北方园艺 2011 (6):218~220

决策树在精准农业中的应用现状与发展趋势

史　众[1,2]，陈立平[2]，陈天恩[2]

(1.首都师范大学 信息工程学院，北京 100048；2.国家农业信息化工程技术研究中心，北京 100097)

摘　要：决策树算法是一种重要的分类方法，是数据挖掘领域研究热点之一。现介绍了决策树算法当前的研究状况。从土壤质量等级划分、自然灾害分析、遥感影像分类及耕地质量分析等方面介绍决策树算法在精准农业中的应用现状，对决策树算法在精准农业中的应用前景进行了展望。

关键词：决策树；分类；精准农业

中图分类号：S-1　文献标识码：A　文章编号：1001－0009(2011)16－0218－03

我国是农业大国，幅员辽阔，土壤类型众多，作物品种复杂，病虫害发生频繁且症象不断变化。我国曾进行过各种农业普查，科技人员积累了大量与农业生产过程密切相关的属性数据和空间数据，这些数据具有多维性、海量性、空间性、时间性等特点，它们真实、具体的反映了农业生产作业的本质状况，是指导区域精准作业的宝贵财富[1]。目前的数据库系统可以高效实现数据的录入、查询、统计等功能，但无法发现数据之间的关系，无法预测未来发展趋势，因而无法指导农业生产。数据挖掘技术可以解决农业领域“数据丰富知识贫乏”的状况。

数据挖掘(Data mining，DM)，是从大量的、不完全的、模糊的、随机的数据中，提取隐含其中的、人们不知道的、具有潜在利用价值的信息和知识的过程[2]，挖掘过程如图1所示。数据挖掘提出至今，已有顶级学术会议VLDB，ICDE，SIGMOD召开研讨数据挖掘技术，典型的有影响的数据挖掘工具也很多，如SAS公司的Enterprise Miner，IBM公司的Intelligent Miner，SPSS公司的Clementine等。数据挖掘方法很多，决策树，关联规则，模糊聚类，神经网络，支持向量机等，其中决策树方法是应用广泛的分类方法之一。2006年，国际权威学术组织IEEE International Conference on Data Mining(ICDM)从18个候选算法中评出数据挖掘十大经典算法，它们分别是C4.5，k-Means，SVM，Apriori，EM，PageRank，AdaBoost，kNN，Naïve Bayes和CART，其中C4.5和CART是常用的决策树算法，可见决策树算法的重要性。

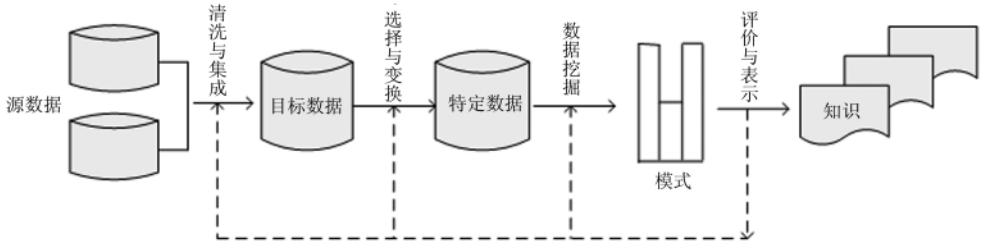

图1　数据挖掘过程

精准农业技术体系从实施过程来分大致包括农田信息获取、信息管理与分析、决策分析、决策的实施四大部分，决策树算法在这4部分中均有应用[3]。RS(遥感)是属于农田信息的获取手段之一，决策树算法可以对遥感影像图进行分析。GIS是农田信息管理和分析手段，目前决策树可以针对GIS处理的数据进行分类。以决策树为中心构建的决策支持系统也很多，由此可见决策树在精准农业中应用十分广泛的[4]。

1　决策树算法

决策树(Decision tree)算法是用于分类和预测的主要技术。决策树由节点和分枝组成，其中节点分为根节点、内部节点和叶子节点，根节点和内部节点对应于待分类对象的属性，叶子节点代表一种可能的分类结果。决策树的分枝代表一个测试输出。如图2所示，A为根节点，B、C…F为叶子节点，a、b…f为内部节点。它采用自顶向下的递归方式，在决策树的内部节点进行属性值的比较并根据不同的属性值判断从该节点向下的分支[5]，所以从根到叶节点就对应着一条合

第一作者简介：史众(1986-)，女，硕士，研究方向为数据挖掘。E-mail：99tiantian@163.com。

收稿日期：2011－05－04

取规则,整棵树就对应着一组析取表达式规则。图 2 中从 A 到 B 位一个条件测试。它是一种逼近离散函数值的方法,分类精度高,操作简单,因而成为实用的并且比较流行的数据挖掘算法[6]。构造最优的决策树问题已经被证明是 NP-完全问题,因此典型的决策树学习算法采取用启发式策略进行构造。

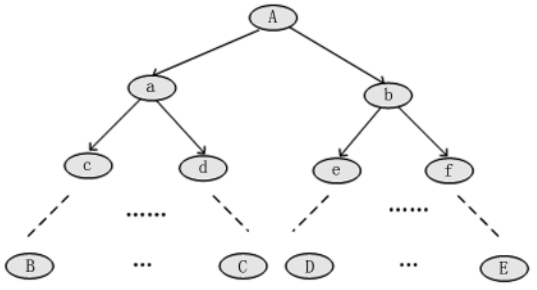

图 2 决策树模型

在决策树的构建过程通常考虑 3 个方面。特征选择:根据某种策略选择属性进行分裂。常用的是信息增益或纯性度量方法,包括信息增益率,基于距离的度量(Distance measure),χ^2 统计,最小二乘法、正交法(Ortogonality measure)等。不同的属性分裂方法有不同的效果,尤其是对于多值属性。节点分裂:根据选定属性的不同取值将节点分开。当该节点是连续属性时,将其离散化处理。离散化方法很多,Dougherty 提出的全局离散化方法,C 4.5 算法提出基于某个结点的实例进行局部离散化方法等。剪枝:是最常用的简化决策树的方法,当训练集中含有噪声时,可能使生成的决策树出现过拟合(Over fitting)现象,通常采用预剪枝(Pre-pruning)和后剪枝(Post-pruning)2 种方法克服噪声。

2 决策树在精准农业的应用

精准农业是 20 世纪 80 年代由英美等发达国家率先发展起来的跨学科综合技术,其特点是通过全球定位系统(GPS)、遥感(RS)、地理信息系统(GIS)、自动化技术和网络技术并结合农学、地学、生态学规律和模型,根据土壤特性和作物生长发育的需要,精细准确的调整各种农艺措施,以降低物资消耗、增加利润、保护生态环境,实现农业可持续发展。

2.1 决策树在土壤分类方面的应用

土壤质量的评价与检测是评价和重新设计可持续性土地利用系统的基础。传统的地学统计方法需要大量野外采样,而基于景观建模方法的分类以统计学为基础,难以处理字符型数据,并且完全依赖数据本身的学习模型

任舟桥等在海南琼海市 3 个试验区内,选取有机质含量、地形坡度、速效磷、速效钾等属性作为土地适宜性评价测试属性,用 C4.5 算法将土地适宜性划分成 4 个等级,因而随着土地资源数据的变更能快速更新土地适宜性评价数据[7]。孙微微等用 C4.5 方法,选取高程、地面坡度、土壤 pH、地貌类型等属性,以文献[8]中的土壤质量评价方法作为分类属性,对广东省部分地区土壤质量等级进行有效划分[9]。周斌等利用浙江省龙游县试验区土壤性质数据,用 C5.0 算法将土壤性质含量与景观属性(包括地形、地质等)联系起来,建立了 PH、OM、速效 P、速效 K 4 种决策树模型,并采用前剪枝与后剪枝结合的方法对决策树进行剪枝。该方法适合于具有高度空间变异性地区的土壤调查和制图,有助于决策管理[10]。

2.2 在自然灾害方面的应用

农业对自然灾害是比较敏感的,尤其是气象灾害和病虫害,它们直接影响着农业粮食安全,因此掌握灾害发生的特点和规律,来提高防灾减灾能力,在精准农业的研究中有十分重要的意义。司巧梅用 C4.5 算法对牡丹江部分县区 2008 年 6～8 月风雹灾害数据构造决策树,按照地区、时间、受灾人口等属性将数据划分无经济损失、较大经济损失、很大经济损失 3 个等级。并试验证明该方法对灾害分类正确性达 90%以上[11]。金海月等运用 ID3 的改进算法 IBLE,以黄瓜、玉米叶部常见病害作为研究对象,试验结果显示该算法正确进行病例识别,正确率达 92%以上[12]。

2.3 在遥感影像分类方面的应用

随着雷达、红外、光电、卫星等宏观和微观传感器的使用,遥感影像数据的数量、大小和复杂性都在飞快的增长,已经远远超出人的分析和解译能力。用户不可能详细的分析所有这些数据,并提取感兴趣的空间知识。遥感影像分类一直是遥感技术领域研究的一项重要内容。将光谱数据与其它辅助数据结合,发展综合信息复合的方法可以大大提高分类的精度,是提高遥感应用性的有效途径之一,因而从大量数据中自动获取知识是研究热点。目前决策树方法在遥感影像分类的研究很多。赵萍等人利用分类回归树方法对江苏省江宁试验区的土地利用/覆被情况进行分类,选择了地理坐标、纹理特性(均值、方差等)、地形因子等 16 个测试变量,用 3 340 个样本将试验区分为水体、居民地、道路、荒草地、水田、林地和阴影 7 类,试验证明,基于 CART 的方法分类精度比传统的监督分类和逻辑通道法有较大提高,并且计算简单,运行速度快[13]。张丽娜等利用 ETM 数据,选取 SAVI、NDMI、纹理等属性对张北地区土壤盐碱化进行研究,将试验区分为重度盐碱、轻度盐碱、水体、植被和裸地 5 个类别,有效的提取试验区的盐碱地信息,并根据实际地况,对盐碱地成因做了分析[14]。麦吉尔大学麦克唐纳分校的农学研究中心利用 CART 方法对高光谱遥感数据分析,对试验区玉米是否使用除草剂,氮肥施用量进行高效鉴别[15]。针对目前 LUCC(土地利用/覆盖)信息提取中自动化程度低,精度与可靠性不足等缺陷,王萍首先建立综合数据库,然后在区域范围内利用决策树技术进行变化信息分层提取技术,形成分层次决策树的土地利用/土地覆盖变化信息提取技术方法和流程[16]。

2.4　耕地质量评价中的应用

耕地质量评价是根据耕地构成因素组合特征的差异或直接根据单位面积耕地产量的高低或耕地价值的多少,对耕地质量进行等级的划分。耕地地力划分是耕地质量评价的重要方面。曹丽英利用 ID3 算法对吉林省德惠市的土壤数据进行分析,根据有机质、全氮、速效磷、速效钾将土壤分成 6 个等级[17]。田剑等利用改进的 C5.0 算法对广东省龙川县建立耕地评价模型,预测精度达 94.92%,试验结果表明,运用决策树进行耕地评价是可行的,其建立的评价模型具有精度高、鲁棒性和易理解性等特点[18]。

2.5　其它应用

精准农业中的产量图分析是获取影响农田产量差异主要因子的重要手段,是精准农业决策信息获取的切入点和突破口。薛正平等人运用决策树和图层叠置的方法,以宁夏农垦局暖泉农场采样数据作为分析对象,研究得出该地的土壤养分状况,进而得出提高单产的主要措施[19]。

3　小结

对数据挖掘中决策树算法的研究现状,及已成熟的决策树分类算法及构建方式进行了简单论述。决策树归纳方法在农业上已经有了广泛的应用,并且有了许多成熟的系统。但是目前决策树算法在精准农业上只是取得初步成果,还有大量的理论和方法需要深入研究。

在理论上,国内外文献中对决策树算法研究主要有以下几个方面:一是对大数据集的适应性。ID3、C4.5、C5.0 等算法都限制训练样本驻留内存,这一限制制约了算法的可伸缩性,是决策树应用中必须面对和解决的关键问题。这方面的尝试很多,比较有代表性的研究是 SLIQ、SPRINT,雨林等算法,他们强调了决策树对大训练集的适应性。二是与其它数据挖掘算法结合,例如粗糙集、贝叶斯、关联规则等。三是对原有算法的改进,例如属性选择,叶子数目,属性和类标签的多值问题,树的整体性能优化等。从有偏的数据中学习:农业数据极易受到其他因素影响,因此要从这些并不能反映实际情况的数据中学习,构建准确率高的决策树。针对农业数据的自身特点,海量、动态变化、不确定等,处理海量数据,学习变化,从而动态指导农业生产.

参考文献

[1] 陈桂芬.面向精准农业的空间数据挖掘技术研究与应用[D].长春:吉林大学,2009.

[2] 史忠植.知识发现[M].北京:清华大学出版社,2002:22-45.

[3] Zhao C J. Progress of agricultural information technology[M]. International Academic Publishers,2000.

[4] 赵春江,薛绪掌,王秀,等.精准农业技术体系的研究进展与展望[J].农业工程学报,2003,19(4):7.

[5] 巩吉璋.决策树分类算法在银行个人信用评级中的应用[D].广州:暨南大学,2008.

[6] 郑明超.数据挖掘技术中分类算法的比较分析[D].兰州:兰州商学院,2007.

[7] 任舟桥,刘耀林,焦利民.基于决策树的土地适应性评价[J].国土资源科技管理,2007(3):5.

[8] 胡月明,万洪富,吴志峰,等.基于 GIS 的土壤质量模糊变权评价[J].土壤学报,2001,38(3):266.

[9] 孙微微,胡月明,刘才兴,等.基于决策树的土壤质量等级研究[J].华南农业大学学报,2005,26(3):108.

[10] 周斌,王繁.基于决策树模型的土壤性质空间判断[J].土壤通报,2004,35(4):385.

[11] 司巧梅.基于决策树的农业气象灾害等级预测模型[J].安徽农业科学,2010,38(9):4925.

[12] 金海月,宋凯.决策树算法在农业病害诊断中的应用[J].当代农机,2007(5):76-77.

[13] 赵萍,傅云飞,郑刘根,等.基于分类回归树分析的遥感影像土地利用/覆被分类研究[J].遥感学报,2005,9(6):708.

[14] 张丽娜,程晓,伍吉仓,等.基于决策树分类的张北土壤盐碱化研究[J].安徽农业科学,2009,37(25):12103.

[15] Waheed T,Bonnell R B,Prasher S O,et al. Measuring performance in precision agriculture:CART-A decision tree approach[J]. Agricultural Water Management,2006:173.

[16] 王萍.遥感土地利用/土地覆盖变化信息提取的决策树方法[D].济南:山东科技大学,2004.

[17] 鲁丽英.决策树在耕地地力等级评价中的应用研究[D].长春:长春理工大学,2009.

[18] 田剑,胡月明,王长委,等.聚类支持下决策树模型在耕地评价中的应用[J].农业工程学报,2007,23(12):58.

[19] 薛正平,邓华,杨星卫,等.基于决策树和图层叠置的精准农业产量图分析方法[J].农业工程学报,2006,22(8):140.

Application Situation and Advance of Decision Tree in Precision Agriculture

SHI Zhong[1,2],CHEN Li-ping[2],CHEN Tian-en[2]

(1. College of Information Technology, Capital Normal University, Beijing 100048; 2. National Agricultural Information Engineering Research Center, Beijing 100097)

Abstract: Decision tree algorithm is an important classification method. It's a hotspot in data mining areas. This paper firstly introduced the current research status of decision tree algorithm, then discussed the application status from soil quality hierarchy, natural disasters analysis, remote sensing image classification and cultivated land quality analysis, the application in precision agriculture was prospected at last.

Key words: decision tree; classification; precision agriculture