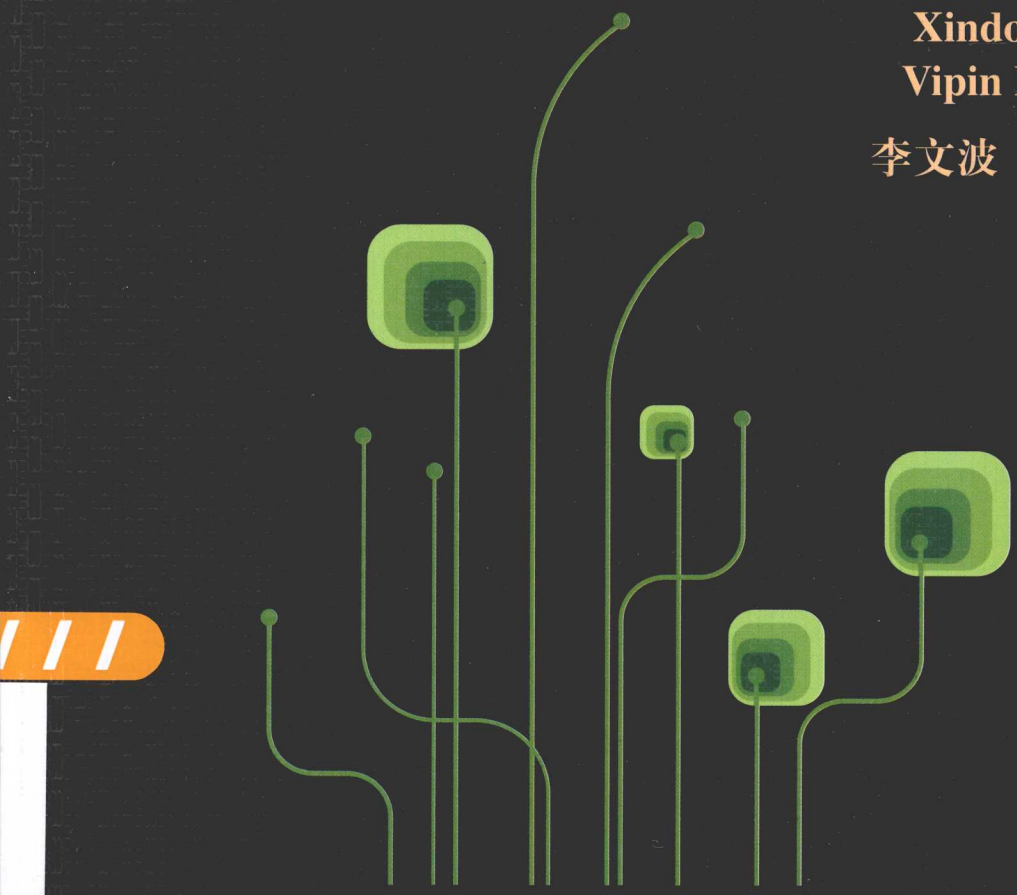
CRC Press
Taylor & Francis Group

世界著名计算机教材精选

数据挖掘十大算法

Xindong Wu
Vipin Kumar 编著

李文波 吴素研 译



THE TOP TEN ALGORITHMS IN DATA MINING

清华大学出版社

世界著名计算机教材精选

数据挖掘十大算法

Xindong Wu
Vipin Kumar 编著
李文波 吴素研 译

清华大学出版社
北 京

The Top Ten Algorithms in Data Mining
by Xindong Wu, Vipin Kumar
EISBN: 978-1-4200-8964-6

图书在版编目（CIP）数据

数据挖掘十大算法/(美)吴信东，(美)库玛尔编著；李文波，吴素研译．—北京：清华大学出版社，2013.5
书名原文：The top ten algorithms in data mining
(世界著名计算机教材精选)
ISBN 978-7-302-31061-7

Ⅰ．①数…　Ⅱ．①吴…　②库…　③李…　④吴…　Ⅲ．①数据采集　Ⅳ．①TP274

中国版本图书馆 CIP 数据核字(2012)第 304133 号

责任编辑：龙启铭
封面设计：傅瑞学
责任校对：焦丽丽
责任印制：沈　露

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社　总　机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：北京富博印刷有限公司
装 订 者：北京市密云县京文制本装订厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：10.5　　**字　　数**：256 千字
版　　次：2013 年 5 月第 1 版　　**印　　次**：2013 年 5 月第 1 次印刷
印　　数：1～3000
定　　价：39.00 元

产品编号：046002-01

译者序

数据挖掘这一学科近年来发展十分迅速，不仅产生了大量不同类型的挖掘算法，而且也表现出与机器学习等学科深度融合的态势。无论是从事研究的专家学者还是从事应用的开发人员都十分希望能一窥其大略，从而比较准确地把握数据挖掘领域当前的主干技术，并比较全面地了解当前的发展趋势。

当前，在市场上流通的数据挖掘方面的著作已经不算少了，主要是两大类：一类是具有完整体系的教材类图书，一类是面向特定领域的应用型图书。前者主要是服务教学，所以侧重原理、逻辑严谨，但是通常对数据挖掘的前沿介绍比较欠缺。后者往往集中于介绍某一领域的问题和方法，或者是关于某些典型工具的使用方法，其优点在于直观有效，但相对于整个数据挖掘领域其覆盖面偏小。

为此，很有必要对整个数据挖掘领域的近期发展和前沿成果进行梳理，而这一类信息往往散见于相关的大量学术期刊和会议文集中，限于视野和精力，任何个人都难以完成这一任务。在此基础上，还需要对当前庞大的数据挖掘知识体系进行恰当的取舍和凝练，这一工作必须依靠该领域的高水平学者。所以，国际数据挖掘社区合众人之力，在 2006 年推出了 *The Top Ten Algorithms in Data Mining* 这一继往开来之作。该书列举了评选出来的十个最具影响力的数据挖掘算法：C4.5、k-means、SVM、Apriori、EM、PageRank、AdaBoost、kNN、Naive Bayes 和 CART。我们认为该书有其鲜明特色：

第一，立意承前启后，推出的时机恰当。该书的内容涵盖了分类、聚类、统计学习、关联分析和链接分析等重要主题在近年来的发展，这不但对数据挖掘的研究和发展十分重要，也将数据挖掘推动到更大范围的真实应用中，激励更多数据挖掘领域的学者对这些算法的作用和新问题进行深入探索。

第二，汇集群体智慧，具有很高权威性。参评人员囊括了历届 ACM KDD 创新奖和 IEEE ICDM 研究贡献奖得主这些顶尖学者，以及 SIGKDD、ICDM 和 SDM 这三大数据挖掘学术会议的程序委员会的全体委员。此外，还组织了专题会邀请了一百多位领域专家进行开放研讨。

第三，执行过程严谨，确保内容高品质。第一阶段是由顶尖学者推荐算法并提供算法名称、简要理由和代表文献这些必要信息，第二阶段用 google scholar 对每个提名算法进行客观地引用验证和排序，第三个阶段由数据挖掘社区的专家和相关领域的专家进行投票，获得完全一致的结果。最后，邀请资深学者撰写上榜算法的介绍并集结成书。

本书的翻译工作由中科院软件研究所李文波和北京市科学技术情报研究所吴素研共同完成，我们非常希望能为国内数据挖掘方面的工作略尽一点绵薄之力，但是由于水平有限，译作难免有错漏之处，请不吝指出。

最后，我们还要感谢国家自然科学基金（项目编号：61003117）、ISTIC-Thomson Reuters 科学计量学联合实验室开放基金（项目编号：IT2011003）、国家 863 计划（项目编号：2013AA01A603）和国家软科学计划（项目编号：2009GXQ6D154）对我们工作的支持。

前　言

在香港举办的2006年度IEEE数据挖掘国际会议(ICDM, http://www.cs.uvm.edu/~icdm/)上,与会专家遴选出了十个最具影响力的数据挖掘算法,也就是本书所列的十个算法:C4.5、k-means、SVM、Apriori、EM、PageRank、AdaBoost、kNN、Naive Bayes和CART。

遴选过程第一步,在2006年9月,我们邀请ACM KDD创新奖得主和IEEE ICDM研究贡献奖得主每人推荐十个最著名的数据挖掘算法,并提供以下信息:

(a) 算法名称;

(b) 算法简介;

(c) 代表文献。

我们还要求每个被提名的算法都应被数据挖掘领域的学者广泛引述和使用,每位推荐人提名的算法集应能代表数据挖掘的不同领域。除一人外其他所有专家都给予了回复。

遴选过程第二步,在2006年10月,我们用Google Scholar对每项提名进行了验证,去除了引用数低于50的提名,将保留下的所有提名(共18个)分成十个主题:关联分析、分类、聚类、统计学习、装袋推举、序列模式、集成挖掘、粗糙集、链接挖掘和图挖掘。对于某些算法,如k-means,不要求提供发明该算法的原始文献,但需要提供阐述该算法重要性的近期论文。可从ICDM站点(http://www.cs.uvm.edu/~icdm/algorithms/CandidateList.shtml)上找到这些代表性文献。

遴选过程第三步,我们动员了研究社区的很多人参与,其中包括KDD-06 (the 2006 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining)、ICDM'06 (the 2006 IEEE International Conference on Data Mining)和SDM'06 (the 2006 SIAM International Conference on Data Mining)的程序委员会的全体委员,以及ACM KDD创新奖得主和IEEE ICDM研究贡献奖得主。请每位参与人员从18个候选算法中选出不超过10个最知名算法,结果在ICDM'06的"数据挖掘十大算法"专题研讨会上公布。

2006年12月21日,在ICDM'06的一个专题讨论会上,邀请145名与会专家对这18个候选算法公开投票,从中选出十个得票最高的算法,得到和上面遴选第三步完全一致的结果。这个3小时的专题研讨会是ICDM'06的一个环节,在同一地点并行召开的还有Web Intelligence(WI'06)和Intelligent Agent Technology(IAT'06)的共7个论文展示环节,共吸引到了145名学者参与。

在ICDM'06之后,我们邀请了这十大算法的原创作者和专题研讨会部分发言人共同撰写了一篇期刊论文对每个算法的内容、影响进行介绍,对其现状和未来趋势加以评述。这篇期刊论文于2008年1月发表在*Knowledge and Information Systems*[1]上。本书是该期刊论文的扩展,每章介绍一个算法,内容包括算法描述、可用软件、示例应用、高级主题和习题等部分。

本书的每一章都邀请两位独立审稿人和本书的一位编辑来审核,有的章节在此基础上

还要在最终定稿前再重审一遍。

我们希望这十个算法的遴选能有助于在世界范围推动数据挖掘的应用，激励更多数据挖掘领域的学者去扩大这些算法的影响，探索新的研究内容。这十个算法覆盖了分类、聚类、统计学习、关联分析和链接分析等重要的数据挖掘研究和发展主题，也对数据挖掘、机器学习和人工智能等学科的课程设计有指导意义。

致　　谢

遴选十大数据挖掘算法缘起于2006年5月曹建农博士和吴信东博士在香港理工大学的一次学术研讨。此间，吴博士做了《数据挖掘研究十大挑战》的报告[2]，之后，Vipin Kumar博士又在KDD-06大会上广邀众学者就算法遴选工作进行了探讨，一时应者云集。

明尼苏达大学计算机科学与工程系职员Naila Elliott为三轮遴选过程收集编纂了算法提名和投票结果。佛蒙特大学计算机科学系职员Yan Zhang负责将十份不同格式的提交文档转换成同一LaTex格式，这是一个非常耗时的工作。

参 考 文 献

[1] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg. Top 10 algorithms in data mining, *Knowledge and Information Systems*, 14(2008), 1: 1-37.

[2] Qiang Yang and Xindong Wu (Contributors: Pedro Domingos, Charles Elkan, Johannes Gehrke, Jiawei Han, David Heckerman, Daniel Keim, JimingLiu, David Madigan, Gregory Piatetsky-Shapiro, Vijay V. Raghavan, RajeevRastogi, Salvatore J. Stolfo, Alexander Tuzhilin, and Benjamin W. Wah). 10 challenging problems in data mining research, *International Journal of Information Technology & Decision Making*, 5, 4(2006), 597-604.

关于作者

吴信东(Xindong Wu)教授 英国爱丁堡大学人工智能学博士,任美国佛蒙特大学计算机科学系主任。吴教授在数据挖掘、知识系统和 Web 信息开发等研究领域内颇有建树,在 IEEE TKDE、TPAMI、ACMTOIS、DMKD、KAIS、IJCAI、AAAI、ICML、KDD、ICDM 和 WWW 等学术会议和期刊上发表了 170 余篇学术论文,另外,还出版了 18 部学术专著和会议文集。他还获得了 IEEE ICTAI-2005 的最佳论文奖和 IEEE ICDM-2007 的最佳理论/算法论文奖亚军。

吴博士是 *IEEE Transactions on Knowledge and Data Engineering* (TKDE,由 IEEE Computer Society 主办)的主编,*IEEE International Conference on Data Mining* (ICDM)的创始人和指导委员会主席,*Knowledge and Information Systems* (KAIS,由 Springer 发行)的创办人和荣誉主编,IEEE Computer Society Technical Committee on Intelligent Informatics (TCII)的创始主席(2002—2006),*Springer Advanced Information and Knowledge Processing* (AI&KP)系列著作的编辑。他还是 ICDM'03 (the 2003 IEEE International Conference on Data Mining)程序委员会主席和 KDD-07 (the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining)程序委员会联合主席。他获得了 2004 ACM SIGKDD 服务奖、2006 IEEE ICDM 杰出服务奖,是 2005 年合肥科技大学"长江学者奖励计划"讲座教授。他还是很多学术会议的特邀专家/专题报告人,如 NSF-NGDM'07、PAKDD-07、IEEE EDOC'06、IEEE ICTAI'04、IEEE/WIC/ACM WI'04/IAT'04、SEKE 2002 和 PADD-97 等。

Vipin Kumar 教授 明尼苏达大学计算机科学与工程系 William Norris 讲席教授、系主任。他于 1977 年获得印度鲁尔基理工学院(正式名称是鲁尔基大学)的电子和通信工程学士学位,1979 年获得荷兰埃因霍温飞利浦国际学院的电子工程硕士学位,1982 年获得马里兰大学帕克分校的计算机科学博士学位。Kumar 教授的研究兴趣主要集中在数据挖掘、生物信息学和高性能计算领域。他提出了评估并行算法可扩展性的恒等效率度量指标,并研发了多款稀疏矩阵分解(PSPASES)和图剖分(METIS, ParMetis, hMetis)的高效并行算法及软件。他发表了 200 多篇研究论文,合编合著了 9 本学术专著,包括被广泛使用的教科书 *Introduction to Parallel Computing* 和 *Introduction to Data Mining*,都由 Addison-Wesley 出版。Kumar 是众数据挖掘和多并行计算领域的学术会议、专题研讨会的主席或共同主席,如 *IEEE International Conference on Data Mining* (2002)、*International Parallel and Distributed Processing Symposium* (2001) 和 *SIAM International Conference on Data Mining* (2001)。Kumar 是 SIAM International Conference on Data Mining 指导委员会共同主席,IEEE International Conference on Data Mining 和 IEEE International Conference on Bioinformatics and Biomedicine 指导委员会委员。Kumar 是 *Journal of Statistical Analysis and Data Mining* 的创始主编之一,*IEEE Intelligent*

Informatics Bulletin 主编和 *Data Mining and Knowledge Discovery* 系列图书(由 CRC Press/Chapman Hall 出版)的编辑。Kumar 还担任很多其他学术刊物的编辑,如 *Data Mining and Knowledge Discovery*、*Knowledge and Information Systems*、*IEEE Computational Intelligence Bulletin*、*Annual Review of Intelligent Informatics*、*Parallel Computing*、*Journal of Parallel and Distributed Computing*、*IEEE Transactions of Data and Knowledge Engineering* (1993—1997)、*IEEE Concurrency* (1997—2000)和 *IEEE Parallel and Distributed Technology* (1995—1997)等。他是 ACM 会士、IEEE 会士、AAAS 会士和 SIAM 会员。Kumar 由于在并行算法设计、图剖分和数据挖掘领域的杰出贡献,获得了 2005 IEEE Computer Society 的技术成就奖。

贡献人员

陈松灿，南京航空航天大学，中国
Joydeep Ghosh，得克萨斯大学奥斯汀分校，得克萨斯州奥斯汀
David J. Hand，帝国理工学院，英国伦敦
Alexander Liu，得克萨斯大学奥斯汀分校，得克萨斯州奥斯汀
刘兵，伊利诺依大学芝加哥分校，伊利诺伊州芝加哥
Geoffrey J. McLachlan，昆士兰大学，布里斯班，澳大利亚
Hiroshi Motoda，大阪大学 ISIR 研究所和 AFOSR/AOARD 空军研究实验室，日本
Shu-Kay Ng，格里菲斯大学，澳大利亚梅多布鲁克
Kouzou Ohara，大阪大学 ISIR 研究所，日本
Naren Ramakrishnan，弗吉尼亚理工，弗吉尼亚州布莱克斯堡
Michael Steinbach，明尼苏达大学，明尼苏达州
Dan Steinberg，Salford Systems 公司，加利福尼亚州圣地亚哥
陈封能，密歇根州立大学，密歇根州东兰辛
薛晖，南京航空航天大学，中国
杨强，香港科技大学，香港九龙清水湾
Philip S. Yu，伊利诺依大学芝加哥分校，伊利诺伊州芝加哥
俞扬，南京大学，中国
周志华，南京大学，中国

目　录

第 1 章　C4.5

Naren Ramakrishnan

1.1 引 言

C4.5 算法[30]是机器学习和数据挖掘领域中的一整套用于处理分类问题的算法。该算法是有监督学习类型的，即：给定一个数据集，所有实例都由一组属性来描述，每个实例仅属于一个类别，在给定数据集上运行 C4.5 算法可以学习得到一个从属性值到类别的映射，进而可使用该映射去分类新的未知实例。如图 1.1 中所示的例子：Day 列是日期序号，几个属性列(Outlook，Temperature，Humidity，Windy)描述了天气条件，Play Gaof? 列对应于类别(是否适合打高尔夫)。图 1.1 中的表格显示的是“训练数据”——每行是一个实例，实例的属性包括：Outlook(光照，三值)、Temperature(温度，连续值)、Humidity(湿度，连续值)、Windy(是否有风，二值)；实例的类别 PlayGolf? 是布尔类型。C4.5 算法用这个训练数据可以学到一个映射，该映射以新实例(其类别未知)的属性值作为输入，输出是对这个实例所属类别的预测。

Day	Outlook	Temperature	Humidity	Windy	Play Golf?
1	Sunny	85	85	False	No
2	Sunny	80	90	True	No
3	Overcast	83	78	False	Yes
4	Rainy	70	96	False	Yes
5	Rainy	68	80	False	Yes
6	Rainy	65	70	True	No
7	Overcast	64	65	True	Yes
8	Sunny	72	95	False	No
9	Sunny	69	70	False	Yes
10	Rainy	75	80	False	Yes
11	Sunny	75	70	True	Yes
12	Overcast	72	90	True	Yes
13	Overcast	81	75	False	Yes
14	Rainy	71	80	True	No

图 1.1 C4.5 算法所用数据集的一个示例

J. Ross Quinlan 设计的 C4.5 算法源于名为 ID3 的一种决策树诱导算法[25]，而 ID3 是被称为“迭代分解器(iterative dichotomizers)”系列算法的第 3 代。决策树相当于将一系列问题组织成树，具体说，每个问题对应一个属性(比如 Outlook)，根据属性值来生成判断分支，一直到决策树的叶节点就产生了类别(此处，就是 PlayGolf?)的预测结果。决策树和你利用的车辆用户手册排查你的车到底出了什么问题没什么不同。C4.5 算法除了能诱导出决策树，还可以将决策树转换成某种具有良好可理解性的规则。特别是进一步看到，通过 C4.5 的后剪枝操作得到的分类器不能再精确地被转换回决策树。

C4.5 算法的发展历史非常生动地展现了不同的学术社区在分类问题研究上的殊途同

归的情形。一方面，ID3 算法和 Original Tree 算法是各自独立发展的，而 Original Tree 算法最初由 Friedman 发明[13]，后来在 Breiman、Olshen 和 Stone 等人的参与下发展成 CART 算法[4]。另一方面，可以看到(从 ID3 发展而来的)C4.5 算法也大量引用了的 CART 算法的文献[30]，其设计理念很大程度上也受到 CART 算法的影响，例如对特殊类型属性的处理方式(为了尽量避免书中不同章节之间内容的重叠，本书在第 10 章中专门对 CART 算法进行介绍，而本章则尽可能缩减关于 CART 算法的论述，仅在必要处点出二者的关联)。此外，在文献[25]和[36]中，Quinlan 还肯定了 CLS(概念学习系统[16])框架对于发展 ID3 算法和 C4.5 算法的重要作用。如今，经典的 C4.5 已经被 See5/C5.0 所取代，后者是 Rulequest Research 公司的商用产品。

数据挖掘社区遴选出的十大算法中有两个都是基于树的算法，这也反映出此类算法在数据挖掘中被使用的广泛程度。最早的决策树仅处理标称/类别的数据类型，而如今已扩展到支持数值、符号乃至混合型的数据类型。具体的应用领域也很广泛，例如临床决策、生产制造、文档分析、生物信息学、空间数据建模(地理信息系统)等。实际上，只要是目标问题的类间边界能用树型分解方式或规则判别方式来确定，就可以使用 C4.5 算法。

1.2　算法描述

我们谈到 C4.5 时通常并不是指一个单一的算法，而是泛指基本的 C4.5、C4.5-no-pruning 以及拥有多重特性的 C4.5-rules 等诸多变体的一整套算法。下面先介绍最基本的 C4.5 算法，然后再进一步讨论其特性。

算法 1.1 在从宏观上描述了 C4.5 的工作原理。所有的树诱导方法都大体上遵循一种统一的递归模式，即：首先，用根节点表示一个给定的数据集；然后，从根节点开始在每个节点上测试一个特定的属性，把节点数据集划分成更小的子集，并用子树表示；该过程就要一直进行，直到子集成为“纯”的，就是说直到子集中的所有实例都属于同一个类别，树才停止增长。

算法 1.1　C4.5(D)

```
Input: an attribute-valued dataset D
  1: Tree={}
  2: if D is "pure" OR other stopping criteria met then
  3:    terminate
  4: end if
  5: for all attribute a∈D do
  6:    Compute information-theoretic criteria if we split on a
  7: end for
  8: a_best = Best attribute according to above computed criteria
  9: Tree=Create a decision node that tests a_best in the root
 10: D_v = Induced sub-datasets from D based on a_best
 11: for all D_v do
 12:    Tree_v = C4.5(D_v)
 13:    Attach Tree_v to the corresponding branch of Tree
 14: end for
 15: return Tree
```

图 1.1 给出的是用于演示 C4.5 算法的 golf 数据集，该数据集已被打包在 C4.5 软件的安装程序中了。如前所述，该示例的目的是演示根据给某日的天气状况来预测该日是否适合打高尔夫球。要注意在这个示例中，有些特征值是连续型，而另一些是类别型。

图 1.2 显示的是基于图 1.1 中的“训练数据”用 C4.5 算法（采用默认选项）诱导出的树。下面介绍从数据中诱导出这样的树所面临的各种选择。

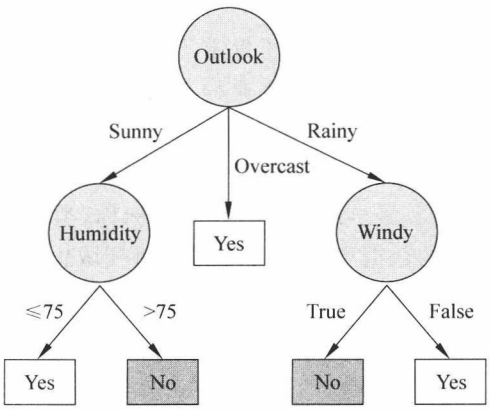


图 1.2 C4.5 算法基于图 1.1 的数据集获得的决策树

• 有哪些可能测试类型？

如图 1.2 所示，C4.5 算法并不限定只能进行二元测试，也允许有多于两项输出的测试。如果属性为布尔型，测试就会诱导出两个分支；如果属性是类别型，就需要进行多值测试；当然，也可以将这些值分组以生成更少的选项，此时每个选项相当于一个属性值，如此即可用这些选项来进行预测。如果属性是数值型的，测试就是{$\leqslant\theta$? 或$>\theta$?}形式的二元值，其中 θ 是属性的一个合适的判定阈值。

• 如何对测试进行选择？

C4.5 算法使用增益（gain）、增益率（gain ratio）等信息论准则来对测试进行选择。增益被定义为“执行一个测试所导致的类别分布的熵的减少量”，增益准则的一个缺陷在于她过于偏向选择具有更多输出结果的测试，而增益率具有克服这一偏差的优点，所以 C4.5 算法默认的测试选择准则是增益率。在树增长的每一步中，C4.5 算法都要选择具有最符合准则的那个测试。

• 如何选取测试的阈值？

如前所述，对于布尔型和类别型的属性，测试所用的值也就仅仅是该属性的可能取值。但对于数值型属性，相应测试就需要确定阈值，就是要对该属性的取值进行排序进而求出合适的切分点以最大化上述测试选择准则。Fayyad 和 Irani[10]的研究显示不是所有的前后相继的值都需要被考虑在内，例如：对于一个连续型属性的两个相继的值 V_i 和 V_{i+1}，如果这两个取值对应的实例属于同一类别，那么将它们分裂开来是不能提高信息增益（或增益率）的。

• 如何决定停止树生长？

通常，如果某节点的一个分支所辖的全部实例都是“纯”的，则这个分支就被确定为一个叶节点；另一种可能的终止条件是，如果该分支覆盖的实例总数已经低于预定的阈值。