

文章编号: 1674-7054(2019)04-0380-07

基于大数据挖掘的橡胶园耕地地力评价

彭金莲¹, 唐朝胜¹, 蒋菊生², 赵春梅³

(1. 海南大学 计算机与网络空间安全学院, 海口 570228; 2. 海南省农垦科学院, 海口 570206;
3. 中国热带农业科学院 橡胶研究所 海口 571101)

摘要: 为了客观评价海南垦区胶园耕地的地力, 笔者使用 K-Means, Two-Step, Kohonen 和 TwoStep-AS 等 4 种聚类算法, 对海南农垦胶园测土调查数据进行大数据分析, 根据轮廓系数(Silhouette Coefficient)的大小, 确定 TwoStep-AS 聚类模型对海南农垦胶园耕地地力指标进行聚类分析, 经过反复迭代, 最终将海南农垦胶园耕地划分为 3 种类型, 并筛选出 5 个对海南垦区胶园耕地类型有影响的 5 个地力指标, 即年平均降雨量、有效土层厚度、全氮、北纬和成土母质。

关键词: 橡胶园; 大数据挖掘; 地力评价; 轮廓系数

中图分类号: S-3

文献标志码: A

DOI: 10.15886/j.cnki.rdsxb.2019.04.012

我国耕地地力评价主要方法有模糊数学法、层次分析法、指数和法等, 我国第 2 次土壤普查多采用指数和法评定耕地地力等级。近年来, 陈桂芬等^[1]应用 K-means 聚类方法、Johnson 粗糙集属性约简算法与 C4.5 决策树算法优化地力评价。蔡丽霞等^[2]应用 C4.5 决策树、K-means 和 DBSCAN 聚类算法, 对土壤养分数据进行对比分析, 依靠大量土壤肥力状况以预测未来土壤肥力的变化趋势。闫一凡^[3]认为评价指标筛选目前尚无国际通用的标准, 中国已建立了包括 7 类共 64 项指标的公用指标体系; 当前主流的地力评价方法(如中国农业部推荐的特尔斐一层次分析法等)仍存在人为随意性等不足, 推荐将不同评价方法(特别是特尔斐一层次分析法和分类与回归树模型法的组合)组合起来用于实际工作, 可取长补短, 进一步提高结果的可靠性。建议建立起统一的地力等级定义, 引入分类与回归树、人工智能、机器学习等新方法, 并将其与现有评价方法(如特尔斐法、多元回归等)集成, 充分挖掘数据、实现优势互补, 认为是未来耕地地力评价研究的重点和发展趋势^[4-7]。海南农垦测土配方施肥项目从 2006 年开始试点, 历时 10 年, 实施范围为海南农垦原有的 92 个农场(分 10 个片区), 面积大约 2 万 hm^2 , 依据测土配方调查数据进行了橡胶园耕地地力评价。采用了专家打分法和特尔斐法, 10 个片区分别建立了耕地地力评价指标体系, 10 个片区采用的耕地地力评价指标不尽相同, 耕地地力评价指标的选择与重要度由专家打分确定, 人为主观影响明显, 无法对海南垦区胶园耕地地力进行统一评价。笔者对海南垦区胶园大数据进行挖掘, 旨在筛选出对海南胶园耕地类型有影响的 5 个地力指标, 以便客观评价海南垦区胶园的耕地地力。

1 研究对象与技术方法

1.1 研究对象 本研究通过整合海南农垦测土配方施肥 10 个片区(项目场)的属性数据库、空间数据库和测土配方施肥数据库, 创建全垦区统一的耕地地力评价数据库; 对海南垦区测土配方施肥原始数据进行数据探索, 通过数据分析, 探索土壤内在的分类规律, 筛选出与垦区耕地地力相关的主要因素, 以确定海南垦区的耕地地力指标; 使用 4 种聚类模型, 利用轮廓系统进行适用性评估, 最终确定一种聚类算法作为耕地地力分类模型来划分胶园耕地类型, 从而实现对海南橡胶园耕地地力进行统一的综合评价。

收稿日期: 2018-10-19

修回日期: 2019-08-15

基金项目: 海南省自然科学基金项目(20156246)

作者简介: 彭金莲(1962-), 女, 教授。研究方向: 数据库与数据挖掘。E-mail: jl_peng@163.com

1.2 技术方法 使用 IBM 公司开发的企业级大型数据挖掘工具 SPSS Modeler 进行数据挖掘,技术路线参照跨行业数据挖掘标准流程 CRISP-DM 模型设计而成,此数据挖掘标准流程由问题理解、数据理解、数据准备、建立模型、模型评估和成果发布等 6 个阶段构成。

1.2.1 分类型指标的数据预处理 通过对原始数据的审核,发现能够作为地力评价潜在的分类指标为:地貌类型、地形部位、坡向、成土母质和质地等。但由于原始数据时间跨度大、涉及人员多,许多数据没有严格按照操作规程填写,导致各指标的类别数超出规程规定,产生了地貌类型 71 类,地形部位 95 类,坡向 71 类,成土母质 41 类,质地 64 类,所以数据质量较差,因此,必须按照农业部测土配方施肥野外调查操作规程要求调整。其中,地形部位原来只有其他部位和山腰,已经没有数据挖掘价值。调整前后各指标及其类别值见表 1。调整后,地貌类型调整为 3 种类别,分别是山地、丘陵、平地。坡向调整为 9 种类别,分别是北、东北、东、东南、南、西南、西、西北和平地等。成土母质调整为 8 种类别,分别是安山岩、变质岩、花岗岩、浅海沉积物、砂页岩、石灰岩、玄武岩和云母片岩风化物。质地调整为 6 种类别,分别是粘土、粘壤土、中壤土、砂壤土、砂土和其他等。分类型地力指标的缺失值处理依据不同情况进行了不同的处理。地形部位只有 4 个片区农场有调查数据,缺失值占 74%,已经没有数据挖掘价值。地貌类型的缺失值为 11.44%,只有 1 个片区农场没有调查此类数据,通过 C&RT 模型进行缺失值插补。质地指标 60% 是其他类别,已经没有分类价值。

表 1 地力评价各分类指标调查数据调整前后类别数
Tab.1 Number of various indexes of survey data before and after adjustment

项目 Items	地貌类型 Landform types	地形部位 Landform position	坡向 Aspect	成土母质 Soil parent material	质地 Soil texture
调整前类别数 Number of indexes before adjustment	71	95	71	41	64
合并调整后类别数 Number of indexes after adjustment	3	丢弃 Abandoned	9	8	6

1.2.2 连续型地力指标的预处理 连续型地力指标的预处理相对分类型指标来说要复杂很多,包括离群值、极值、缺失值的数据预处理,也包括数据分布和数据分箱处理。连续型地力指标共有 10 个,分别是 pH 值、有机质、全氮、有效磷、速效钾、北纬、东经、海拔、常年平均降雨量和有效土层厚度。从描述性统计分析,查看均值、标准差、最大值和最小值。一般情况下,若标准差远远大于均值,可粗略判定数据存在异常值。采用 3 倍标准差上最大值替换离群值和丢弃极值的方法。为了提高数据挖掘效率,基于对连续型地力指标进行离散化处理和适应某些模型只能处理分类型变量的要求,减少变量取值数以实现样本量的缩减。参考《海南岛土系概论》^[8],对海南农垦胶园土壤肥力的养分指标进行分组,如表 2 所示,分组给出的区间范围是大于左边的区间数和小于右边的区间数。在 SPSS Modeler 中对连续数据分组的方法有固定宽度、分位数、等级、平均值/方差和最优等 5 种方法,根据调查样本的数据分布,确定海拔高度和有效土层厚度采用样本量固定宽度分级,年平均降雨量、东经和北纬采用数值固定宽度分组,各立地指标分组标准如表 3 所示,分组给出的区间范围是大于左边的区间数和小于右边的区间数。

表 2 土壤养分指标分组
Tab.2 Grouping of indexes of soil nutrients

分组 Group	有机质/(g·kg ⁻¹) Organic matter	全氮/(g·kg ⁻¹) Total nitrogen)	有效磷/(g·kg ⁻¹) Available phosphorus	速效钾/(g·kg ⁻¹) Available potassium	pH
1	> 30	> 1.10	> 40	> 200	< 3.50
2	25 ~ 30	0.85 ~ 1.10	20 ~ 40	150 ~ 200	3.50 ~ 4.5
3	20 ~ 25	0.65 ~ 0.85	10 ~ 20	100 ~ 150	4.50 ~ 5.50
4	15 ~ 20	0.50 ~ 0.65	5 ~ 10	50 ~ 100	5.50 ~ 6.50
5	10 ~ 15	0.10 ~ 0.50	3 ~ 5	30 ~ 50	6.50 ~ 7.50
6	5 ~ 10	0.05 ~ 0.10	1 ~ 3	20 ~ 30	7.50 ~ 8.50
7	< 5	< 0.05	< 1	< 20	> 8.50

表 3 立地指标分组标准表
Tab. 3 Grouping of site indexes

分组 Group	年平均降水量/mm Mean annual precipitation	纬度/° Latitude	经度/° Longitude	有效土层厚度/cm Effective soil layer thickness	海拔/m Altitude
1	< 1 500	18.19 ~ 18.64	108.72 ~ 109.28	40 ~ 55	0 ~ 45
2	1 500 ~ 1 600	18.64 ~ 19.09	109.28 ~ 109.85	55 ~ 70	45 ~ 90
3	1 600 ~ 1 700	19.09 ~ 19.55	109.85 ~ 110.42	70 ~ 85	90 ~ 135
4	1 700 ~ 1 800	19.55 ~ 20.00	110.42 ~ 110.98	85 ~ 100	135 ~ 180
5	1 800 ~ 1 900			> 100	180 ~ 225
6	1 900 ~ 2 000				225 ~ 270
7	> 2 000				> 270

1.2.3 聚类模型算法 聚类分析将一批样本数据,在没有先验知识的前提下,根据数据的诸多特征,按照其在特征上亲疏自动分组,使得组内个体的结构特征具有较大的相似性,组间个体特征相似性较低。聚类与分类不同,聚类分析的输入数据集是一组未标记的对象,也就是说此时输入的对象还没有被进行任何分类,聚类的目的是根据一定的规则合理地进行分组或聚类,并用显式或隐式的方法描述不同的类别。由于分析可以采用不同的算法,所以对于相同的数据集合可能有不同的划分。从聚类的原理角度来看,聚类的算法可以分为:划分聚类、层次聚类、基于密度的聚类和基于网格聚类等方法。本研究分别使用 K-Means, Two-Step, Kohonen 和 TwoStep-AS 等 4 种聚类算法,对海南农垦测土调查数据进行聚类分析,最终,根据轮廓系数(Silhouette Coefficient)的大小确定 TwoStep-AS 聚类模型为海南垦区耕地进行分类。

1.2.3.1 TwoStep-AS 聚类模型算法介绍^[9] SPSS Modeler 中的 TwoStep-AS 聚类算法,它的优势至少表现在以下几个方面:1) 可同时基于类别变量和连续变量进行聚类;2) 可自动确定最终的分类型数;3) 可处理大型数据集。

TwoStep-AS 聚类基本步骤如下:

1) 预聚类: 采用“贯序”方式将样本粗略划分成 L 个子类。开始阶段视所有数据为一个大类。读入一个样本数据后,根据“亲疏程度”决定该样本应派生出一个新类,还是应合并到已有的某个子类中。这个过程反复进行,最终形成 L 个类。预聚类过程聚类数目不断增加。

2) 聚类: 在预聚类的基础上,再根据“亲疏程度”决定哪些子类可以合并,最终形成 L 类。聚类数目不断减少的过程,随着聚类的进行,类内部的差异性将不断增大。

在距离测量有“欧氏距离”和“对数似然”两种算法作为聚类变量相似度的测量形式;如果聚类变量均为数值型,采用欧氏距离,若为混合型采用对数似然距离。在聚类准则有贝叶斯(BIC)和 Akaike 信息准则(AIC),作为聚类个数的判断依据。同其他统计方法一样,TwoStep-AS 聚类算法也有严苛的适用条件,它要求模型中的变量独立,类别变量是多项式分布,连续变量符合正态分布。

注意,与其他的聚类分析算法一样,所得到的模型可以在一定程度上依赖于训练数据的顺序。重新排序数据和重建模型可能会导致不同的最终群集模型。对记录顺序的解决方案的鲁棒性可以通过多次拟合一个模型来评估,每次使用不同的随机顺序记录。

评价模型质量的方法采用轮廓系数(Silhouette Coefficient)^[10-14],是聚类效果好坏的一种评价方式。最早由 Leonard Kaufman, Peter J. Rousseeuw 在 1986 年提出。它结合内聚度和分离度两种因素。可以用在相同原始数据的基础上用来评价不同算法,或者算法不同运行方式对聚类结果所产生的影响。

对数据中每一个数据样本与本结果簇中其他样本间的相似性以及该数据样本与其他结果簇中样本的相似性进行定量分析,并将定量结果进行某种形式的组合,得到聚类结果优劣的评价。

对于第 i 个样本,计算该样本到其所属簇中所有样本的平均距离,记为 $a(i)$,对于第 i 个样本和不包

含该样本的任意簇,计算该样本到给定簇的中所有样本的平均距离,所有簇的平均距离最小者记为 $b(i)$ 。对于第 i 个样本,其轮廓系数的计算如下公式(1)所示。

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (1)$$

式中: $a(i) = \text{average}(i \text{ 向量到所有它属于的簇中其他点的距离})$; $b(i) = \min(i \text{ 向量到所有非本身所在簇的点的平均距离})$ 。

$a(i)$ 越小,说明样本 i 越应该被聚类到该簇。 $b(i)$ 越大,说明样本 i 越不属于其他簇。

轮廓系数 $s(i) = +1$ 时,表示样本 i 与其他簇中的对象相异性较大,聚类合理。

轮廓系数 $s(i) = 0$ 时,表示样本 i 分类不明显,在两个簇的边界上。

轮廓系数 $s(i) = -1$ 时,表示样本 i 被分配到一个错误的簇中,轮廓系数的值在 -1 和 $+1$ 之间变化。

所有样本的 $s(i)$ 的均值称为聚类结果的轮廓系数,是该聚类是否合理、有效的一种度量方式。

1.2.3.2 评价指标的重要度算法 考察地力指标的重要性,可使用 Pearson 样本相关系数、似然比 (Likelihood Ratio) 卡方值和 Cramer's V 系数。SPSS Modeler 计算 Cramer's V 系数^[9],见公式 2。

$$V = \sqrt{\frac{\chi^2}{n \min[(R-1), (C-1)]}} \quad (2)$$

式中: χ^2 是 Pearson 卡方统计量, R 和 C 分别表示列联表的行数和列数。Cramer's V 系数取值范围在 $0 \sim 1$, 越接近 1, 越说明输入变量与输出变量的相关性越强,输入变量越重要。重要度在 $0 \sim 0.2$ 范围为差,重要度在 $0.2 \sim 0.6$ 范围为中,重要度在 $0.6 \sim 1$ 范围为良。

2 结果与分析

2.1 地力指标的筛选结果 初次运算,输入地力特征参数包括成土母质、地貌类型、坡向、pH 值、全氮、速效钾、有机质、年平均降水量、北纬、东经、有效土层厚度、海拔等在内的 14 个指标,特征重要度排名前 10 的指标如图 1 所示。年平均降雨量特征重要度为 1,全氮特征重要度为 0.67,有效土层厚度特征重要度为 0.58,纬度特征重要度为 0.58,成土母质特征重要度为 0.51,有机质特征重要度为 0.46,经度特征重要度为 0.36,地貌类型特征重要度为 0.21,坡向特征重要度为 0.18,有效磷特征重要度为 0.11。

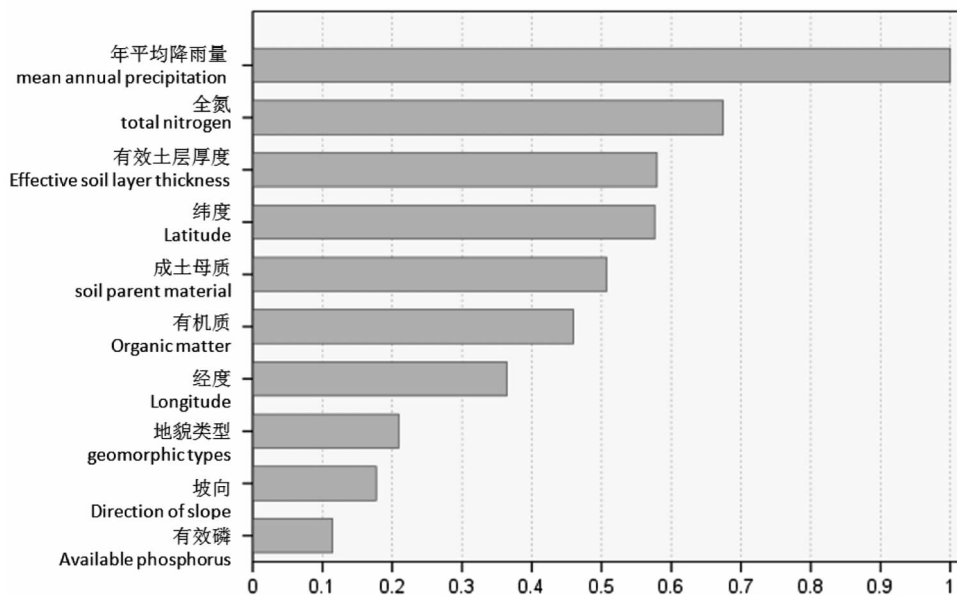


图1 初次运算的地力指标的特征重要度

Fig. 1 Characteristic importance of land productivity at the initial operation

根据特征重要度的大小,选择特征重要度 >0.2 以上的地力指标作为模型聚类的特征值,经过反复迭

代筛选,以轮廓系数最优为依据,最终确定地力指标为年平均降雨量、全氮、有效土层厚度、纬度和成土母质(图2),年平均降雨量重要度为1,纬度的重要度为0.9,全氮的重要度为0.75,有效土层厚度的重要度为0.59,成土母质的重要度为0.57。从图2可以看出,影响海南胶园地力的5个指标中,土壤方面有3个,分别是全氮、成土母质、有效土层厚度,这3个因子一起综合在一定程度上代表了不同土壤的养分总量,因有机质含量与全氮有较大的相关性,选择特征重要度 >0.2 重新迭代时就隐含了;地理方面有1个指标,就是纬度,这是一个影响气温和热量的十分重要的因子;气象方面有一个指标,就是年均降雨量,影响的是土壤水分。可见,水、热、养分,作为热带树种橡胶树来说,是其生长发育和高产高效的前提条件,也就是构成地力的重要因子。

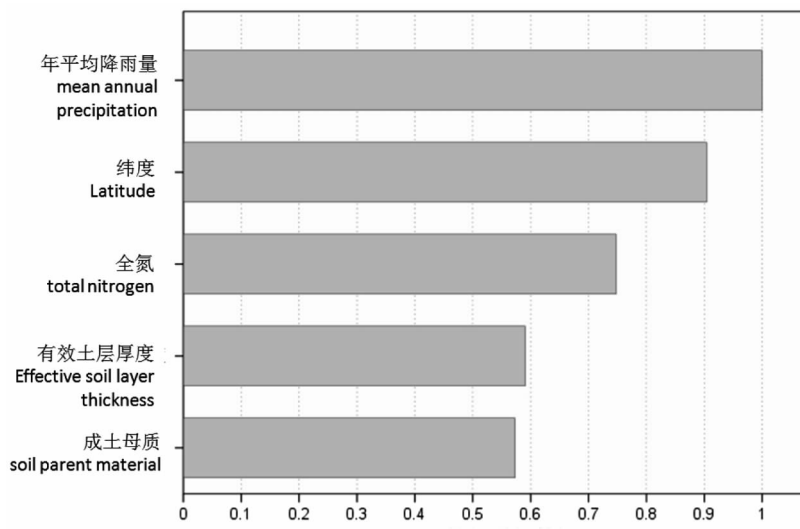


图2 最终确定的地力指标的特征重要度

Fig. 2 Characteristic importance of land productivity at the final operation

2.2 TwoStep-AS 模型聚类结果 根据最终确定的地力指标(常年降雨量、纬度、全氮、有效土层厚度和成土母质)建立 TwoStep-AS 聚类模型。模型参数信息如下:最小常规聚类数2类,最大常规聚类数15类。自适应特征选择,信息标准采用贝叶斯信息标准(BIC),距离测量采用对数似然。最终模型常规聚类数为3个,即将海南垦区十大片区的耕地划分为3种类型。3种耕地类型的概要特征如表4所示,每种耕地类型最重要的特征在表中以粗体显示。第1种耕地类型最重要的特征是年平均降雨量在1 600~1 700 mm,全氮质量含量在 $0.5 \sim 0.65 \text{ g} \cdot \text{kg}^{-1}$,有效土层厚度在70~100 cm,成土母质为玄武岩。第2种耕地类型最重要的特征是年平均降雨量大于2 000 mm,全氮质量含量在 $0.05 \sim 0.10 \text{ g} \cdot \text{kg}^{-1}$,有效土层厚度在55~70 cm,成土母质为花岗岩。第3种耕地类型最重要的特征是年平均降雨量在1 700~1 800 mm,全氮质量含量在 $0.65 \sim 0.85 \text{ g} \cdot \text{kg}^{-1}$,有效土层厚度在70~100 cm,成土母质为花岗岩。

表4 3种耕地类型的聚类中心

Tab. 4 The clustering center for three types of farmland

地力指标 Land productivity indexes	Cluster - 1	Cluster - 2	Cluster - 3
年平均降雨量/mm Mean annual precipitation/mm	1 600 ~ 1 700	> 2 000	1 700 ~ 1 800
纬度/(°) Latitude/(°)	19.55 ~ 20.00	18.19 ~ 18.64	19.09 ~ 19.55
全氮/($\text{g} \cdot \text{kg}^{-1}$) Total nitrogen/($\text{g} \cdot \text{kg}^{-1}$)	0.50 ~ 0.65	0.05 ~ 0.10	0.65 ~ 0.85
有效土层厚度/cm Effective soil layer thickness/cm	70 ~ 100	55 ~ 70	70 ~ 100
成土母质 Soil-forming parent material	玄武岩 Basalt	花岗岩 Granite	花岗岩 Granite

注:最重要的特征用粗体表示

Note: The most important characteristics are in bold.

2.3 模型评估 在 SPSS Modeler 中,提出了 2 种度量聚类效果的指标,一个是优度,另一个是重要度。优度是一种测量聚类内聚性和分离性的指标。优度在 $-1 \sim 0.2$ 范围为差,优度在 $0.2 \sim 0.5$ 范围为中,优度在 $0.5 \sim 1$ 范围为良。重要度是一种测量聚类内聚性的指标。重要度在 $0 \sim 0.2$ 为差,重要度在 $0.2 \sim 0.6$ 范围为中,重要度在 $0.6 \sim 1$ 范围为良。轮廓系数是聚类效果好坏的一种评价方式,轮廓系数在 $-1 \sim 0.2$ 范围为差,在 $0.2 \sim 0.5$ 范围为中,在 $0.5 \sim 1$ 范围为良。用 TwoStep-AS 聚类算法为基于海南农垦测土调查数据进行分类时,总体模型优度(平均轮廓系数)为 0.38。各聚类的优度和重要度如表 5 所示。从表 5 的数据来看,各聚类类型的优度为中等,而重要度为良好,说明用 TwoStep-AS 聚类算法对海南农垦测土调查数据进行聚类时内聚性一般,而分离性很好。

表 5 模型质量
Tab.5 Model quality

聚类类型 Clustering types	样地数/个 Number of sample plots	优度 Goodness of fit	重要度 Importance
Cluster - 1	1 715	0.167	0.78
Cluster - 2	2 023	0.498	1.00
Cluster - 3	1 205	0.528	0.567

3 结 论

1) 本研究使用基于大数据的 TwoStep-AS 聚类模型对海南农垦测土调查数据进行聚类分析,将海南垦区十大片区划分为 3 种耕地类型。第 1 种耕地类型最重要的特征是年平均降雨量在 $1\,600 \sim 1\,700\text{ mm}$,全氮质量含量在 $0.5 \sim 0.65\text{ g} \cdot \text{kg}^{-1}$ 范围,有效土层厚度在 $70 \sim 100\text{ cm}$ 范围,成土母质为玄武岩。第 2 种耕地类型最重要的特征是年平均降雨量大于 $2\,000\text{ mm}$,全氮质量含量 $\text{g} \cdot \text{kg}^{-1}$ 在 $0.05 \sim 0.10\text{ g} \cdot \text{kg}^{-1}$ 范围,有效土层厚度在 $55 \sim 70\text{ cm}$,成土母质为花岗岩。第 3 种耕地类型最重要的特征是年平均降雨量在 $1\,700 \sim 1\,800\text{ mm}$,全氮质量含量在 $0.65 \sim 0.85\text{ g} \cdot \text{kg}^{-1}$ 范围,有效土层厚度在 $70 \sim 100\text{ cm}$,成土母质为花岗岩。2) 影响海南胶园耕地的最重要的地力指标分别是年平均降雨量、全氮、北纬、有效土层厚度、成土母质。3) 基于数据挖掘的地类划分没有人为主观因素的干扰,与作物产量无关,以耕地本身特征内在的凝聚和分离为依据,排除了人为主观因素,这种分类将更有助于人们对耕地性质的认识,有利于土地的区划和综合利用。4) 本研究结果为建立统一的海南垦区胶园地力评价指标体系提供了初步的方法和结果,并为下一步开展系统性的研究打下了基础,其结果的准确性和适用性还有待今后在深入研究和应用中得到检验。

参考文献:

[1]陈桂芬,马丽,董玮,等. 聚类、粗糙集与决策树的组合算法在地力评价中的应用[J]. 中国农业科学, 2011, 44(23) : 4833 - 4840.

[2]蔡丽霞,陈桂芬. 基于数据挖掘算法土壤肥力状况的分析研究[J]. 中国农机化学报, 2014, 35(3) : 252 - 255.

[3]闫一凡,刘建立,张佳宝,等. 耕地地力评价方法及模型分析[J]. 农业工程学报, 2014, 30(5) : 204 - 210.

[4]张海涛,周勇,汪善勤,等. 利用 GIS 和 RS 资料及层次分析法综合评价江汉平原后湖地区耕地自然地力[J]. 农业工程学报, 2003, 19: 219 - 223.

[5]BAJA, S, CHAPMAN D. M, DRAGOVICH D. Spatial based compromise programming for multiple criteria decision making in land use planning[J]. Environmental Modeling and Assessment, 2007, 12: 171 - 184.

[6]LIU WENXIN, HE XIULI. Suitability analysis of scale management in Northeast China based on efficiency evaluation[J]. Asian Agricultural Research, 2018, 10(10) : 30 - 34.

[7]黄艳. 平潭综合实验区耕地调查与地力评价[J]. 土壤科学, 2017, 5(2) : 13 - 21.

[8]龚子同,张甘霖,漆智平. 海南岛土系概论[M]. 北京: 科学出版社, 2004: 78 - 80.

[9]薛薇,陈欢歌. SPSS Modeler 数据挖掘方法及应用(第 2 版) [M]. 北京: 电子工业出版社, 2014: 232 - 257.

- [10] KAUFMAN L, ROUSSEEUW P J. Finding groups in data: an introduction to cluster analysis [M]. Lanada: John Wiley&Sons, 2009.
- [11] ARANGANAYAGI S, THANGAVEL K. Clustering categorical data using silhouette coefficient as a relocating measure [J]. Computational Intelligence and Multimedia Applications(ICCIMA) , 2008: 13 – 17.
- [12] 塞巴斯蒂安·拉施卡. Python 机器学习 [M]. 北京: 机械工业出版社, 2017: 200 – 203.
- [13] 朱连江, 马炳先, 赵学泉. 基于轮廓系数的聚类有效性分析 [J]. 计算机应用, 2010, 30(2) : 139 – 141.
- [14] 王学贺. 一种基于改进微粒群和轮廓系数的划分聚类方法 [J]. 云南民族大学学报(自然科学版) , 2016, 25 (4) : 367 – 371.

Evaluation of Land Productivity in Rubber Plantations Based on Data Mining

PENG Jinlian¹, TANG Zhaosheng¹, JIANG Jusheng², ZHAO Chunmei³

(1. School of Computer Science & Cyberspace Security, Hainan University, Haikou, Hainan 570228, China; 2. Hainan State Farms Academy of Sciences, Haikou, Hainan 570206, China; 3. Rubber Research Institute, CATAS, Danzhou, Hainan 571737, China)

Abstract: The data of soil survey in rubber plantations in Hainan State Farms was analyzed to evaluate the land productivity in the rubber plantations by using four clustering algorithms (K-Means, Two-Step, Kohonen and TwoStep-AS) . According to the Silhouette coefficients, TwoStep-AS clustering models were established for cluster analysis of the indexes of the land productivity in the Hainan State Farm rubber plantations. After repeated iterations, the land in rubber plantations of Hainan State Farms was divided into three types, and the indexes affecting the land type of the rubber plantations were selected, including annual average rainfall, effective soil thickness, total nitrogen, north latitude, and soil-forming parent material.

Keywords: rubber plantation; data mining; evaluation of soil fertility; Silhouette coefficient

(责任编辑: 叶 静)