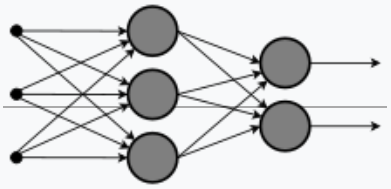
維基百科
自由的百科全書

决策树

维基百科，自由的百科全书

机器学习与数据挖掘



范式

监督学习 · 無監督學習 · 線上機器學習 · 元学习 · 半监督学习 · 自监督学习 · 强化学习 · 基于规则的机器学习 · 量子機器學習

问题

统计分类 · 生成模型 · 迴歸分析 · 聚类分析 · 降维 · 密度估计 · 异常检测 · 数据清洗 · 自动机器学习 · 关联规则学习 · 語意分析 · 结构预测 · 特征工程 · 表征学习 · 排序学习 · 语法归纳 · 本体学习 · 多模态学习

监督学习

(**分类** · **回归**)

学徒学习 · 决策树学习 · 集成学习 （Bagging · 提升方法 · 随机森林） · *k*-NN · 線性回歸 · 朴素贝叶斯 · 人工神经网络 · 邏輯斯諦迴歸 · 感知器 · 相关向量机（RVM） · 支持向量机（SVM） · 迁移学习 · 微调

聚类分析

BIRCH · CURE算法 · 层次 · *k*-平均 · Fuzzy · 期望最大化（EM） · DBSCAN · OPTICS · 均值飘移

降维

因素分析 · CCA · ICA · LDA · NMF · PCA · PGD · t-SNE · SDL

结构预测

圖模式 （貝氏網路 · 條件隨機域 · 隐马尔可夫模型）

异常检测

RANSAC · *k*-NN · 局部异常因子 · 孤立森林

人工神经网络

自编码器 · 認知計算 · 深度学习 · DeepDream · 多层感知器 · RNN （LSTM · GRU · ESN · 储备池计算） · 受限玻尔兹曼机 · GAN · SOM · CNN （U-Net） · Transformer （Vision transforme） · 脉冲神经网络 · Memtransistor · 电化学RAM（ECRAM）

强化学习

Q学习 · SARSA · 时序差分（TD） · 多智能体 （Self-play） · RLHF

与人类学习

主动学习 · 众包 · Human-in-the-loop

模型诊断

学习曲线

数学基础

内核机器 · 偏差–方差困境 · 计算学习理论 · 经验风险最小化 · 奥卡姆学习 · PAC学习 · 统计学习 · VC理论

大会与出版物

NeurIPS · ICML · ICLR · ML · JMLR

相关条目

人工智能术语 · 机器学习研究数据集列表 · 机器学习概要

决策论中（如风险管理），**决策树**（Decision tree）由一个决策图和可能的结果（包括资源成本和风险）组成，用来创建到达目标的规划。决策树建立并用来辅助决策，是一种特殊的树结构。决策树是一个利用像树一样的图形或决策模型的决策支持工具，包括随机事件结果，资源代价和实用性。它是一个算法显示的方法。决策树经常在运筹学中使用，特别是在决策分析中，它帮助确定一个能最可能达到目标的策略。如果在实际中，决策不得不在没有完备知识的情况下被在线采用，一个决策树应该平行概率模型作为最佳的选择模型或在线选择模型算法。决策树的另一个使用是作为计算条件概率的描述性手段。

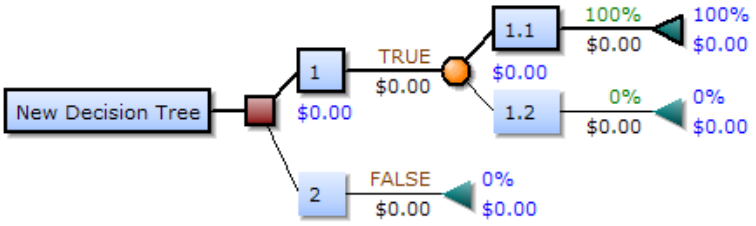
简介

机器学习中，**决策树**是一个预测模型；他代表的是对象属性与对象值之间的一种映射关系。树中每个节点表示某个对象，而每个分叉路径则代表某个可能的属性值，而每个叶节点则对应从根节点到该叶节点所经历的路径所表示的对象的值。决策树仅有单一输出，若欲有複数输出，可以建立独立的决策树以处理不同输出。 数据挖掘中决策树是一种经常要用到的技术，可以用于分析数据，同样也可以用来作预测。

从数据产生决策树的机器学习技术叫做**决策树学习**,通俗说就是**决策树**。

一个决策树包含三种类型的节点：

1. 决策节点：通常用矩形框来表示
2. 机会节点：通常用圆圈来表示
3. 终结点：通常用三角形来表示



决策树学习也是数据挖掘中一个普通的方法。在这裡，每个决策树都表述了一种树型结构，它由它的分支来对该类型的对象依靠属性进行分类。每个决策树可以依靠对源数据库的分割进行数据测试。这个过程可以递归式的对树进行修剪。 当不能再进行分割或一个单独的类可以被应用于某一分支时，递归过程就完成了。另外，随机森林分类器将许多决策树结合起来以提升分类的正确率。

决策树同时也可以依靠计算条件概率来构造。

决策树如果依靠数学的计算方法可以取得更加理想的效果。 数据库已如下所示：

(x, y) = (x1, x2, x3…, xk, y)

相关的变量Y表示我们尝试去理解，分类或者更一般化的结果。 其他的变量x1, x2, x3等则是帮助我们达到目的的变量。

类型

决策树有幾種產生方法：

- **分类树**分析是当预计结果可能为離散类型（例如三個種類的花，输贏等）使用的概念。
- **回归树**分析是当局域结果可能为实数（例如房价，患者住院时间等）使用的概念。
- **CART**分析是结合了上述二者的一个概念。CART是Classification And Regression Trees的缩写.

- **CHAID**（Chi-Square Automatic Interaction Detector）

建立方法

1. 以資料母群體為根節點。
2. 作單因子變異數分析等，找出變異量最大的變項作為分割準則。（決策樹每個葉節點即為一連串法則的分類結果。）
3. 若判斷結果的正確率或涵蓋率未滿足條件，則再依最大變異量條件長出分岔。

在教学中的使用

决策树，影响性图表，应用函数以及其他的决策分析工具和方法主要的授课对象是学校里商业、健康经济学和公共卫生专业的本科生，属于运筹学和管理科学的范畴。

举例阐述

下面我们用例子来说明：

小王是一家著名高尔夫俱乐部的经理。但是他被雇员数量问题搞得心情十分不好。某些天好像所有人都來玩高尔夫，以至于所有员工都忙的团团转还是应付不过来，而有些天不知道什么原因却一个人也不来，俱乐部为雇员数量浪费了不少资金。

小王的目的是通过下周天气预报寻找什么时候人们会打高尔夫，以适时调整雇员数量。因此首先他必须了解人们决定是否打球的原因。

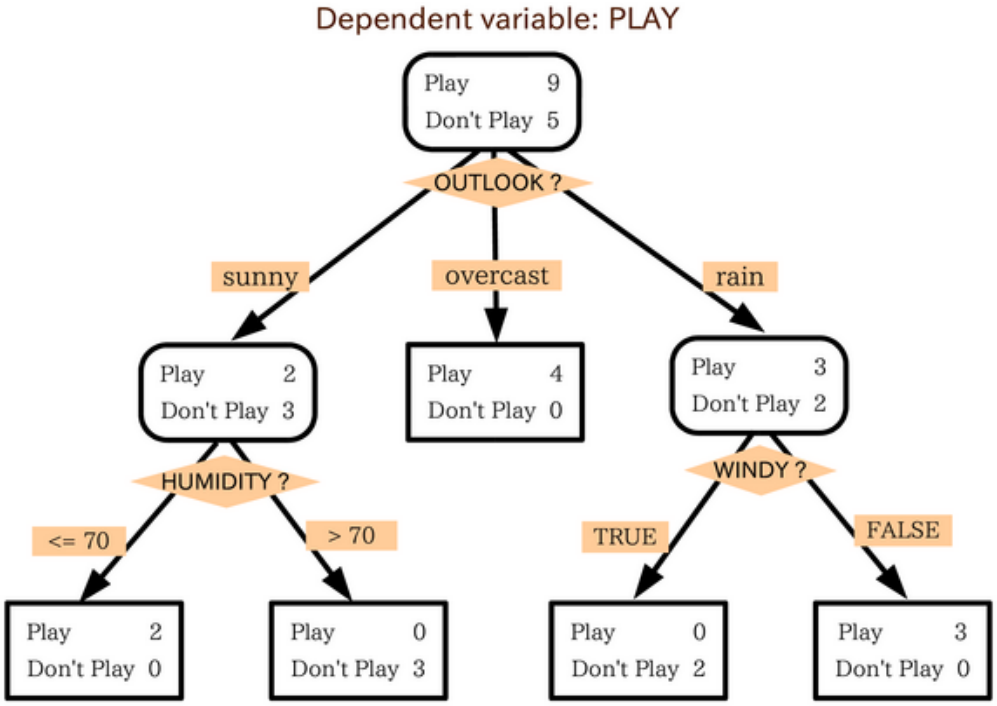
在2周时间内我们得到以下记录：

天气状况有晴，云和雨；气温用华氏温度表示；相对湿度用百分比；还有有无风。当然还有顾客是不是在这些日子光顾俱乐部。最终他得到了14行5列的数据表格。

Play golf dataset

Independent variables				Dep. var
OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
sunny	85	85	FALSE	Don't Play
sunny	80	90	TRUE	Don't Play
overcast	83	78	FALSE	Play
rain	70	96	FALSE	Play
rain	68	80	FALSE	Play
rain	65	70	TRUE	Don't Play
overcast	64	65	TRUE	Play
sunny	72	95	FALSE	Don't Play
sunny	69	70	FALSE	Play
rain	75	80	FALSE	Play
sunny	75	70	TRUE	Play
overcast	72	90	TRUE	Play
overcast	81	75	FALSE	Play
rain	71	80	TRUE	Don't Play

决策树模型就被建起来用于解决问题。



决策树是一个有向无环图。根结点代表所有数据。分类树算法可以通过变量outlook，找出最好地解释非独立变量play（打高尔夫的人）的方法。变量outlook的范畴被划分为以下三个组：

晴天，多云天和雨天。

我们得出第一个结论：如果天气是多云，人们总是选择玩高尔夫，而只有少数很着迷的甚至在雨天也会玩。

接下来我们把晴天组的分为两部分，我们发现顾客不喜欢湿度高于70%的天气。最终我们还发现，如果雨天还有风的话，就不会有人打了。

这就通过分类树给出了一个解决方案。小王（老板）在晴天，潮湿的天气或者刮风的雨天解雇了大部分员工，因为这种天气不会有人打高尔夫。而其他的天气会有很多人打高尔夫，因此可以雇用一些临时员工来工作。

公式

C4.5和C5.0生成树算法使用熵。这一度量是基于資訊科學理论中熵的概念。

$$I_E(i) = -\sum_{j=1}^{m} f(i,j) \log_2 f(i,j)$$

决策树的优点

相对于其他数据挖掘算法，决策树在以下几个方面拥有优势：

- **决策树易于理解和实现.**人们在通过解释后都有能力去理解决策树所表达的意义。
- **对于决策树，数据的准备往往是简单或者是不必要的.**其他的技术往往要求先把数据一般化，比如去掉多余的或者空白的属性。
- **能够同时处理数据型和常规型属性。**其他的技术往往要求数据属性的单一。
- **是一个白盒模型**如果给定一个观察的模型，那么根据所产生的决策树很容易推出相应的逻辑表达式。
- **易于通过静态测试来对模型进行评测。**表示有可能测量该模型的可信度。
- **在相对短的时间内能够对大型数据源做出可行且效果良好的结果。**

决策树的缺点

决策树：

- 对于那些各类别样本数量不一致的数据，在决策树当中信息增益的结果偏向于那些具有更多数值的特征。

决策树的剪枝

剪枝是决策树停止分支的方法之一，剪枝有分预先剪枝和后剪枝两种。预先剪枝是在树的生长过程中设定一个指标，当达到该指标时就停止生长，这样做容易产生“视界局限”，就是一旦停止分支，使得节点N成为叶节点，就断绝了其后继节点进行“好”的分支操作的任何可能性。不严格的说这会已停止的分支会误导学习算法，导致产生的树不纯度降差最大的地方过分靠近根节点。后剪枝中树首先要充分生长，直到叶节点都有最小的不纯度值为止，因而可以克服“视界局限”。然后对所有相邻的成对叶节点考虑是否消去它们，如果消去能引起令人满意的不纯度增长，那么执行消去，并令它们的公共父节点成为新的叶节点。这种“合并”叶节点的做法和节点分支的过程恰好相反，经过剪枝后叶节点常常会分布在很宽的层次上，树也变得非平衡。后剪枝技术的优点是克服了“视界局限”效应，而且无需保留部分样本用于交叉验证，所以可以充分利用全部训练集的信息。但后剪枝的计算量代价比预剪枝方法大得多，特别是在大样本集中，不过对于小样本的情况，后剪枝方法还是优于预剪枝方法的。

由决策树扩展为决策圖

在决策树中所有从根到葉節點的路径都是透過“與”（AND）运算连接。在决策图中可以使用“或”来连接多于一个的路徑。

决策树的实现

Bash

决策树的代码实现可参考：决策树算法实现——Bash (http://liuzhiqiangruc.iteye.com/blog/1601922)（页面存档备份 (https://web.archive.org/web/20151125045121/http://liuzhiqiangruc.iteye.com/blog/1601922)，存于互联网档案馆）

相关条目

- 贝叶斯定理
- 贝叶斯概率
- 專家系統
- 真值表
- 运筹学
- 形态学分析
- 决策表
- 马尔科夫链

参考文献

- [1] T. Menzies, Y. Hu, Data Mining For Very Busy People. IEEE Computer, October 2003, pgs. 18-25.
- [2]决策树分析 (http://www.mindtools.com/pages/article/newTED_04.htm)（页面存档备份 (https://web.archive.org/web/20060208152747/http://www.mindtools.com/pages/article/newTED_04.htm)，存于互联网档案馆）mindtools.com
- [3]J.W. Comley and D.L. Dowe (http://www.csse.monash.edu.au/~dld)（页面存档备份 (https://web.archive.org/web/20060212185445/http://www.csse.monash.edu.au/~dld)，存于互联网档案馆）, "Minimum Message Length, MDL and Generalised Bayesian Networks with Asymmetric Languages", 第十一章 (https://web.archive.org/web/20060909003748/http://mitpress.mit.edu/catalog/item/default.asp?ttype=2&tid=10478&mode=to

c) (pp265-294) in P. Grunwald, M.A. Pitt and I.J. Myung (eds)., Advances in Minimum Description Length: Theory and Applications (https://web.archive.org/web/20060619060230/http://mitpress.mit.edu/catalog/item/default.asp?sid=4C100C6F-2255-40FF-A2ED-02FC49FEBE7C&ttype=2&tid=10478), M.I.T. Press, April 2005, ISBN 0262072629. （本论文把决策树作为贝叶斯网络使用Minimum Message Length (http://www.csse.monash.edu.au/~dld/MML.html)（页面存档备份 (https://web.archive.org/web/20060209113220/http://www.csse.monash.edu.au/~dld/MML.html)，存于互联网档案馆） (MML的内部结点).早期版本：Comley and Dowe (2003) (http://www.csse.monash.edu.au/~dld/Publications/2003/Comley+Dowe03_HICS2003.ref)（页面存档备份 (https://web.archive.org/web/20060209003640/http://www.csse.monash.edu.au/~dld/Publications/2003/Comley+Dowe03_HICS2003.ref)，存于互联网档案馆）, .pdf (http://www.csse.monash.edu.au/~dld/Publications/2003/Comley+Dowe03_HICS2003_GeneralBayesianNetworksAsymmetricLanguages.pdf)（页面存档备份 (https://web.archive.org/web/20060210161715/http://www.csse.monash.edu.au/~dld/Publications/2003/Comley+Dowe03_HICS2003_GeneralBayesianNetworksAsymmetricLanguages.pdf)，存于互联网档案馆）.）

- [4]P.J. Tan and D.L. Dowe (http://www.csse.monash.edu.au/~dld)（页面存档备份 (https://web.archive.org/web/20060212185445/http://www.csse.monash.edu.au/~dld)，存于互联网档案馆） (2004), MML Inference of Oblique Decision Trees (http://www.csse.monash.edu.au/~dld/Publications/2004/Tan+DoweAI2004.pdf)（页面存档备份 (https://web.archive.org/web/20160806013805/http://www.csse.monash.edu.au/~dld/Publications/2004/Tan+DoweAI2004.pdf)，存于互联网档案馆）,人工智能讲义3339, Springer-Verlag, pp1082-1088 (http://www.csse.monash.edu.au/~dld/Publications/2004/Tan+Dowe2004.ref)（页面存档备份 (https://web.archive.org/web/20051226173202/http://www.csse.monash.edu.au/~dld/Publications/2004/Tan+Dowe2004.ref)，存于互联网档案馆）.（此论文使用Minimum Message Length.）
- [5] Eruditionhome (http://www.eruditionhome.com/datamining)（页面存档备份 (https://web.archive.org/web/20060212195322/http://www.eruditionhome.com/datamining)，存于互联网档案馆）数据挖掘资源大词典
- [6]决策树基础 (http://www.vanguardsw.com/DpHelp4/dph00075.htm) （页面存档备份 (https://web.archive.org/web/20180201095438/http://www.vanguardsw.com/DpHelp4/dph00075.htm)，存于互联网档案馆） vanguardsw.com
- [7]General Morphological Analysis: A General Method for Non-Quantified Modelling (http://www.swemorph.com/pdf/gma.pdf) （页面存档备份 (https://web.archive.org/web/20210222003847/http://www.swemorph.com/pdf/gma.pdf)，存于互联网档案馆） From the Swedish Morphological Society (http://www.swemorph.com)（页面存档备份 (https://web.archive.org/web/20111129005904/http://www.swemorph.com/)，存于互联网档案馆）
- [8]decisiontrees.net Interactive Tutorial (https://web.archive.org/web/20190831011603/http://www.decisiontrees.net/)
- [9] [Morgan.Kaufmann.Data.Mining.Practical.Machine.Learning.Tools.and.Techniques.Second.Edition]Jun.2005, 非常好的一本介绍决策树的书，是一本可以从基础学起的好书，另外介绍的weka决策树软件也不错。是JAVA编的。
