中国农学通报 2019,35(13):152-156
Chinese Agricultural Science Bulletin

基于主成分与支持向量机的邵阳县烟草产量预测

张 泰[1,2],张 莉[3],彭佳红[1]
([1]湖南农业大学信息科学技术学院,长沙 410128;[2]湖南农业大学图书馆,长沙 410128;
[3]湖南应用技术学院信息工程学院,湖南常德 415000)

摘 要:为探索准确预测邵阳县烟草产量的方法,首先对邵阳县70个植烟区土壤样本中的碱解氮、有效磷、速效钾等19个养分指标进行主成分分析,得出邵阳县烟草产量主要受有机质、有效锌、有效硼、有效锰、有效硫、交换性钙、全钾、有效铁和速效钾等9个养分含量的影响。在此基础上,基于支持向量机回归算法SVR对70个植烟区的烟草产量进行回归预测。结果发现,以主成分分析后的9个养分指标作为特征变量预测得到的烟草产量的均方误差明显小于以19个养分指标作为特征变量预测得到的烟草产量的均方误差。同时,对比SVR算法和随机森林回归算法发现,SVR算法的预测精度明显优于随机森林回归算法。基于主成分与支持向量机的回归算法是预测邵阳县烟草产量的有效方法。

关键词:支持向量机;随机森林;主成分分析;烟草产量;土壤养分
中图分类号:S-3 **文献标志码:**A **论文编号:**casb18030137

Prediction of Tobacco Yield in Shaoyang Based on Principal Component Analysis and SVR

Zhang Tai[1,2], Zhang Li[3], Peng Jiahong[1]
([1]*College of Information Science and Technology*, *Hunan Agricultural University*, Changsha 410128;
[2]*The Library of Hunan Agricultural University*, Changsha 410128;
[3]*College of Information Engineering*, *Hunan Applied Technology University*, Changde Hunan 415000)

Abstract: To accurately predict the yield of tobacco in Shaoyang, soil samples from 70 tobacco growing areas were taken as the research objects. Firstly, the principal component analysis (PCA) was used to analyze the impact of 19 soil nutrient indexes on tobacco yield. The results showed that the yield of tobacco in Shaoyang was mainly affected by 9 soil nutrient indexes, including organic matter, available zinc, available boron, available manganese, available sulfur, exchangeable calcium, total potassium, available iron and available potassium. After that, support vector regression (SVR) was used to predict the tobacco yield of the 70 tobacco growing areas in Shaoyang. The results showed that the mean square error (MSE) of prediction results with 9 soil nutrient indexes was significantly less than the MSE of prediction results with 19 soil nutrient indexes. Finally, compared with that of random forest regression algorithm, the prediction accuracy of SVR was obviously better. The method based on principal component analysis and support vector regression is effective to predict tobacco yield in Shaoyang.

Keywords: support vector regression; random forest; principal component analysis; tobacco yield; soil nutrient

0 引言

“良由地气使然也,根长全赖地肥力,气厚半借土膏腴”,土壤肥力对于烟草产量和品质的影响力可见一斑。土壤肥力不仅影响烟草的长势,还影响烟叶的口

基金项目:湖南省科技厅项目“湘中生态公益林重点建设区植被复技术研究”(S2006N332)。
第一作者简介:张泰,男,1978年出生,湖南长沙人,馆员,在读硕士,研究方向为数据挖掘与智能决策。通信地址:410128 湖南省长沙市芙蓉区 湖南农业大学图书馆,E-mail:26712970@qq.com。
通讯作者:彭佳红,女,1962年出生,湖南江华人,教授,博士,研究方向为数据挖掘与智能决策。通信地址:410128 湖南省长沙市芙蓉区 湖南农业大学信息科学技术学院,E-mail:pjh719@163.com。
收稿日期:2018-03-26,**修回日期:**2018-05-18。

感。土壤肥力与土壤有机质和矿物质的含量存在显著的相关关系[1]。分析植烟土壤有机质和矿物质含量与烟草产量的关系,不仅能预测烟叶的产量为烟农提前预判烟草市场,还能对优质烟草的规模化区域种植起到重要参考作用[2-3]。

邵阳县是邵阳市烤烟生产第一大县,烤烟产业又是邵阳县四大支柱产业之一[4]。研究邵阳县烤烟种植区土壤养分与烟草产量的关系,预测邵阳县植烟区烟草产量有利于调控烟叶种植计划,促进当地农业耕地资源合理化利用。张慎等[5]对邵阳主产烟区植烟土壤环境进行了系统分析,并对邵阳植烟土壤重金属污染状况进行了测定,证明了邵阳植烟土壤有机质、有效磷等7个土壤养分含量处于中等到丰富水平,且土壤处于"无污染"状态,烟叶的安全性较好。李永富[6]通过采样分析了邵阳植烟区土壤有效锌含量的丰缺状况、空间分布和演变趋势,得出邵阳植烟区有效锌呈斑状分布的特点。邓小华等[7]分别采用传统统计学方法和地统计学方法分析了邵阳植烟区土壤有机质含量丰缺状况和有机质与土壤其他养分的定量关系,得出邵阳植烟区土壤有机质总体上适宜,土壤碱解氮、有效磷等养分函数与有机质含量存在显著正相关关系。上述研究分析了邵阳植烟区土壤养分含量的空间分布,并得到了邵阳植烟区适宜种植烟草的结论,但并没有深入定量分析土壤养分与烟草产量的关系,笔者在上述研究的基础上分析邵阳县植烟区土壤养分对烟草产量的决定性影响,并根据土壤养分分布情况对烟草产量进行定量预测。

1 材料与方法

1.1 数据来源

2016年采集了邵阳县70个植烟区土壤样本,通过对样本进行理化分析得出,邵阳县植烟区土地利用类型都为水田;地貌类型以丘陵为主(占比94.29%)、山地为辅(占比5.71%);母岩都为石灰岩,母质都为坡积物;耕层厚度20~32 cm;土壤土质主要可以分为壤质黏土、粉砂质黏土、粉砂质黏壤土、黏壤土、粉砂质壤土、壤土6类。测定了土壤样本中19个养分指标的含量,包括pH、碱解氮、有效磷、速效钾、全氮、全磷、全钾、有机质、有效铁、有效锰、有效铜、有效锌、有效硼、有效钼、有效硫、交换性钙、交换性镁、氯离子、阳离子交换量(X_1~X_{19})。其中有4个土壤样本中未检测出氯离子,为便于数据处理,将其值赋值为最小值0.01 mg/kg。此外,收集了这70个植烟区近年的烟草产量(Y)。

1.2 方法与步骤

1.2.1 主成分分析与SVR算法 主成分分析(Principal Component Analysis, PCA)[8]是把一组具有一定相关性的多个原始指标通过线性变换为另一组新的不相关的少数指标的特征降维方法,新的指标按照方差依次递减的顺序排列,排第一的指标称为第一主成分,排第二的指标称为第二主成分,依次类推;其中每个主成分都能反映原始指标的大部分信息,且所含信息互不重复[9]。利用主成分分析法可对样本特征进行降维处理,将冗余数据产生的噪声影响降低[10]。邵阳县植烟区土壤养分正交变换矩阵见表1。

支持向量机(Support Vector Machine, SVM)[11]是20世纪末才出现的热门机器学习方法,具有极佳的泛化能力。SVR(Support Vector Regression)[12]是支持向量机的回归模型,它非常适合解决非线性回归问题。SVR通过训练样本训练学习,调整参数,从而找到一个模型,使得所有训练样本的预测值与实际值偏差的总和最小,进而将预测样本输入上述模型进行实际预

表1 邵阳县植烟区土壤养分正交变换矩阵

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
第1主成分	0.3455	0.3514	-0.1602	-0.1489	0.3772	0.0133	-0.1828	0.3853	-0.1558	-0.0695	-0.0223
第2主成分	0.1578	-0.1963	-0.2431	0.0618	-0.2138	-0.1809	-0.0177	-0.1796	-0.3789	-0.1159	-0.4767
第3主成分	-0.0936	-0.0522	-0.4769	-0.3963	-0.0160	-0.2962	-0.0661	0.0037	-0.0172	0.1821	0.0903
第4主成分	-0.1031	-0.0679	-0.0273	0.2501	-0.1017	0.3398	-0.3683	-0.0097	-0.2135	0.5845	-0.2149
第5主成分	-0.1155	0.1729	0.0899	0.2042	0.0920	-0.3839	-0.1091	0.1381	-0.0290	-0.0308	-0.1853
第6主成分	-0.3488	0.2755	-0.1113	-0.3848	0.0834	0.35	-0.0947	0.1499	0.1439	-0.0952	-0.1617
第7主成分	0.0682	0.0924	-0.1279	0.0797	0.1350	0.1805	0.5765	0.0194	-0.1988	0.1224	-0.07291
第8主成分	0.0473	0.0864	-0.2487	-0.1846	-0.0586	-0.3819	0.3197	0.0284	-0.0380	0.4408	-0.0034
第9主成分	0.0005	-0.0756	0.0483	0.1029	-0.0069	0.1306	0.4148	0.0075	0.4178	0.1594	0.0756
第10主成分	0.0938	-0.1157	0.2055	-0.5669	-0.1335	0.2585	0.0049	-0.0885	-0.2493	-0.0709	-0.1488

续表1

	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}	X_{19}	贡献/%	累计/%
第1主成分	0.0010	-0.1011	-0.1165	-0.1031	0.319	0.2213	0.2345	0.3391	0.2771	0.2771
第2主成分	-0.4189	-0.0395	-0.3091	-0.2869	0.1100	-0.0229	-0.047	-0.0947	0.1707	0.4478
第3主成分	-0.0519	-0.4876	0.2399	0.1504	-0.1443	-0.2326	0.2438	-0.1228	0.1013	0.5492
第4主成分	-0.0213	-0.1263	-0.0349	0.3172	-0.1274	0.2568	0.0598	0.1498	0.0807	0.6298
第5主成分	-0.0502	0.3669	-0.2035	0.4935	0.0281	-0.3840	0.3281	-0.0777	0.0630	0.6928
第6主成分	-0.1911	0.0934	-0.3741	-0.0844	-0.4083	-0.1710	-0.1849	0.0541	0.0599	0.7528
第7主成分	-0.2976	0.0429	0.3023	0.2375	0.0544	-0.2880	-0.2479	0.3675	0.0478	0.8005
第8主成分	0.2689	0.4163	-0.1720	-0.0950	-0.1904	0.3311	-0.1212	0.0388	0.0406	0.8411
第9主成分	-0.3970	-0.0966	-0.2426	-0.0591	-0.0107	0.2029	0.5600	-0.0457	0.0375	0.8786
第10主成分	-0.0652	0.3924	0.3242	0.1119	0.0407	0.1236	0.3521	-0.1248	0.0251	0.9038

测[13]。它由以下几步组成：(1)参数寻优，以n-fold交叉验证方法找到最优的核函数参数，得到svmtrain方法；(2)将训练样本输入svmtrain方法进行训练，确立回归模型svmpredict；(3)将预测样本输入svmpredict方法进行实际预测。支持向量机有5种核函数，而径向基核函数被很多学者验证为最具有普遍适应性的核函数[14-15]。本研究也采用径向基核函数来进行SVR回归预测。

利用主成分分析算法和支持向量回归组合算法对邵阳县70个植烟区烟草产量进行回归预测，模型如图1所示。

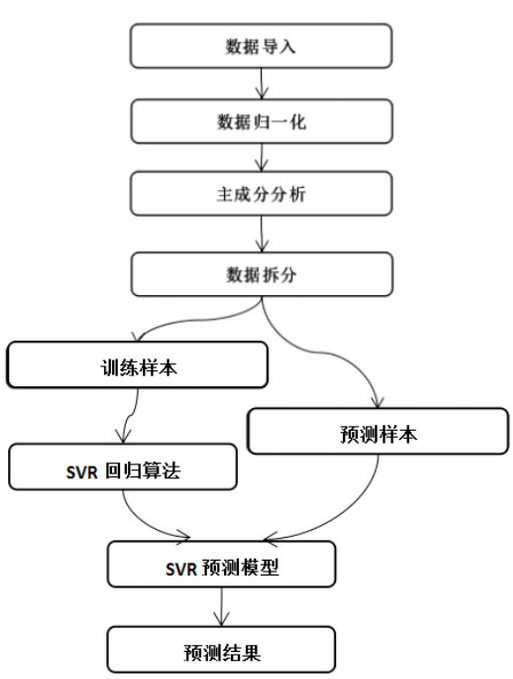

图1 PCA-SVR回归预测模型

1.2.2 数据导入 将70个土壤样本数据按X_1、X_2、…、X_{18}、X_{19}、Y的列排序方式导入表“data”作为初始数据。其中X_i为自变量特征，Y为因变量。

1.2.3 数据归一化 对样本数据中的19个特征列进行归一化处理，转换如式(1)。

$$Y=\frac{x-V_{\min}}{V_{\max}-V_{\min}} \quad \cdots\cdots (1)$$

式中，$V_{\max}$、$V_{\min}$分别为特征列中的最大值和最小值，x为初始值，y为归一后的值[16]。数据归一化的目的是降低样本数据间数量级差异对预测结果产生的噪声影响[17]。

1.2.4 数据拆分 对70个样本进行拆分，其中训练样本69个、预测样本1个。首先训练样本(既有自变量，也有因变量)输入SVR模型进行学习训练，得到SVR回归模型，然后再将预测样本(只有自变量)输入SVR回归模型进行回归预测，预测因变量。

1.2.5 评价标准 为验证模型的实际预测能力，对所有70个样本都进行实际预测。预测结果的优劣采用均方误差(mean squared error, *MSE*)作为度量指标，如式(2)。

$$MSE=\frac{\sum(y_i-y_i')}{n} \quad \cdots\cdots (2)$$

式中，y_i'为预测值，y_i为真值，n=70。*MSE*越小，说明预测模型的整体预测结果与实际值越接近[18]。

1.2.6 对比模型——随机森林回归模型 随机森林算法具有预测精度高、适合各种数据集、能处理维度较高的数据集、对于噪音数据具有一定的容错能力等优点[19]，也时常被用做回归预测[20-21]。为验证本算法的有效性，选取随机森林回归算法作为对比模型，同样对70个样本数据进行实际预测，然后对比2种算法的预测结果。

随机森林回归算法[22](random decision forests)指用随机的方式建立起包含多个没有关联的决策树森林。当一个预测样本输入随机森林之后,让森林中的每一棵决策树对样本的因变量进行预测,每一个决策树都会给样本一个预测结果,随机森林认为所有决策树预测结果的平均值就是样本的实际因变量值。随机森林中的每一棵决策树为二叉树,其生成遵循自顶向下的递归分裂原则,单棵决策树的算法可以选择c4.5、cart或id3[23]。

2 结果与分析

2.1 主成分分析结果

从主成分分析的正交变换矩阵(表1)可以看出,第1主成分受有机质的影响最大,贡献率达到38.53%;第2主成分受有效锌的影响最大,贡献率达到41.89%;第3主成分受有效硼的影响最大,贡献率达到48.76%;第4主成分收有效锰的影响最大,贡献率达到58.45%;第5主成分受有效硫的影响最大,贡献率达到49.35%;第6主成分受交换性钙的影响最大,贡献率达到40.83%;第7主成分受全钾的影响最大,贡献率达到57.65%;第8主成分受有效锰的影响最大,贡献率达到44.08%;第9主成分受有效铁的影响最大,贡献率达到41.78%;第10主成分受速效钾的影响最大,贡献率达到56.69%。其中第4主成分和第8主成分都受有效锰的影响最大,说明邵阳植烟区烟草产量主要受有机质、有效锌、有效硼、有效锰、有效硫、交换性钙、全钾、有效铁和速效钾这9个养分指标含量的影响。

2.2 SVR回归预测结果

表2是由19个养分指标作为样本特征,利用SVR预测的邵阳县植烟区烟草产量结果。表3是由主成分分析降维后9个养分指标作为样本特征,利用随机森林预测的邵阳县植烟区烟草产量结果。表4是由主成分分析降维后的9个养分指标作为样本特征,利用SVR预测的邵阳县植烟区烟草产量结果。

表2 邵阳县植烟区烟草产量SVR回归预测结果
(19个养分指标)

土壤样本	烟草产量实际值/(kg/hm²)	烟草产量预测值/(kg/hm²)	*MSE*
1	1950	1869	
2	2100	2163	
i	…...	…...	223.7
69	2025	2110	
70	2400	2091	

对比表2和表4发现,利用主成分分析对样本进行特征降维后的预测结果的*MSE*明显小于未经过特征降维的预测结果,这进一步说明邵阳植烟区烟草产量主要受有机质、有效锌、有效硼、有效锰、有效硫、交换性钙、全钾、有效铁和速效钾这9个养分含量的影响,其他10个养分指标对烟草产量预测会产生噪声干扰作用。

对比表3和表4发现,SVR预测结果的MSE明显小于随机森林的预测结果,这说明利用SVR对邵阳县植烟区的烟草产量进行预测比随机森林算法更有效。

表3 邵阳县植烟区烟草产量随机森林回归预测结果
(主成分分析后的9个养分指标)

土壤样本	烟草产量实际值/(kg/hm²)	烟草产量预测值/(kg/hm²)	*MSE*
1	1950	1869	
2	2100	2163	
i	…...	…...	186.3
69	2025	2110	
70	2400	2091	

表4 阳县植烟区烟草产量回归预测结果
(主成分分析后的9个养分指标)

土壤样本	烟草产量实际值/(kg/hm²)	烟草产量预测值/(kg/hm²)	*MSE*
1	1950	1869	
2	2100	2163	
i	…...	…...	55.5
69	2025	2110	
70	2400	2091	

3 结论

测定了邵阳县70个植烟区土壤样本中19个养分指标的含量,经过主成分分析降维得出了决定邵阳县植烟区烟草产量的9个影响因子,利用9个影响因子结合SVR算法对邵阳植烟区烟草产量进行回归预测。通过纵向和横向对比发现,SVR算法在邵阳县植烟区是有效的烟草产量预测方法。即只需测定土壤中9个养分指标的含量就能通过主成分分析和SVR算法预测烟草的产量,为烟草产量的预测提供了一条新的思路,也为邵阳植烟区烟农进行土壤合理利用和施肥提供了重要参考。

4 讨论

中国是世界上最大的烟草生产国和消费国。烟草产量的精准预测对于烟草生产和消费都有重大指导意义。国内一些学者对国内烟草产量预测进行了研究。曾志三等[24]以福建省宁化县为例，通过分析该县1996—2004年烟草产量数据，建立基于灰色系统的烟草产量预测GM(1,1)模型。王红影等[25]以中国1981—2014年的烟草产量数据为样本，基于ARMA模型对烟草产量进行了预测。上述研究局限于只利用历年烟草产量预测未来烟草产量，未考虑土壤养分含量等外在因素对烟草产量的影响，且预测方法都是线性模型，说服力不足。

刘晓宇[26]考虑了以氮、钾、磷肥的施肥量作为烟草产量的影响因素，通过MATLAB工具箱建立了黑龙江省烟草产量多元二次回归模型，并对2009年、2010年的烟草产量进行了实际预测，但施肥量并非土壤肥力的单一来源，不能完全代表土壤养分含量对烟草产量的影响，且该预测方法还是局限于线性回归模型。

烟草产量受土壤养分含量的直接影响。土壤养分包括有机质、氮、磷、钾、钙、镁 、硫、铁、硼、钼、锌、锰、铜和氯等多种元素[27]，每种养分都对烟草产量具有一定的影响，但这些元素之间又存在很大的相关性，所以，本研究最初尽可能多地测定了土壤中的19个养分指标含量作为影响烟草产量的特征集，然后采用主成分分析法消除特征集间的相关性、冗余性，得到了影响烟草产量的包含9个养分指标的特征子集。既考虑了土壤养分对烟草产量的全面性影响，又考虑了土壤养分对烟草产量的精准性影响。

烟草产量与土壤养分含量间的关系并不是简单的线性关系，而是复杂的非线性关系，但国内利用非线性方法预测烟草产量的研究并不多见。笔者选择的支持向量机模型对于处理非线性回归问题非常擅长，但作为烟草产量的预测方法在国内应属首次，对于烟草产量预测领域的研究具有一定参照意义和对比意义。

虽然本研究以邵阳县植烟区烟草产量为例进行分析预测，但提出的基于主成分和支持向量机的烟草产量预测方法同样可以推广应用到其他地区的烟草产量预测。

参考文献

[1] 金亚波,李桂湘,韦建玉,等.云南玉溪植烟区气候-土壤因子聚类分析[J].土壤通报,2010,41(2):275-281.

[2] 李强,周冀衡,张一扬,等.基于地统计学的曲靖植烟土壤主要养分丰缺评价[J].烟草农学,2012,11:69-73.

[3] 邹娟,鲁剑巍,周先竹,等.湖北省主要植烟区土壤肥力状况及分析[J].长江流域资源与环境,2015,24(3):504-509.

[4] 邹凯,肖钦之.邵阳烟区气象因素与烤烟化学因子相关性分析[J].湖南农业科学,2017,9:24-27.

[5] 张慎,刘建丰,于庆涛.邵阳主产烟区植烟土壤特征及安全性评价[J].作物研究,2014(s1):817-820.

[6] 李永富.湖南省邵阳烟区土壤有效锌含量时空特征及其影响因素[J].中国烟草学报,2015,21(1):53-58.

[7] 邓小华,邓井青,宾波,等.邵阳植烟土壤有机质含量时空特征及其他土壤养分的关系[J].烟草科技,2014,6:82-86.

[8] 李航.统计学习方法[M].北京:清华大学出版社,2012:41-43.

[9] Moore B. Principal component analysis in linear systems: Controllability, observability, and model reduction[J].IEEE Transactions on Automatic Control,2003,26(1):17-32.

[10] 李靖华,郭耀煌.主成分分析用于多指标评价的方法研究——主成分评价[J].管理工程学报,2002,16(1):39-43.

[11] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York: Spring Verlag Press,1995:11-23.

[12] Li Y, Shao X, Cai W. A consensus least squares support vector regression (LS-SVR) for analysis of near-infrared spectra of plant samples[J].Talanta,2007,72(1):217-22.

[13] 王笑冰,张红燕,谢元瑰,等.基于GS-SVR的耕地面积预测及其驱动因子分析[J].中国农学通报,2013,29(23):210-215.

[14] 袁哲明,张永生,熊洁仪.基于SVR的多维时间序列分析及其在农业科学中的应用[J].中国农业科学,2008,41(8):2485-2492.

[15] 杨柳,刘艳芳.将微粒群和支持向量机用于耕地驱动因子选择的研究[J].武汉大学学报:信息科学版,2010,35(2):248-251.

[16] 李星,陈渊,张永生,等.基于支持向量回归与地统计学的多维时间序列分析[J].中国农学通报,2011,27(29):133-138.

[17] 汤荣志,段会川,孙海涛.SVM训练数据归一化研究[J].山东师范大学学报:自然科学版,2016,12(15):114-117.

[18] Breiman L. Random forests[J].Machine learning,2001,45(1):5-32.

[19] 曹正凤.随机森林算法优化研究[D].北京:首都经济贸易大学,2014.

[20] 崔东文.随机森林回归模型及其在污水排放量预测中的应用[J].供水技术,2014,8(1):31-36.

[21] 顾娟,林敏,鞠桂玲.基于随机森林回归的军械器材需求预测[J].自动化应用,2017(9).

[22] Ho, Tin Kam. Random Decision Forest[A].In: Proceedings of the 3rd International Conference on Document Analysis and Recognition[C].Montreal, Canada,1995:278-282.

[23] 徐鹏,林森.基于C4.5决策树的流量分类方法[J].软件学报,2009,20(10):2692-2704.

[24] 曾志三,顾明.GM(1,1)模型对烟草产量的灰色预测[J].山地农业生物学报,2006,25(4):293-296.

[25] 王红影,马成文.基于ARMA模型的中国烟草产量预测[J].齐齐哈尔工程学院学报,2016(3):62-65.

[26] 刘晓宇.黑龙江烟草产量预测及病害预警方法研究[D].哈尔滨:东北农业大学,2012.

[27] 胡克林,陈德立.农田土壤养分的空间变异性特征[J].农业工程学报,1999,15(3):33-38.