm for Petroleum Reservoir Characterization. *Journal of Petroleum Science and Engineering*, **151**, 480-487. https://doi.org/10.1016/j.petrol.2017.01.024

[6] Scarpone, C., Schmidt, M.G., Bulmer, C.E. and Knudby, A. (2017) Semi-Automated Classification of Exposed Bedrock Cover in British Columbia's Southern Mountains Using a Random Forest Approach. *Geomorphology*, **285**, 214-224. https://doi.org/10.1016/j.geomorph.2017.02.013

[7] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324

[8] 王浩. 基于随机森林的网络攻击检测方法[J]. 软件, 2016, 37(11): 60-63.

[9] 彭程, 文雨, 李楚畅. 基于决策树算法的医疗大数据[J]. 信息技术与信息化, 2018, 222(9): 70-74.

[10] Lado, L.R., Heng, T. and Renter, H.I. (2008) Heavy Metals in European Soils: A Geostatistical Analysis of the FOREGS Geochemical Database. *Geoderma*, **48**, 189-199. https://doi.org/10.1016/j.geoderma.2008.09.020

[11] 赵帅, 李妍君, 熊伟丽. 基于 KPCA-Bagging 的高斯过程回归建模方法及应用[J]. 控制工程, 2019, 26(1): 131-136.

[12] Brouers, F. and Al-Musawi, T.J. (2018) Brouers-Sotolongo Fractal Kinetics versus Fractional Derivative Kinetics: A New Strategy to Analyze the Pollutants Sorption Kinetics in Porous Materials. *Journal of Hazardous Materials*, **350**, 162-168. https://doi.org/10.1016/j.jhazmat.2018.02.015