基于信息论的决策树算法探讨

张　彦，刘暾东，李茂青

（厦门大学 信息科学与技术学院 自动化系，福建　厦门　361005）

摘　要：信息论是数据挖掘技术的重要指导理论之一，是决策树算法实现的理论依据。决策树算法是一种逼近离散值目标函数的方法，其实质是在实例学习的基础上，得到分类规则。本文简要介绍信息论的基本原理，重点阐述基于信息论的决策树算法，分析了它们目前主要的代表理论以及存在的问题。

关键词：数据挖掘；信息论；决策树；信息熵

中图分类号：O236　文献标识码：A　文章编号：1003-7241(2006)01-0004-04

Decision Tree Algorithm Based on the Information Theory

ZHANG Yan，LIU Tun-dong，LI Mao-qing

(Department of Automation Xiamen University，Xiamen 361005，China)

Abstract: The information theory is one of the basic theories of Data Mining，and also is the theoretical foundation of the Decision Tree Algorithm. Decision Tree Algorithm is a method to approache the discrete-valued objective function. The essential of the method is to obtain a classification rule on the basis of example-based learning.

Key words: Data mining；Information theoretic；Decision tree；Entropy

1　引言

数据挖掘技术是通过统计论中的相关、聚类、回归、判别等方法，寻找数据中隐藏的关联、概念、及模式[1]。目前，数据挖掘算法种类繁多，一般可将其归纳为六大类[2]：信息论方法：利用信息论的原理来构造算法；集合论方法；仿生学方法；公式发现法；统计方法；其他方法(如模糊数学、可视化技术)。

其中，数据挖掘技术中的决策树算法是建立在信息论的基础上，是一种常用于预测模型的算法，它通过将大量数据有目的地分类，从中找到一些具有价值的、潜在的信息。该算法的理论依据充分，具有较高的精度和效率，是一种知识获取的有用工具，因而成为近年来数据挖掘的研究新热点。决策树算法起源于概念学习系统 CLS(concept learning system)，随后 ID3 算法成为一个里程碑[3][4]，然后又演化为能处理连续属性的 C4.5 和 C5.0[5]。此外，目前经典的决策树算法还有 CART (Classification And Regression Trees)[6] 和 CHAID (Chi squared Automatic Interaction Detection)[7]。在此基础上人们又做了大量的改进和变种，如 FACT[8]、GINI[9]、SEE5[10]、IBLE[11]、SLIQ[12]-[14]、LBET[15] 和 SPRINT[16] 算法，但基本还是以前面的算法为代表。这些算法都支持分类方法，少数还可以用于回归方法[17]。根据分割方法的不同，这些决策树算法可分为两类：基于信息论的(Information Theory)的方法和最小 GINI 指标(lowest GINI index)方法。对应前者的经典算法有 ID3、C4.5 和 IBLE。基于这三种经典算法，人们又提出了自己的改进算法，如基于概率的决策树构造算法 PID[18]，MID3 算法[19]，LBET 方法，MedGen 算法[20] 等。

收稿日期：2005-08-25

2　信息论原理简介

信息论是 20 世纪中叶从通信中发展起来的理论，是概率论与通信技术相结合的边缘学科。自香农在 1948 年发表奠定信息论基础的“通信的数学理论”一文以来，信息论学科迅速发展并已延伸到许多领域中。

2.1　信息论基本概念[21][22]：

(1) 离散信源数学模型：消息 u_i ($i=1, 2, 3, \cdots r$)的发生概率 $P(u_i)$ 组成的信源模型

$$[U, P]=\begin{bmatrix} u_1 & u_2 & \mathrm{L} & u_r \\ P(u_1) & P(u_2) & \mathrm{L} & P(u_r) \end{bmatrix},\ P(Ui)\geqslant 0,\ \text{且满足}\ \sum_{i=1}^{r} P(u_i)=1 \qquad (1)$$

(2) 互信息 $I(U, V)$：$H(U)$ 代表接收到输出符号集 V 以前

关于输入符号集 U 的平均不确定，而 $H(U|V)$ 代表接收到符号集 V 后关于输入符号 U 的平均不确定性。

$$I(U,V)=H(U)-H(U|V)=\log\frac{1}{p(ui)}-\log\frac{1}{p(ui|vi)} \quad (2)$$

(3) 信道容量 C: 给定信道的互信息 $I(U,V)$ 是 $P(U)$ 的上凸函数，由上凸函数的性质知道，一定存在一概率分布使得 $I(U,V)$ 达到最大。这个最大的互信息就称为信道容量，记为 C。

$$C=\max_{P(U)}\{I(U,V)\} \quad (3)$$

(4) 信息增益率 gain _ ration: 实验证明采用“信息增益率”比采用“信息增益”更好，能够克服 ID3 方法选择偏向于取多值的属性的不足。

$$\text{gain\_ratio}=I(C,V)/H(V) \quad (4)$$

2.2 决策树算法的信息论原理[23]

决策树算法是利用信息论原理对大量样本的属性进行分析和归纳而产生的。对于一个给定的样本分类所需的期望信息:

$$I(s_1,s_2,\cdots,s_m)=-\sum_{i=1}^{m}p_i\log(p_i) \quad (5)$$

在树的每个节点上使用信息增益来度量选择测试属性，则由该测试属性划分成子集的熵或期望信息:

$$E(A)=\sum_{i=1}^{m}\frac{s_{1j}+s_{2j}+\cdots+s_{mj}}{s}I(s_{1j},s_{2j},\cdots,s_{mj}) \quad (6)$$

决策树的根结点是所有样本中信息量最大的属性。树的中间结点是该结点为根的子树所包含的样本子集中信息量最大的属性。即取下式(信息编码)最大:

$$Gain(A)=I(s_1,s_2,\cdots,s_m)-E(A) \quad (7)$$

3 基于信息论的决策树算法

决策树算法实质是在实例学习的基础上，得到分类规则。具体地说，决策树学习是一种逼近离散值目标函数的方法，该方法将训练数据中学习到的函数表示为树[24]。决策树以图形或文本形式的规则来描述或预测数据，但一般只能有一个依赖变量，多个依赖变量需要多个决策树；且数据类型一般不能是连续的[25]。

3.1 ID3 算法

自 1986 年 J. R. Quinlan 在 Machine Learning Journal 发表题为“Induction of decision trees”的论文，引入一种新的 ID3 算法后，ID3 算法成为了国际上最早、最有影响的决策树方法。该方法实质是构造一个熵值下降平均最快的判定树[26]。该算法基本思想是: 将一棵决策树看作一个信源，通过依次将每一个特征的不同特征值作为信宿的表现状态，来判断哪一个特征的信息量最大，从而得到对产生分类规则最关键的特征[17][27]。

ID3 是一个典型的决策树学习系统，它以信息熵作为分离目标评价函数，采用自顶向下不可返回的策略，搜出全部空间的一部分，它确保决策树建立最简单，每次所做的测试数据最少。ID3 算法的基础理论清晰，使得算法较简单，学习能力较强，且构造的决策树平均深度较小，分类速度较快，特别适合处理大规模的学习问题。

但是 ID3 算法也存在着缺点[5][28]:

(1) ID3 算法注意力集中在特征的选择上，且偏向于选择特征值数目较多的特征，而特征值数目较多的特征却不总是最优的特征，这样不太合理。

(2) 用互信息作为特征选择量上存在一个假设，即训练例子集中的正、反例的比例应该与实际问题领域里正、反例的比例相同。一般情况下，不能保证相同，这样计算训练集的互信息就存在偏差。

(3) ID3 对噪声较为敏感，训练集中正例与反例的比例很难控制。

(4) 学习简单的逻辑表达能力差。

(5) 当训练集增加时，ID3 的决策树会随之变化。这对渐进学习是不方便的。

(6) ID3 在建树时，每个节点仅含一个特征，特征之间的相关性强调不够。

ID3 算法适用于数量较大的决策判断系统和大型的数据库系统。在这些系统中，其优势将会得到更好的体现。ID3 引入后不久，Schlimmer 和 Fisher 在 ID3 的基础上构造了 ID4 算法，允许递增式地构造决策树。1988 年，Utgoff 也提出 ID5 算法，它允许通过修改决策树来增加新的训练实例，而无需重建决策树。

3.2 C4.5 算法

在 ID3 算法的基础上，J. R. Quinlan 于 1993 年在其“Programs for Machine Learning”一书中，对 ID3 算法进行了补充和改进，提出了又一流行的 C4.5 算法。C4.5 算法继承了 ID3 全部优点，且克服了 ID3 在应用中的不足，主要体现在以下几方面[29]:

(1) 用信息增益率来选择属性，克服了 ID3 用信息增益选择属性时偏向于选择取值多的属性的不足;

(2) 在树构造过程中或者构造完成之后，使用不同的修剪技术以避免树的不平衡;

(3) 能够完成对连续属性的离散化处理;

(4) 能够对不完整数据进行处理;

(5) K 次迭代交叉验证;

(6) C4.5 采用的知识表示形式为决策树，并能最终可以形成产生规则。

此外，C4.5 算法可通过使用不同的修剪技术以避免树的不平衡。即通过剪枝操作删除部分节点和子树以避免“过度适合”，以此消除训练集中的异常和噪声。

C4.5 算法代表着基于决策树的方法的里程碑[30]。但是，C4.5 算法同样存在不足:①C4.5 算法采用分而治之的策略所得到决策树不一定是最优的;②采用一边构造决策树，一边进行评价的方法，使决策树的结构调整、性能改善较困难;③仅考虑决策树的错误率，未考虑树的节点、深度，而树的结点个数代表了树的规模，树的平均深度对应着决策树的预测速度;④对属性值分组时逐个探索，没有一种使用启发式搜索的机制，分组效率较低[5];⑤Quinlan 经典的展示 C4.5 算法结果的方法，是将结果树逆时针旋转 90 度，以文本形式输出，很不直观[31]。

C4.5算法特别适用于挖掘数据量多，且对效率和性能要求高的场合。C5.0算法是C4.5的商业改进版，它利用boosting技术把多个决策树合并到一个分类器，使得在大数据量情况下，效率和生成规则的数量与正确性都有显著的提高。

3.3 IBLE算法

国内于90年代初，研究出基于信道容量的IBLE(Information－Based Learning from Example)算法[10]，较之ID3每次只选一个特征作为决策树的结点的方法，IBLE算法选一组重要特征建立规则，更有效地正确判别，克服了ID3依赖正反例比例的缺点[32]。

IBLE算法的基本思想是利用信道容量，寻找数据集中信息量从大到小的多个字段，并由这些字段取值来建立决策规则树的一个结点。根据该结点中指定字段取值的权值之和与两个阀值比较，建立左、中、右三个分枝。在各分枝子集中重复建树结点和分枝过程，这样就建立了决策规则树。

IBLE算法的优点在于它不依赖类别先验概率，特征间为强相关，具有直观的知识表示[10]，获得的知识与专家知识在表示和内容上有较高的一致性。因此，IBLE算法特别适合于处理大规模的学习问题，其形成系统可作专家系统的知识获取工具[15]。

3.4 改进的决策树算法

以ID3为代表构造决策树的算法把研究重点放在属性的选择上，这一研究方式受到了许多有关学者的关注与怀疑。针对这一情况，人们都在此基础上提出了自己的改进思想。

洪家荣等从事例学习最优化的角度[33－35]分析了决策树归纳学习的优化原则，提出了一种新的基于概率的决策树构造算法PID[18]。PID在决策树的规模和精度方面优于ID3，但是在训练速度和测试速度上比ID3慢，并且PID决策树上的某些属性可能重复使用。针对ID3算法选择属性较多的属性这一缺点，刘小虎等[19]提出了ID3算法的优化算法－MID3算法。该算法改进了选择新属性的启发式函数，取得了比ID3更好的分类效果。MID3算法还引入概率按最小信息熵对候选属性的扩展来弥补学习简单的逻辑表达能力差这一缺点。较之传统的ID3算法，文[15]中提出的LBET方法的最大优点在于接受和记忆数据的信息量增加。而曲开社等人就Quinlan的ID3算法中信息熵标准有倾向于取值较多的属性的缺陷，在计算信息熵时引入了用户兴趣度，改进了ID3算法，使决策树减少了对取值较多的属性的依赖性[32]。同样，王静红和李笔为了克服ID3算法偏向于选择取值多的，但在实际问题中对分类意义并不大的属性作为测试属性的缺点，引入了选取优值法的概念来对ID3算法进行改进[36][37]。此外，对于Quinlan的ID3算法中采用的信息增益并非最优启发式这一缺点，吴艳艳提出了将决策树的基本建树思想ID3算法与对象决策属性化简的粗集理论相结合的一种新型的决策树建树方法，该方法的提出使数据挖掘的效果更简单、更容易理解[38－39]。

MedGen算法是由Micheline K等人在1997年提出的[20]，它基于C4.5算法的思想，在决策树的生成和评测上均采用C4.5的方法，但在选择属性建立决策树之前，采用面向属性规约的方法和相关分析方法，在此基础上构建决策树。对于C4.5算法的不足，肖勇等提出了利用遗传算法构造决策树的算法。该算法的过程，是利用遗传算法从上一代决策树群体经过遗传算子的操作，产生下一代群体演化直到满足终止条件。虽然遗传算法的处理比较费时，但是该算法在各方面均比C4.5有提高[40]。

4 基于信息论决策树算法的展望

基于信息论的决策树技术已经取得了较大的发展，已被广泛地应用[5][17][41][42][43－44]，其主要特点如下：

(1) 产生的规则能够轻易地转化为可以理解的规则，如转换为“IF－THEN”这种形式的关联规则，也可以被翻译成自然语言或SQL语句[45]；

(2) 在进行分类时所需的计算量不大；

(3) 该算法具有较高的分类精确度；

(4) 模型效率高，对训练集数据量较大的情况较为适合；

(5) 既支持离散数据也支持连续数据；

(6) 能够清楚地指出哪一个数据域对于预测或决策是最重要的；

(7) 不需要对数据的性质做预先的假设，也不需要受训数据外的知识；

(8) 能够使用包含数值型和分类型数据的数据集建立模型；

(9) 基于信息论的决策树已经成功地应用于现实问题[46]。

当然，基于信息论的决策树算法还存在许多问题，仍然需要在很多方面进一步研究、发展、解决、改进，主要包括：

(1) 实验证明了要找到这种最优的决策树是NP问题，必须寻找各种启发算法以寻求较优的解[5][47]。

(2) 决策树的好坏，不仅影响了分类的效率，而且影响分类的准确率。更优的启发式函数和评价函数有待进一步研究[48]。

(3) 理想的决策树分为三种：叶结点数最少；叶子结点深度最小；叶结点数最少且叶子结点深度最小。力求理想的决策树算法是今后发展的新方向[27]。

5 结论

本文简要介绍信息论的基本原理，重点阐述基于信息论的决策树算法，分析了它们目前主要的代表理论以及存在的问题，提出了对信息论决策树算法的展望。笔者曾利用基于信息论的决策树算法解决天气分类问题，并将该算法应用于教师课堂教学评估系统，受益于信息论决策树算法的优点，取得了良好的效果。随着信息资源日益增长，决策树算法和信息理论研究的深入，两者必将更加紧密结合在一起，多种方法的集成将是数据挖掘发展的一个方向。

6 参考文献:

[1] JIAWEI HAN. 数据挖掘—概念与技术[M]. 北京: 机械工业出版社, 2001.

[2] 薛禾生. 知识发现与数据挖掘[J]. 中国医学统计, 2000, 7(3): 164—165.

[3] J. R. QUINLAN. Induction of decision trees[J]. Machine Learning, 1986, 1(1): 81—106.

[4] J. R. QUINLAN. Programming for Machine Learning[M]. CA: San Mateo, 1993.

[5] J. R. QUINLAN. C4.5: Programs for machine learning[M]. Los Altos, California: Morgan Kaufmann Publishers, Inc. 1993.

[6] 邵华, 赵宏. 一种与神经元网络杂交的决策树算法[J]. 小型微型计算机系统, 2001, (8): 964—966.

[7] 陈文伟, 黄金才编著. 数据仓库与数据挖掘[M]. 人民邮电出版社, 2004.

[8] L. BRIEMAN, J. FRIEDMAN, R. OLSHEN, and C. STONE. Classification of regression tress[M]. Wiley Computer Publishing, 1997.

[9] PEILEI TU, JENYAO CHUNG. A New Decision Tress Classification Algorithm for Machine Learning[A]. Proc. of the 1992 IEEE Int. Conf. On Tools with AI Arlington[C], VA, Nov. 1992: 370—377.

[10] ROIGER R J, A ZARBOD C, SANT R. A majority rules approach to data mining[A]. In: Proc of Intelligent Information Systems IIS' 97[C]. Grand Bahama Island, 1997, 100—107.

[11] 杨林, 富元斋, 等. 基于神经网络的分类算法的改进[J]. 计算机工程与应用, 2002, (5): 71—73.

[12] USAMA M F GREGORY P S PADHRAIC S. Advances in Knowledge Discovery and Data Ming[M]. Massachusetts: The MIT Press. 1996.

[13] BRACHMAN R. J, ANAND T. The process of knowledge discovery in databases. In Advances in Knowledge Discovery and Data Ming[M]. AAA MIT Press, 1996.

[14] FAYYAD. U. M, PIATETSKY ShHAPIRO. G. From data Ming to knowledge discovery: an overview A[M]. In Advances in knowledge Discovery and Data Ming AAA I Press and the MIT Press, 1996.

[15] 徐凌宇, 杜庆东, 等. 一种基于互信息的特征跃迁示例学习法[J]. 2001, (6): 653—657.

[16] DAVID HAND, HEIKKI MANNILA, PADHRAIC SMYTH 著. 银奎, 廖丽, 宋俊, 等译. 数据挖掘原理[M]. 机械工业出版社, 2003.

[17] 姚晔, 李翔. 决策树算法的教育应用探讨[J]. 江西师范大学学报, 2004, (8): 317—320.

[18] 洪家荣, 丁明峰, 等. 一种新的决策树归纳学习方法[J]. 计算机学报, 1995, 18 (6): 471—474.

[19] 刘小虎, 李生. 决策树的优化算法[J]. 软件学报, 1998, 9 (10): 797—800.

[20] KAMBER M, WINSTONE L, GONG W. Generalization and Decision Tree Induction: Efficient Classification in Data Mining[A]. In: Proc. of 1997 Int. Workshop Research Issues on Data Engineering[C]. Birmingham (England), April 1997, 111—120.

[21] 周荫清编著. 信息理论基础[M]. 北京: 北京航空航天出版社.

[22] 唐朝京, 雷菁编著. 信息论与编码基础[M]. 北京: 国防科技大学出版社, 2002.

[23] RICHARD J. ROIGER, MICHAEL W. GEATZ. 著. 翁敬农译. 数据挖掘教程[M]. 北京: 清华大学出版社, 2003.

[24] KASS, G. V. An exploratory technique for investigating large quantities of categorical data[J]. Applied Statics, 1980, (29): 119—127.

[25] AGRAWAL R, MANNILA H, SCRIKEANT R, TOIVONEN H, VERKAMO AI. Fast Discovery of Association Rule[A]. In Fayyad et al. (eds), Advances in Knowledge Discovery and data Mining[C]. AAAI Press/ The MIT Press, 1996, 307—328.

[26] 高波, 陈广学, 等. EIS 环境下的数据挖掘技术的研究[J]. 华中理工大学学报, 1998, (5): 78—80.

[27] 谭旭, 王丽珍, 等. 利用决策树发掘分类规则的算法研究[J]. 云南大学学报, 2000, 22(6): 415—416.

[28] MOTOHIDE UMANO, HIROTAKA OKAMOTO, ITSUO HATONO, HIROYUKI TAMURA. Fuzzy Decision Trees by Fuzzy ID3 Algorithm and Its Application to Diagnosis Systems[A]. Proceedings of the 2004 IEEE International Conference on e—Technology, e—Commerce and e—Service (EEE' 04)[C].

[29] 邵峰晶, 于忠清. 数据挖掘原理[M]. 北京: 中国水利出版社, 2003.

[30] 朱明. 数据挖掘[M]. 合肥: 中国科技大学出版社, 2002.

[31] HOLLAND J H. Adaptation in Natural and Artificial System[M]. Cambridge, MA: MIT Press, 1992.

[32] 曲开社, 成文丽, 等. ID3 算法的一种改进算法[J]. 计算机工程与应用, 2003, (25): 104—107.

[33] AKESHI F, YASUHIKO M, MORISHITA S. Constructing efficient decision trees by using optimized numeric association rules[A]. Proceedings of 22nd VLDB Conference. India: Very Large Data Base Endowment[C], 1996: 146—155.

[34] JOHANNES G, RAGHU R, GANTI VENKATESH. RAIN FOREST. A Framework for Fast Decision Tree Construction of Large Datasets[A]. Proceedings of 24th VLDB conference. New York: Very Large Data Base Endowment[C], 1998: 417—427.

[35] YASUHIKO M, TAKESHI F, MATSUZAWA HIROFUMI. Algorithms for Mining Association Rules for Binary Segmentations of Huge Categorical Databases[A]. Proceedings of 24th VLDB conference[C]. New York: Very Large Data Base Endowment, 1998: 381—391.

[36] 王静红, 李笔. 基于决策树的一种改进算法[J]. 电讯技术. 2004, (5): 175—178.

[37] 王静红, 王熙照, 等. 决策树算法的研究及优化[J]. 微机发展, 2004, (9): 30—32.

[38] SHUTZU TSAI, CHAOTUNG YANG. Decision Tree Construction for Data Mining on Grid Computing[A]. 2004 IEEE International Conference[C] on 28—31 March 2004 Page(s): 441—447.

[39] 吴艳艳. 粗集结合决策树的一种数据挖掘算法[J]. 计算机工程与科学, 2004, 26(2): 60—62.

[40] 肖勇, 陈意云. 用遗传算法构造决策树[J]. 计算机研究与发展, 1998, 35 (1): 49—52.

[41] 周晓宇, 李慎之, 等. 数据挖掘技术初探[J]. 小型微型计算机系统, 2002, 23(3): 342—346.

[42] S. J. YEN, A. L. P. CHEN. An efficient approach to discovering knowledge from large databases[A]. In 4th Intl Conf. Parallel and Distributed Info. Systems[C], December 1996.

[43] SRIKANT R, AGRAWAL R. Mining generalized association rules [A]. In: Proc. 21. VLDB[C], 1995: 407—419.

[44] 中国人民大学统计学系数据挖掘中心. 数据挖掘中的决策树技术及其应用[J]. 统计与信息论坛, 2002, 17(2): 4—10.

[45] CLAUDE SEIDMAN 著. 刘艺, 王鲁军, 将丹丹, 等译. SQL Server 2000 数据挖掘技术指南[M]. 北京: 机械工业出版社, 2002.

[46] L. WEHENKEL, M. PAVELLA. Advances in decision trees applied to power system security assessment[A]. IEE 2nd International Conference on Advances in Power System Control, Operation and Management[C], December 1993, HongKong. 47—53.

[47] HYAFIL L, RIVEST R L. Constructing Optimal Binary Decision Trees Is NP—Complete[J]. Info Proc Letters, 1976, 5(1): 15—17.

[48] JIAWEI HAN, MICHELINE KAMBER 著, 范明, 孟小峰译. 数据挖掘概念与技术[M]. 2001: 185—218. 54—61.

作者简介: 张彦(1980—), 女, 硕士研究生, 主要研究方向为数据挖掘技术、数据仓库、神经网络。