

基于决策树分类技术的遥感影像分类方法研究

姜丽华^{1,2} 杨晓蓉^{1,2}

(1.中国农业科学院 农业信息研究所 网络技术研究室,北京 100081;

2.农业部重点开放实验室 智能化农业预警技术重点开放实验室,北京 100081)

摘 要 采用决策树分类技术对遥感影像进行分类,阐述了决策树算法结构和原理,讨论了 C4.5 基本原理以及新技术 Boosting 方法,探讨了决策树在遥感数据分类方面的优势,从而提高了遥感影像的分类精度。

关键词 决策树;分类;遥感影像

中图分类号 TP7

文献标识码 A

文章编码 :1672-6251(2009)10-0034-03

Classification Methods of Remote Sensing Image Based on Decision Tree Technologies

Jiang Lihua^{1,2}, Yang Xiaorong^{1,2}

(1.Agriculture Information Institute, Chinese Academy of Agriculture Sciences, Beijing 100081, China; 2.Key Laboratory of Digital Agricultural Early-warning Technology, Ministry of Agriculture, Beijing 100081, China)

Abstract: Decision tree classification algorithms have significant potential for remote sensing data classification. It was advanced to adopt decision tree technologies to classify remote sensing images in this paper. First, the algorithms structure and the algorithms theory of decision tree were discussed. Second, C4.5 basic theory and boosting technology were explained. The decision tree technologies have several advantages for remote sensing application by virtue of their relatively simple, explicit and intuitive classification structure.

Key words: decision tree; classification; remote sensing image

1 引言

遥感信息的提取与分类一直是遥感技术领域研究的一项重要内容。遥感分类应用中,监督与非监督分类的传统分类方法^[1]以及人工神经网络分类、专家系统分类^[2-3]等新方法都以影像光谱特征为基础。然而,由于影像本身存在“同物异谱、异物同谱”现象,这种纯粹依赖地物光谱特征的分类方法往往会出现较多的错分、漏分情况。众多研究表明,结合影像光谱信息和其他辅助信息,可以大大提高分类精度^[4]。

决策树分类作为一种基于空间数据挖掘和知识发现(Spatial Data Mining and Knowledge Discovery, SDM&KD)的监督分类方法,突破了以往分类树或分类规则的构建要利用分类者的生态学和遥感知识先验确定,其结果往往与其经验和专业知识水平密切相关的问题。它通过决策学习过程得到分类规则并进行分类,分类样本属于严格“非参”,不需要满足正态分

布,可以充分利用 GIS 数据库中的地学知识辅助分类,大大提高了分类精度^[5-6]。目前,决策树分类方法已经开始应用于各种遥感影像信息提取和土地利用、土地覆盖分类中^[7-11]。在美国 USGS、EPA 等部门联合实施的“美国土地覆盖数据库”计划(NLCD 2001)中,决策树分类技术不仅被应用于土地分类,而且应用于城市密度信息提取和林冠密度信息提取,土地利用分类精度达到了 73~77%,城市密度信息提取精度达到 83~91%,树冠精度在 78~93%,其制图效率比原有的方法提高了 50%,完全满足大规模土地分类数据产品生产要求。

决策树学习方法是解决实际问题中分类问题的数据挖掘方法之一,能够从无次序、无规则的事例中推理出决策树表达形式的分类规则。决策树学习的一个最大优点就是学习过程中不需要操作人了解很多背景知识,只要训练样本能够用“属性—结论”的方式表

收稿日期 2009-07-30

基金项目:国家科技支撑计划项目(2006BAD10A06);中国农业科学院农业信息研究所公益型科研院所基本科研专项资金;科研院所技术开发研究专项“基于智能检索的西藏科技资源共享技术”

作者简介:姜丽华(1980-),女,硕士,助理研究员,研究方向:信息网络。

达出来, 就能使用该算法来学习。决策树学习获得的分类知识易于表达和应用, 目前国外已经有学者利用决策树学习方法获取知识并应用于空间分析与研究过程。

2 决策树算法

决策树 (Decision Tree) 是通过对训练样本进行归纳学习, 生成决策树或决策规则, 然后使用决策树或决策规则对新数据进行分类的一种数学方法。决策树是一个树型结构, 它由一个根节点 (Root node)、一系列内部节点 (Internal) 及叶节点 (Leaf nodes) 组成, 每一个节点有一个父节点和两个或者多个子节点, 节点间通过分支相连。决策树的每个内部节点对应一个非类别属性或属性的集合 (也可以称为测试属性), 每条边对应该属性的每个可能值。决策树的叶结点对应一个类别属性值, 不同的叶结点可以对应相同的类别属性值。决策树除了以树的形式表示外, 还可以表示为一组 IF-THEN 形式的产生式规则。决策树中每条由根到叶的路径对应着一条规则, 规则的条件是这条路径上所有节点属性值的取舍, 规则的结论是这条路径上叶节点的类别属性。与决策属性比, 规则更简洁, 也便于人们理解、使用和修改, 可以构成专家系统的基础, 因此在实际应用中更多地使用规则。本文主要介绍在遥感应用中广泛使用的一种算法—分类回归树 (CART), 以及决策树的另一种算法—C4.5 算法。

2.1 分类回归树 (CART)

分类回归树 CART (Classification and Regression Tree) 为一种通用的树生长算法, 由 Breiman 等人提出, 是一种监督分类方法, 它利用训练样本来构造二叉树并进行决策分类。其特点是充分利用二叉树的结构 (Binary Tree-structured), 即根节点包含所有样本, 在一定的分割规则下根节点被分割成两个子节点, 这个过程又在子节点上重复进行, 成为一个回归过程, 直至不可再分成子节点为止。构造 CART 树采用的思路: 在整体样本数据的基础上, 生成一个多层次、多叶节点的大树, 以充分反映数据之间的联系 (这时树生长为考虑噪声, 往往反映的是训练过度情况下的数据联系), 然后对其进行删减, 产生一系列子树, 从中选择适当大小的树, 用于对数据进行分类, 具体可分为树生长和树剪枝两部分。

2.1.1 树生长 树节点处的一次判别称为一个分支, 它对应于将训练样本划分成的子集, 根节点处的分支对应于全部训练样本, 其后每一次判别都是一次训练

子集划分过程, 因此构造树的过程实际上就是一个属性查询产生分割规则的过程。在本文中, CART 采用了一种称为“节点不纯度”的指标: 用 $i(N)$ 表示节点 N 的“不纯度”, 当节点上的模式数据均来自同一类别时, $i(N) = 0$; 而若数据所属类别均匀分布时, $i(N)$ 应当很大, 分割规则即是基于不纯度函数的极小值而产生的。当前流行的两种“不纯度”测量函数为:

(1) “熵不纯度” (Entropy Impurity), 亦称为信息量不纯度 (Information Impurity):

$$i(N) = -\sum_j P(w_j) \log_2 P(w_j) \quad (1)$$

其中, $p(w_j)$ 是节点 N 处属于 w_j 类模式样本数占总样本数的概率。根据熵的特性, 如果所有模式的样本都来自同一类别, 则不纯度为零, 否则是大于零的正值, 当所有类别以等概率出现时, 熵取最大值。

(2) 方差不纯度—“Gini 不纯度”, 根据多分类问题中节点样本来自不同类别与总体分布方差有关而提出:

$$i(N) = -\sum_{i \neq j} P(w_i) P(w_j) = 1 - \sum_j P^2(w_j) \quad (2)$$

“Gini 不纯度”的意义即为当节点 N 的类别标识任意选取时对应的误差率。

给定一部分树, 目前已经生长到节点 N , 要求对该节点作属性查询时, 一个明显的启发式的思路是选择那个能够使不纯度下降最快的那个查询, 不纯度下降可表示为:

$$\Delta i(N) = i(N) - P_L i(N_L) - (1 - P_L) i(N_R) \quad (3)$$

其中 N_L 和 N_R 分别是左右节点, $i(N_L)$ 、 $i(N_R)$ 是相应的不纯度。 P_L 是当查询 T 被采纳时, 树由 N 生长到 N_L 的概率, 这样最佳的查询值 S 就是那个能最大化 $\Delta i(T)$ 的值。

2.1.2 树剪枝 如果持续生长树, 直到所有的叶节点都达到最小的不纯度为止, 数据一般将被“过拟合”, 那么分类树就退化成方便的查找表, 这样, 对有较大贝叶斯误差的噪声信号灯推广性能就有可能不好, 相反, 如果分支停止的太早, 那么对于训练样本的误差就不够小, 导致分类性能很差。一种主要的停止分支方法就是剪枝 (Pruning), 同时也是为了避免树生长得过分庞大。本文通过最小化如下的定义的全局指标来达到目的:

$$\text{cost} = \alpha \cdot \text{size} + \sum_{\text{叶节点}} i(n) \quad (4)$$

其中 cost 可以理解为该树加权错分率与对复杂度处罚值之和的代价函数, size 表示叶节点数量, 可以

用于衡量这个树分类器的复杂度, α 为一复杂度参数,

$\sum_{\text{叶节点}} i(n)$ 为所有叶节点的不纯度的求和, 表征了使用该分类树对训练样本进行分类时的不确定性。

根据公式 (4), 树剪枝可由以下两步完成:

(1) 在所有互为兄弟的叶节点中, 比较重新合并叶节点后 cost 值的变化;

(2) 删除使值最大减少的叶节点, 若 cost 值不减少, 则不作变化。

重复上述修剪过程, 直到修剪不能再进行。

在剪枝过程中, 训练误差随叶节点个数的增加而减少。测试误差则最初减少, 达到最小值, 然后由于训练数据对树的过分影响, 测试误差又逐渐增加。利用独立的测试数据集进行测试, 选择具有最小测试误差的子树作为最后决策树。本文选择一种启发式验证技术—交叉验证技术 (Cross validation) 来进行最优树的选择: 10—重交叉验证 (10-fold cross validation), 即训练集被随机划分为 10 组数量相等但不相交的数据子集, 分类器要训练 10 次, 每次采用 9 组数据子集进行训练, 余下的 1 组子集用作验证集 (Validation Set), 用于评价测试误差, 估计出的测试误差是 10 组误差的平均。

2.2 C4.5 基本原理

C4.5 为广泛使用的另一种单一决策树生成法, 采用“信息获取率 (Information Gain Ratio)”矩阵来实现分类。它利用训练集, 对每次选取信息获取率最大的但同时获取的信息增益又不低于所有属性平均值的属性作为树的节点, 将每一个可能的取值作为此节点的一个分支, 递归地形成决策树。树生成算法中也采用了 CART 中的“熵不纯度”函数, 而信息增益相当于 CART 中的“不纯度下降差”, 另外增加了“获取率”这一指标, 主要是为了去除高分支属性的影响而对信息增益的一种改进。“获取率”同时考虑了每一次划分所生成的子节点的个数和每个子节点的大小 (包含的数据实例的个数), 考虑的对象主要是一个个划分, 而不再考虑分类所蕴含的信息量。递归的结束条件是子集中的数据记录在主属性上取值都相同, 或没有属性可再供划分使用。

与 CART 不同的是, C4.5 利用了基于分支的统计显著性的误差概率技术来实现剪枝, 另一个显著差别是体现在对缺损模式的处理上, 在训练阶段, C4.5 并没有像 CART 以替代分支 (surrogate split) 来解决分类数据的缺损, 而是以概率加权的方法来处理“属性丢

失”的问题。

2.3 决策树中采用的新技术—Boosting 方法

在决策树分类器设计中, 一种于 20 世纪 90 年代中期在机器学习领域发展的被称之为“Boosting (增强法)”的技术被广泛用来提高分类精度。这种方法可以提高那些较难识别样本的分类准确率, 同时这种技术能降低分类算法对数据噪声和训练样本误差的敏感性。

Boosting 是一种提升任意学习算法准确率的集成学习方法, 可将准确率仅比随机猜测略好的弱学习算法提升为强学习算法, 它的思想来源于可能近似正确 (Probably Approximately Correct, PAC) 学习模型。它利用某种学习算法生成一系列的基分类器, 每个基分类器的训练依赖于在其之前产生的分类器的分类结果, 对训练失败的训练样本赋予较大的权值, 让学习算法在后续的学习中“更加重视”。最终分类器通过多个基分类器的加权投票得到最后的结果, 减少了单个分类器的误差, 提高了分类器的分类准确度。Freund 和 Schapire 在 1995 年根据 Boosting 的基本思想, 提出了最实用的 Boosting 算法—AdaBoost, 并得到了广泛的应用。

3 结论

决策树算法用于遥感数据分类的优势在于对数字特征分布极为复杂的数据集时, 决策树则属于严格“非参”, 对于输入数据空间特征和分类标识, 具有更好的弹性和鲁棒性。因此, 当遥感影像数据特征的空间分布很复杂, 或者源数据各维具有不同的统计分布和尺度时, 用决策树分类法能获得理想的分类结果。

决策树分类法的树状分类结构对数据特征空间分布不需要预先假设某种参数化密度分布, 所以其总体分类精度优于传统的参数化统计分类方法。随着人工智能技术的发展, 当前遥感影像分类的研究也向更高层次发展, 地学知识和地理信息的辅助决策可以大大提高遥感影像分类和信息提取的精度, 其中专家系统是解决这一问题的有效途径。因此, 将决策树算法与基于知识的专家系统相结合应当引起关注。

参考文献

- [1] 李爽, 丁圣彦. 决策树分类方法及其在土地覆盖分类中的应用[J]. 遥感技术与应用, 2002, 17(1): 6~11
- [2] 罗来平, 宫辉力. 遥感图像决策树分类器研究与实现[J]. 遥感信息, 2006, (3): 13~16

(下转第 42 页)

dUser (User user) 方法来进行添加用户的操作。

(3) 应用 Struts

首先要在 Web.xml 中配置控制器 ActionServlet, 然后在 JSP 页面 user_add.jsp 中, 结合 EL 表达式的取值方式和通过 userForm 来收集表单的数据, 如 \$ {userForm.userName}。在 JSP 页面中采用 JavaScript 脚本语言对表单的数据进行客户端的验证, 并以 “user.do?command=add” 提交请求给 ActionServlet 处理。ActionServlet 通过 Struts 的配置文件信息可以找到 UserActionForm 类和 UserAction 类, 从而调用 UserAction 类的 add 方法。

UserAction 类由于要完成多个相关的业务操作, 故采用继承 DispatchAction 类的方式, 其中定义了 add 方法, 相当于 Action 类中的 execute () 方法。

类 UserActionForm 中的属性基本与模型类 User 相一致, 并增加属性 pageNo 和 pageSize, 用于页面的分页。

DRP 分销管理系统是基于 SSH 架构的, 因此系统的灵活性和可扩充性都很好。若需要新的需求, 如增加功能模块的话, 只需编写相应的实体类、相应的业务逻辑接口和实现类, 以及相应的 Action 和 ActionForm 类, 然后添加新的 JSP 页面, 再增加相应的映射

文件和配置文件, 最后分别修改 Struts、Spring、Hibernate 的配置文件即可, 完全不须修改原有的代码, 因此不会对原有的功能模块有任何影响。

5 结束语

本文对当前企业的分销网络管理存在的问题进行了详细的分析, 给出一个基于 DRP 的信息管理系统来实现分销系统各方面的有效管理, 优化了企业分销资源的整体化的调配, 提高了资源使用效率。基于多层 Web 组织模式的 DRP 分销管理系统是将现实的管理流程和计算机技术完美结合起来, 通过先进的 SSH 架构来提高系统的易用性、扩充性、可配置性以及可维护性, 有效地减化管理的复杂度, 缩短企业的供应链, 为管理层提供全面的信息流, 减少了企业运营的成本, 提高了企业的利润。

参考文献

- [1] 霍力, 岑贤生, 于润东. 食品行业 DRP 系统设计与实现[J]. 食品与生物, 2008, (8)
- [2] 郑毅. DRP 系统概述及实现的关键因素[J]. 商业经济, 2005, (9): 117~124
- [3] 王爱民, 刘文华. 基于 Web 的分销供应链 DRP 系统的设计与实现[J]. 计算机工程与应用, 2005, (36)

(上接第 36 页)

- [3] 李飞雪, 李满春. 基于人工神经网络与决策树结合模型的遥感图像自动分类研究[J]. 遥感信息, 2003, (3): 23~25
- [4] 姜青香, 刘惠平. 利用纹理分析方法提取 TM 图像信息[J]. 遥感学报, 2004, 8(5): 458~464
- [5] Friedl M A, Brodley C E, Strahler A H. Maximizing Land Cover Classification Accuracies Produced by Decision Trees at Continental to Global Scales [J]. IEEE Transactions on Geoscience and Remote Sensing, 1999, 37(2): 969~977
- [6] 邱凯昌, 李德仁, 李德毅. 基于空间数据挖掘的遥感图像分类研究[J]. 武汉测绘科技大学学报, 2000, 125(1): 42~48
- [7] Melver D K, Friedl M A. Estimating Pixel-scale Land Cover Classification Confidence Using Non-parametric Machine Learning Methods[J]. IEEE Transaction on Geoscience and Remote Sensing,

2001, 39: 1959~1968

- [8] Melver D K, Friedl M A. Using Prior Probabilities in Decision-tree Remotely Sensed Data [J]. Remote Sensing of Environment, 2002, 81: 253~261
- [9] Zhan X, Sohlberg R A, Townshend J R G, et al. Detection of Land Cover Changes Using MODIS 250 m Data [J]. Remote Sensing of Environment, 2002, 83: 336~350
- [10] Rogan J, Franklin J, Roberts D A. A Comparison of Methods of Monitoring Multi-temporal Vegetation Change Using Thematic Mapper Imagery[J]. Remote Sensing of Environment, 2002, 80(1): 143~156
- [11] 李爽, 张二勋. 基于决策树的遥感影像分类方法研究[J]. 地域研究与开发, 2003, 22(1): 17~21