基于随机森林算法的农产品产量影响因素权重分析 *

胡新祥,赵 霞△,张 乾,李英兰,孔祥盛,马玉婷

(甘肃农业大学 信息科学技术学院,甘肃 兰州 730070)

摘 要:现代化的农业离不开计算机技术的应用。随着计算机技术的飞速发展,当今社会已经进入了大数据时代。大量的数据中存在巨大的价值可以挖掘,数据既是商业资源也是战略资源。面对庞大杂乱的数据资源,分析和利用就显得尤为重要。作为农业大国,中国的农业数据的分析更是国家之命脉,有极大的利用价值。研究目标为寻找一种方法分析不同因素对农产品产量的影响程度,并在图表中直观得呈现出分析结果。实验表明,随机森林算法在农产品产量影响因素的权重分析上有较好的表现,准确率较高。

关键词:农产品;影响因素;决策树;随机森林;机器学习

中图分类号:F324.1

1 概述

几千年来,中国劳动人民过着“靠天收”的生活。农民们根据长久以来的经验总结出了在农耕中各种方式方法。但是这种依靠经验的方法往往会因为一些特殊的因素而受到影响。一旦出现意外,对农户和社会带来的损失可能是不可估量的。进入了新时代,我们可以尝试使用现代技术来对这些影响农作物产量的因素进行科学的分析,让人们更加了解这些因素在农作物产量起到的作用,进而制定出科学的策略来应对一些不可控现象的发生。这既符合大环境趋势,也让理论研究真正的应用到实际社会生产生活之中。在信息时代,计算机技术能够为农产品产量的预测提供更多、更有效的预测方式。利用计算机技术的快速性,国内外的研究者将计算机技术运用到中国农业经济预测的过程中,通过建立相关农产品产量的预测系统,更精确的预测中国农产品产量的变化趋势。

近年来,深度学习等人工智能技术得到了迅速发展,在很多领域都取得了较好的应用效果。其中分类算法在数据挖掘方面应用最为广泛。

常用分类算法有:典型的朴素贝叶斯方法,针对大量数据训练速度较快,并支持增量式训练,对结果的解释便于理解,但在大数据集下才能获得较为准确的分类结果,且忽略了数据各属性值之间的关联性[1];K-最近邻分类算法比较简单,训练过程迅速,抗噪声能力强,新的数据能够直接参与训练集而不需要再次训练,但在样本不平衡时结果偏差较大,且每次分类都需要重新进行一次全局运算[2];决策树分类算法易于理解与解释,可进行可视化分析,运行速度较快,可扩展应用于大型数据库中,但容易出现过拟合问题,且易忽略数据属性间的关联性[3]。

随机森林算法在分类方面表现突出,其避免了决策树分类算法中容易出现的过拟合问题,并在运算量未显著提高的前提下,提高了分类准确率[4]。因此,设计旨在利用随机森林算法实现精准客观且省时省力的分析。

2 研究背景与目的

2.1 研究背景

年甘肃省主要农作物:玉米、高粱、马铃薯、棉花与油料的产量与 10 年间甘肃省各年年均太阳辐射量、年均气温与年均降水量之间的关系。

* **基金项目**:甘肃农业大学学生科研训练计划(SRTP)项目(项目编号:201916056);甘肃省科技厅 2018 年自然科学基金项目(项目编号:18JR3RA79)。

△**通讯作者**:赵霞(1979–),女,副教授,主要从事数据分析与计算机技术的教学与研究。

2.2　研究目的

以 10 年间甘肃省的各年度气象数据为条件,结合产量分析出各种气象因素对不同农作物产量的影响程度。采用 python 语言作为分析工具,采用随机森林算法对数据进行处理与分析。最后得出每一种农作物的产量受各种气候条件影响程度的大小,并用图表的形式直观展现,作为农业生产活动的参考指标。

3　数据的采集与处理

3.1　数据的收集

选取甘肃省 2000 年~2010 年十年间的各类典型农产品产量与各年的年均降水量、年均气温与年均太阳辐射量,数据均来自国家统计局官网。对数据进行整理后在 python 程序中读取并制表,见表 1。

表 1　2000 年~2010 年的数据

	年份	小麦单位面积产量(kg/m²)	玉米单位面积产量(kg/m²)	高粱单位面积产量(kg/m²)	马铃薯单位面积产量(kg/m²)	棉花单位面积产量(kg/m²)	油料单位面积产量(kg/m²)	太阳辐射量(KW.h/m²)	年均温(℃)	年降水量(mm)
0	2000	2232.03	4532.47	3021.12	2517.02	1656.30	1335.36	497.89	14.42	922.410
1	2001	2634.37	4260.57	3345.78	3032.02	1739.53	1176.50	498.66	13.98	957.910
2	2002	2890.08	4352.73	3705.48	2919.78	1721.60	1348.96	503.88	13.97	978.560
3	2003	2834.17	4983.59	4205.10	3022.05	1659.91	1374.26	498.94	13.54	1254.780
4	2004	2917.19	5023.89	4005.15	3110.32	1609.32	1458.12	524.14	13.75	864.280
5	2005	2646.20	5126.14	4171.75	3575.50	1728.16	1529.46	532.04	14.04	962.150
6	2006	2719.76	4222.36	4175.59	3240.57	1677.53	1507.84	532.04	14.37	912.610
7	2007	2417.40	4931.91	4723.29	3141.24	1632.90	1520.20	513.23	13.80	1219.770
8	2008	2967.25	4763.02	4562.12	3264.37	1693.68	1614.12	519.31	14.44	929.130
9	2009	2708.90	4752.20	5083.33	2974.13	1714.15	1663.82	525.62	14.28	907.070
10	2010	2852.27	4672.76	5043.56	2869.32	1577.79	1852.82	473.89	13.65	931.975

3.2　数据预处理

读取数据以后利用 python 对所得数据进行一些预处理动作,目的是为了观察数据是否存在缺失情况与离群数据。都缺失数据与离群数据要进行相应的处理。

首先将各年度的年均气温、降水量与太阳辐射量绘制在二维柱状图中进行观察。

观察 10 年间度甘肃省年均气温的直方图(图 1),数据基本分布在 12℃~14℃左右,无缺失数据与离群数据。

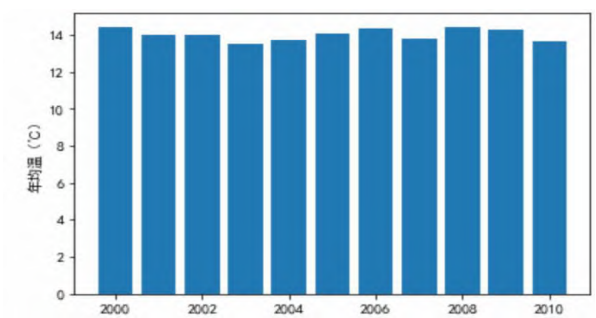

图 1　2000 年~2010 年年均气温柱状图

观察 10 年间度甘肃省年均降水量的直方图(图 2),数据基本分布在 800~1000mm 左右,2002 年与 2007 年降水量有明显增多,无缺失数据。

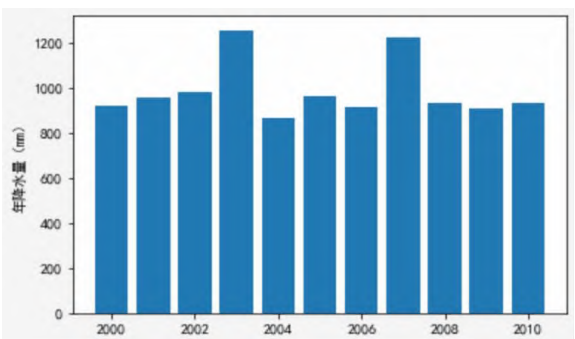

图 2　2000 年~2010 年年均降水量柱状图

观察 10 年间度甘肃省年均太阳辐射量的直方图(图 3),数据基本分布在 500KW.h/m² 左右,无缺失数据与离群数据。

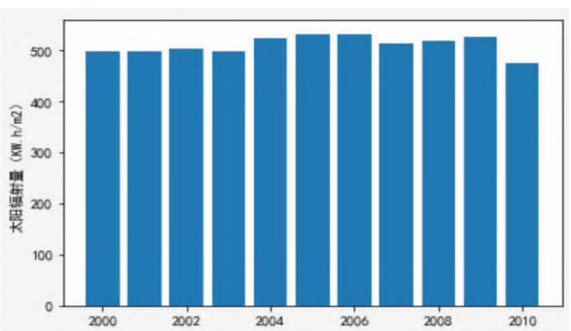

图 3　2000 年~2010 年年均太阳辐射量柱状图

随后将 10 年间甘肃省各类主要农作物的年均产量利用箱型图直观的展现出来(图 4),观察是否有缺失数据与离群数据。

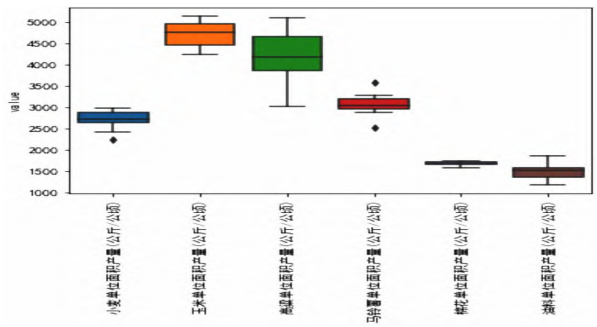

图 4 2000 年~2010 年农作物产量箱型图

3.3 数据的分析

在对数据进行图表直观的分析以后,开始对收集到的数据进行进一步的分析,利用 python 中的 pd.describe()函数对十年间农产品产量与环境量进行计算分析,其意义在于观察这一系列数据的范围。大小、波动趋势等等,便于判断后续对数据采取哪类模型更合适。计算结果见表 2,count 为计数值,mean 为平均值,std 为标准差,min 为最小值,25%为下四分位,50%为中位数,75 为上四分位数,max 为最大值。

表 2 pd.describe()函数对数据处理结果

	小麦单位面积产量 (kg/m²)	玉米单位面积产量 (kg/m²)	高粱单位面积产量 (kg/m²)	马铃薯单位面积产量 (kg/m²)	棉花单位面积产量 (kg/m²)	油料单位面积产量 (kg/m²)	太阳辐射量 (KW.h/m²)	年均温 (℃)	年降水量 (mm)
count	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000
mean	2710.874545	4692.876364	4185.660909	3060.574545	1673.715455	1489.223636	510.876364	14.021818	985.513182
std	223.861895	315.830205	655.587775	266.399289	52.100899	183.449838	18.008514	0.319306	128.425114
min	2232.030000	4222.360000	3021.120000	2517.020000	1577.790000	1176.500000	473.890000	13.540000	864.280000
25%	2640.285000	4442.600000	3855.315000	2946.955000	1644.60000	1361.610000	498.800000	13.775000	917.510000
50%	2719.760000	4752.200000	4175.590000	3032.020000	1677.530000	1507.840000	513.230000	13.980000	931.975000
75%	2871.175000	4957.750000	4642.705000	3190.905000	1717.875000	1571.790000	524.880000	14.325000	970.355000
max	2967.250000	5126.140000	5083.330000	3575.500000	1739.530000	1852.820000	532.040000	14.440000	1254.780000

从分析的结果来看,收集到的各项数据质量较好,都在各自的范围内波动,且无缺失情况。利用这些数据就可以进入到各种环境对产量影响程度的探索阶段。

4 查看相关性

在对收集到的数据进行预处理以后,进入数据相关性的分析工作中。在使用分类算法分析之前,利用 python 中 numpy triu_indices 函数制作数据矩阵,利用 seaborn 绘制数据热力图。这一动作的目的是初步查看各组数据之间的相关性,使用热力图可以更加直观的展现出来,如图 5 所示。

从得到的热力图中可以直观观察到各种农作物与各环境变量之间的相关程度。由图可初步得出:小麦的每公顷产量受太阳辐射量影响程度最大,年均温与年均降水量对其影响程度相当,但次于太阳辐射量的影响程度;玉米与高粱每公顷产量受太阳辐射量与年均降水量的影响程度较大,受年均温的影响程度较小;棉花每公顷产量受太阳辐射量与年均温的影响程度较大,受年均降水量的影响程度较小;三种环境对油料的产量影响程度相当。

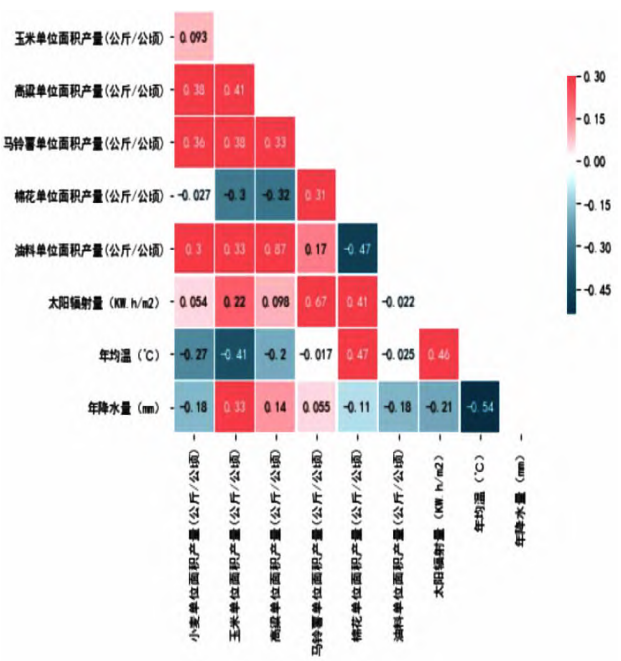

图 5 数据相关性热力图

初步查看到各环境与农作物之间的相关性后,选择一种合适的分类算法对数据进行更加深入的分析,得到各个环境变量对作物产量影响程度的具体权重。

5　随机森林

5.1　决策树

决策树作为随机森林的基分类器,是一种十分常用的分类方法。决策树分类思想实际上是一个数据挖掘过程,其通过产生一系列规则,然后基于这些规则进行数据分析[5]。决策树采用单一决策方式,因此具有以下缺点:一是包含复杂的分类规则,一般需要决策树事前剪枝或事后剪枝;二是收敛过程中容易出现局部最优解;三是因决策树过于复杂,容易出现过拟合问题。为了解决这些缺点,又引入随机森林的概念。

5.2　随机森林

随机森林中的决策树按照一定精度进行分类,最后所有决策树参与投票决定最终分类结果,这是随机森林的核心概念。

随机森林构建主要包括以下 3 个步骤:

1)为 N 棵决策树抽样产生 N 个训练集。每一棵决策树都对应一个训练集,主要采用 Bagging 抽样方法从原始数据集中产生 N 个训练子集。Bagging 抽样方法是无权重的随机有放回抽样,在每次抽取样本时,原数据集大小不变,但在提取的样本集中会有一些重复,以避免随机森林决策树中出现局部最优解问题。

2)决策树构建。该算法为每个训练子集构造单独的决策树,最终形成 N 棵决策树以形成“森林”。节点分裂原则一般采用 CART 算法或 C4.5 算法,在随机森林算法中,并非所有属性都参与节点分裂指标计算,而是在所有属性中随机选择某几个属性,选中的属性个数称为随机特征变量。随机特征变量的引入是为了使每棵决策树相互独立,减少彼此之间的关联性,同时提升每棵决策树的分类准确性,从而提高整个森林的性能。

3)森林形成及算法执行。重复步骤(1)、(2),构建大量决策树,形成随机森林。算法最终输出由多数投票方法实现。将测试集样本输入随机构建的 N 棵决策子树进行分类,总结每棵决策树分类结果,并将具有最大投票数的分类结果作为算法最终输出结果。如图 6 所示。

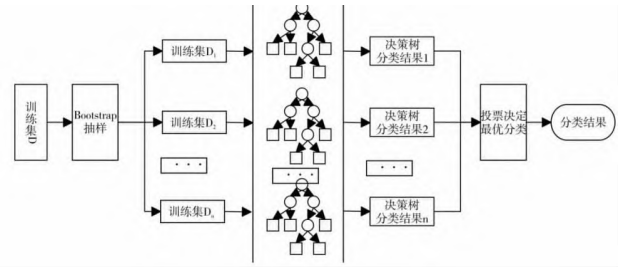

图 6　随机森林算法原理图

5.3　使用随机森林

在程序中构造随机森林模型实现使用随机森林算法对已有数据进行分析,并对得出的果绘制农作物的影响程度的表格,见表 3。

表 3　各个因数影响农作物的程度情况表

	太阳辐射量(KWh/m²)	年均温(℃)	年降水量(mm)
小麦单位面积产量(kg/m²)	0.425988	0.327842	0.246170
玉米单位面积产量(kg/m²)	0.383898	0.431007	0.185095
高粱单位面积产量(kg/m²)	0.558349	0.320426	0.121225
马铃薯单位面积产量(kg/m²)	0.701089	0.155311	0.143600
棉花单位面积产量(kg/m²)	0.338979	0.612493	0.048528
油料单位面积产量(kg/m²)	0.761373	0.195005	0.043622

6　实验分析

由随机森林算法得出的最后结果可以观察到,在此模型中,太阳辐射量、年均气温、年均降水量对小麦单位面积产量的影响程度分别为:0.425988 0.327842 0.246170;对玉米单位面积产量的影响程度分别为:0.383898 0.431007 0.185095;对高粱单位面积产量的影响程度分别为:0.558349 0.320426 0.121225;对马铃薯单位面积产量的影响程度分别为:0.701089 0.155311 0.143600;对棉花单位面积产量的影响程度分别为:0.338979 0.612493 0.048528;对油料单位面积产量的影响程度分别为:0.761373 0.195005 0.043622。

得出的结论与初步查看相关性时,从热力图中的到的大致相关性相吻合。　　(下转第 40 页)

件的要求;施工过程一般工业固体废物和危险废物贮存场地的位置、数量、尺寸、规模、建设标准是否满足环评报告的要求;试运行阶段一般工业固体废物和危险废物贮存场地是否按期建设完成并投入使用[3]。

3.3 环保达标环境监理要点

3.3.1 施工场界达标排放

重点关注施工场界的粉尘、二氧化硫、氮氧化物,施工场界噪声,建筑废料、施工弃渣和施工营地生活垃圾等污染物的排放和处理,是否满足环评报告中提出的建筑施工场界污染物排放标准和污染控制标准。

3.3.2 排放口及厂界达标排放

在各项环保设施严格按照相关技术规范、设计文件、环评文件建设完成,并调试运行正常的情况下,开展污染治理设施排放口和厂界环境监测,重点关注各治理设施废气排放口的废气,各污水处理单元废水排放口的废水,厂界无组织排放废气和厂界噪声,固体废弃物暂时收集贮存场地是否满足环评文件中提出的相应的行业排放标准、综合排放标准和其他控制标准。

3.4 环境风险防范措施环境监理要点

煤化工建设项目环境风险防范措施重点关注污水处理站应急事故池、消防事故废水收集池、初期雨水收集池的位置、应急管网、尺寸、容积是否满足环评报告的要求;各储罐区围堰和围堤的尺寸、规模是否符合环评文件的要求;在各车间及工段内部、煤堆场、主作业区、储罐区、污水处理单元、应急事故池、各类管网的防腐防渗措施是否符合行业技术规范、环评文件的要求[4]。

4 结束语

煤化工建设项目环境监理是为建设单位提供全过程环境监管的专业化环境监理服务,可避免在前期设计阶段、建设实施阶段和试运营阶段,出现批建不符、环保“三同时”制度落实不利、环保不达标等问题,导致项目不能如期竣工验收,并受到相应的处罚。环境监理人员严格按照环保法律法规、技术规范、环境标准及环评报告等,开展环保措施和配套的环保治理设施的全过程监督和管理,对减轻煤化工建设项目实施过程中造成的环境污染有举足轻重的作用。

参考文献:

[1] 蔡同锋.大型石化项目环境监理实践及工作方法探讨[J].环境监控与预警,2015(4):52-56.

[2] 杨凯,朱庚富,胡耘.生活垃圾焚烧发电项目环境监理要点分析[J].山西建筑,北京:人民交通出版社,2013(33):188-189.

[3] 王少斌,张树深.一般工业固废填埋场建设项目环境监理现状与实践[J].环境保护与循环经济,2010(6):38-41.

[4] 苟德国.环境监理在医药化工项目中的应用实践-以防渗设计环境监理审查要点分析为例[J].资源节约与环保,2016(4):107-107+111.

(上接第19页)

说明结论准确可信。同时也验证了随机森林算法在对农产品产量影响因素权重分析中的应用的正确性与有效性。

7 结语

在此次实验中,通过收集到的甘肃省10年间环境变量与主要农作物产量的数据,在进行了数据的预处理与简单的查看相关性后,选择使用随机森林算法模型对一系列数据进行了科学、客观的分析。最后得到了太阳辐射量、年均温、年均降水量对甘肃省六种主要农作物影响程度的具体权重,得到的结果与现实相吻合,且用数据具体的说明的不同环境变量对不同作物的具体影响程度。这一结果在监督算法的保证下真实有效,可以作为农业生产活动的参考指标之一。

参考文献:

[1] LEWIS D D. Naive (Bayes)at forty:The independence assumption in information retrieva[1 C]. European Conference on Machine Learning,1998.

[2] TANG Q Y,ZHANG C X. Data Processing System(DPS) software with experimental design,statistical analysis and data mining developed for use in entomological research [J]. 中国昆虫科学:英文版, 2013,20(2):254-260.

[3] ROMERO C,VENTURA S. Educational data mining:a survey from 1995 to 2005 [J]. Expert Systems with Applications,2007,33(1):135-146.

[4] SVETNIK V,LIAW A,TONG C,et al. Random forest:a classification and regression tool for compound classification and QSAR modelng[J]. Journal of Chemical Information & Computer Sciences, 2003,43(6):1947.

[5] 张琳,陈燕,李桃迎,等. 决策树分类算法研究[J]. 计算机工程,2011,37(13):66-67.