



Decision tree

one of the simplest and yet most useful Machine Learning structures

Epidemiology Student Association

Dianqin Sun

Peking Union Medical College

2019-04-20

内容大纲



Part 1. 什么是决策树

Part 2. 如何生成决策树

Part 3. 决策树在流行病学中的应用

Part 4. 随机森林简介

什么是决策树：举个栗子



问题：

建立预测模型，预测65岁以上的老年人两年内是否会跌倒

背景：

一个随访了两年的以社区为基础的65岁以上老年人群队列(n=1662)

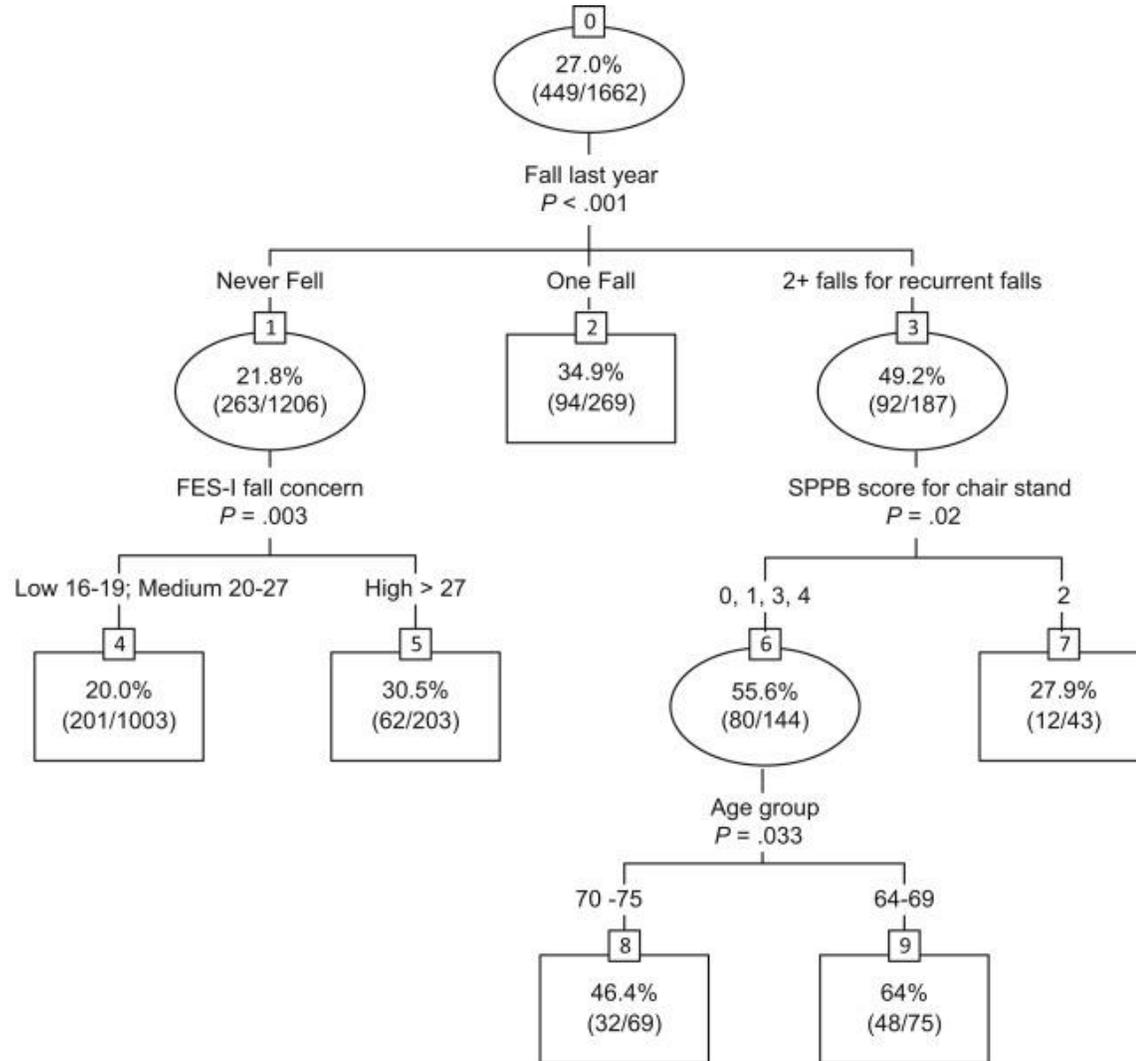
变量：

年龄、性别、体成分、认知功能、是否抑郁、跌倒史、所患慢性病数目、国际版跌倒效能感量表(FES-I)，简易躯体能力测试(SPPB，包括5次坐位站起，串联站立平衡测试，4m行走测试)。

统计分析方法：？

Logistic回归模型

什么是决策树：举个栗子



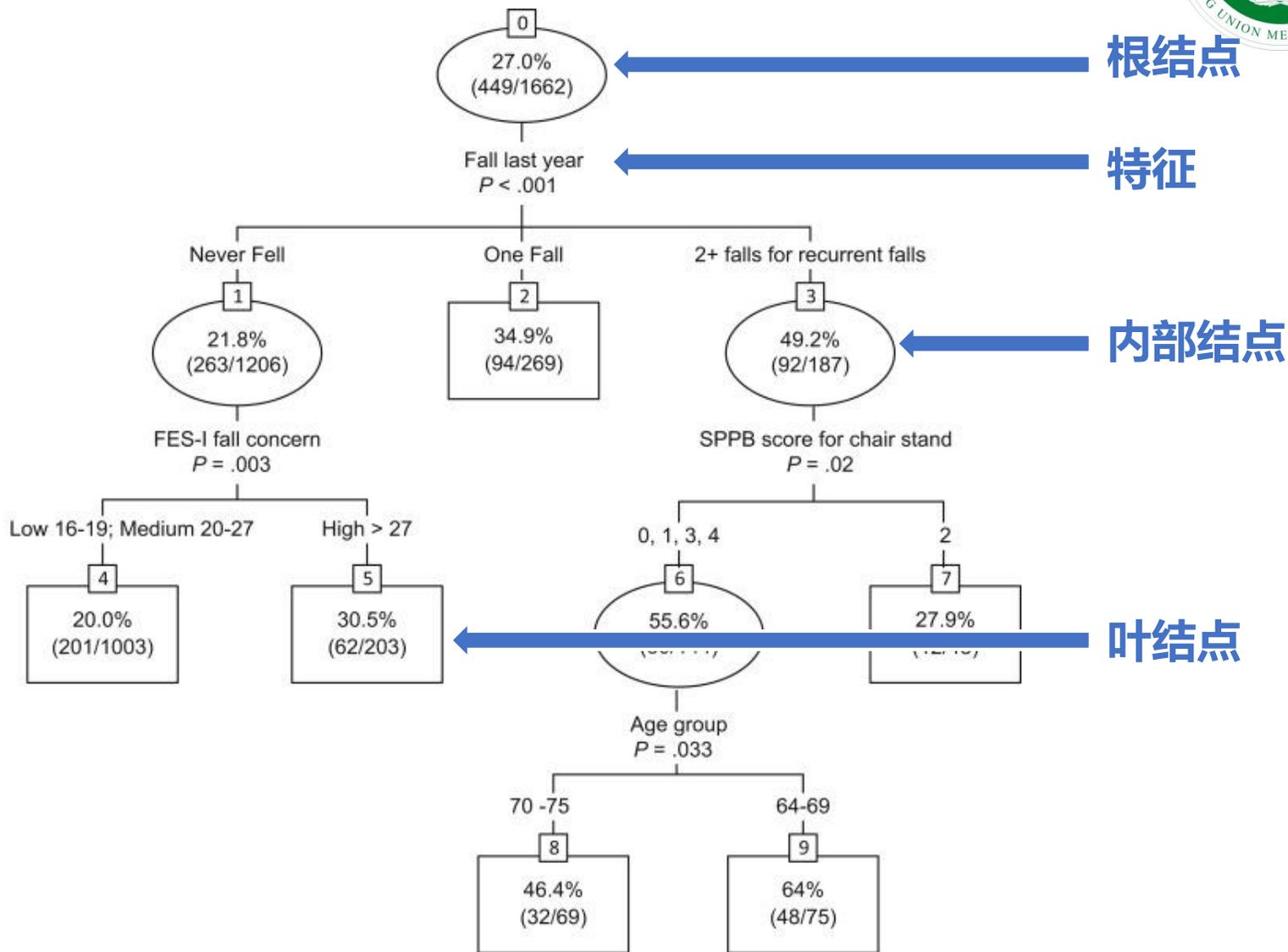
Classification tree using the CHAID method for predicting the risk of falling in community-dwelling older persons at 2-year follow-up. Each divisible group is identified as an ellipse (nodes 0, 1, 3, and 6) and end nodes as a rectangle (2, 4, 5, 7, 8, and 9). Numbers inside nodes represent incidence rates (in percentage), number of fallers, and total number of participants who are indicated by the node.

The classification accuracies were 74.2% for CHAID (10-fold cross-validation) and 63.2% for multiple logistic regression models.



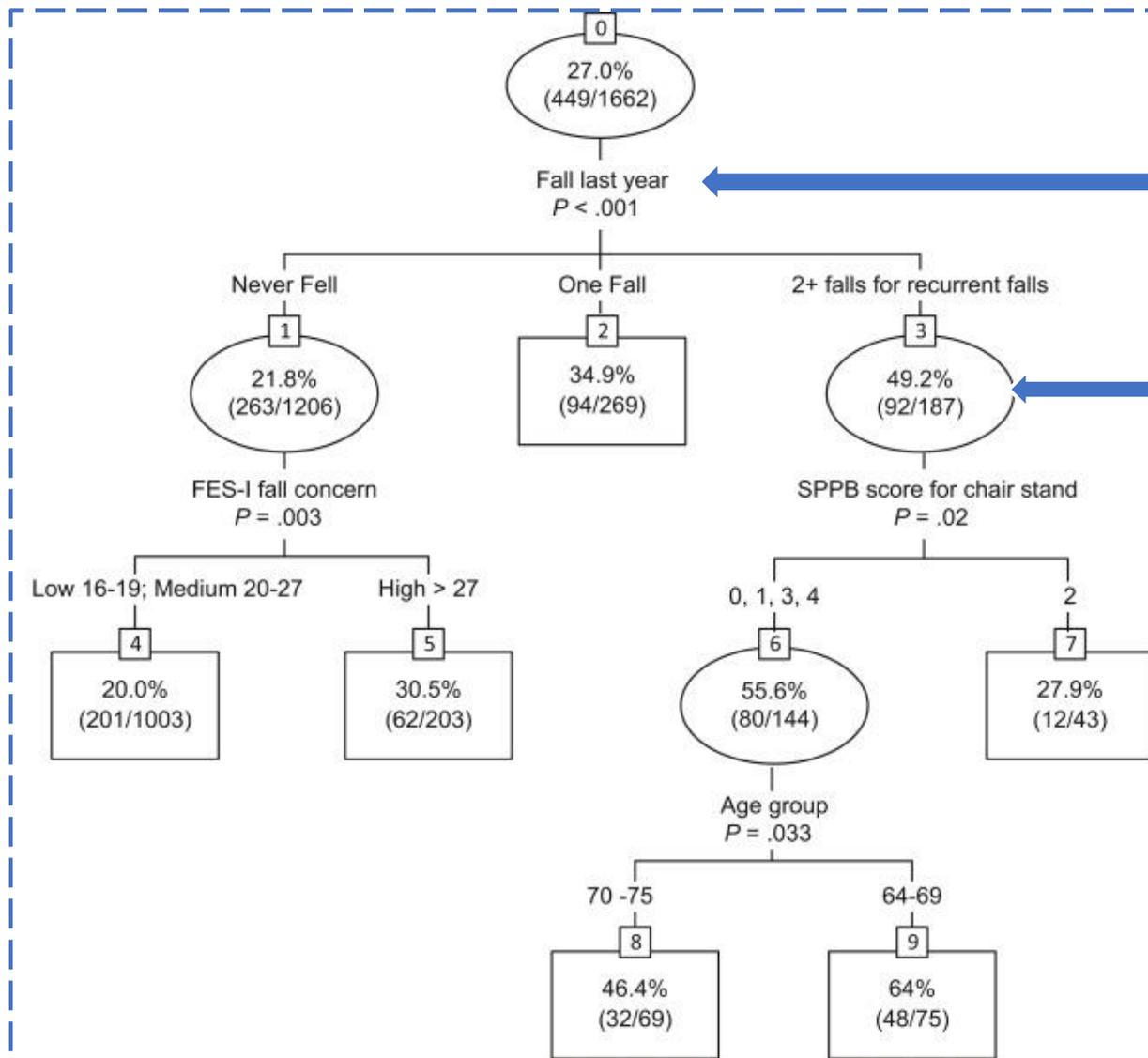
什么是决策树?

Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves).





如何生成决策树-算法构成



选取哪个特征?

特征选择

继续分枝还是停止?

决策树的生成

过拟合?

决策树的剪枝

如何生成决策树-特征选择

信息增益, 信息增益比, 基尼系数, 残差平方和

信息熵 (entropy)

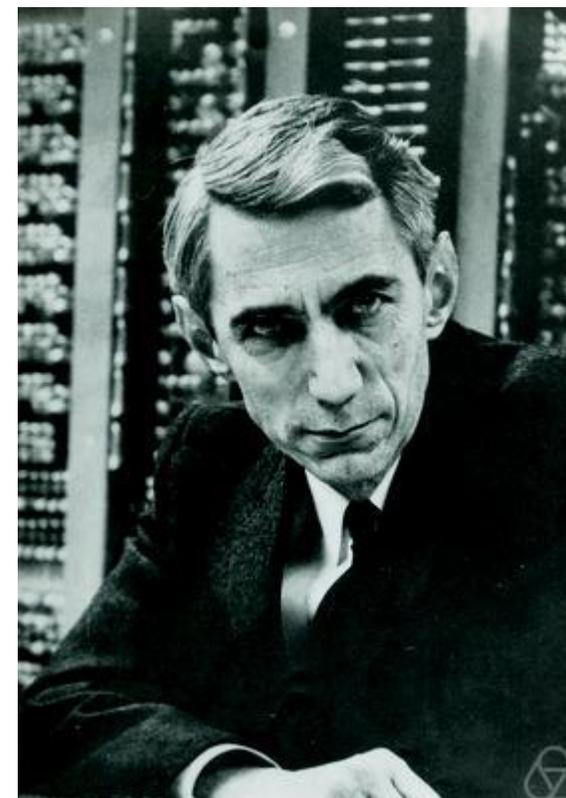
- Shannon的信息论
- 信息是什么?
- 信息是可以量化的东西么?

听君一席话, 胜读十年书。

- 信息量为什么有大有小?

太阳每天从东方升起

高盛预测茅台股价可以冲破1000



如何生成决策树-特征选择

信息增益, 信息增益比, 基尼系数, 残差平方和

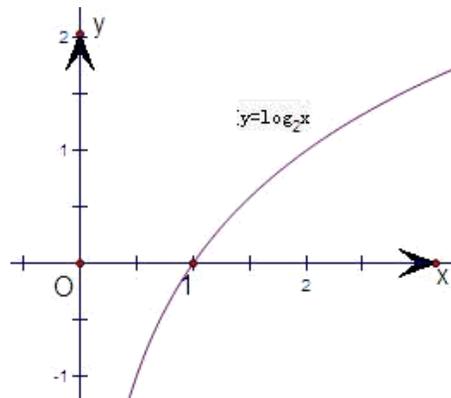
信息熵 (entropy)

- 如何度量信息?

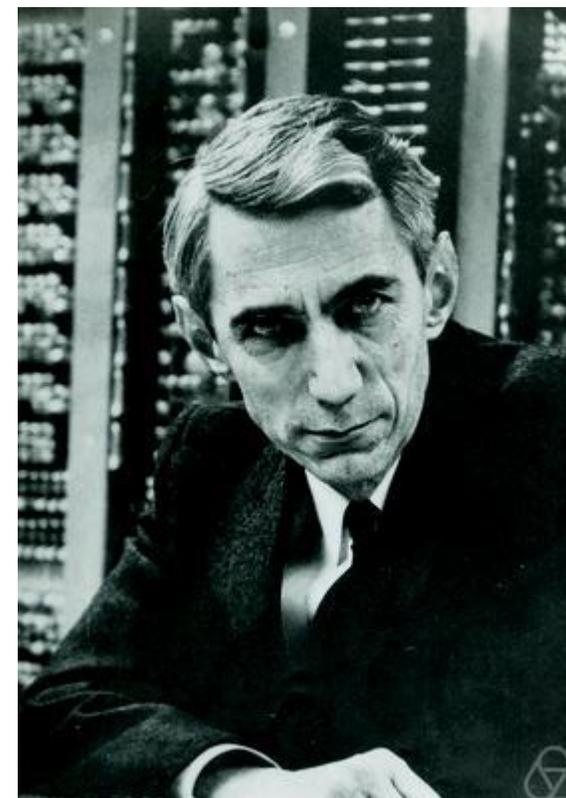
一条信息的信息量大小和它的不确定性有直接的关系。
不确定性的大小和事情可能结果的数量和概率有关。

- 数学公式?

- 非负数
- 可以相加
- 连续依赖于概率
- 与结果数量有关



- 唯一满足的函数: 底数大于1的 $-\log(x)$
- 一个事件的信息量就是这个事件发生的概率的负对数





如何生成决策树-特征选择

信息增益, 信息增益比, 基尼系数, 残差平方和

信息熵 (entropy)

信息熵是跟所有可能性有关系的。每个可能事件的发生都有个概率。信息熵就是平均而言发生一个事件我们得到的信息量大小。所以数学上, 信息熵其实是信息量的期望。

“信息熵” (information entropy) 是度量样本集合纯度最常用的一种指标。假定当前样本集合 D 中第 k 类样本所占的比例为 p_k ($k = 1, 2, \dots, |\mathcal{Y}|$), 则 D 的信息熵定义为

$$\text{Ent}(D) = - \sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k . \quad (4.1)$$

$\text{Ent}(D)$ 的值越小, 则 D 的纯度越高。



如何生成决策树-特征选择

信息增益, 信息增益比, 基尼系数, 残差平方和

信息增益 (information gain)

假定离散属性 a 有 V 个可能的取值 $\{a^1, a^2, \dots, a^V\}$, 若使用 a 来对样本集 D 进行划分, 则会产生 V 个分支结点, 其中第 v 个分支结点包含了 D 中所有在属性 a 上取值为 a^v 的样本, 记为 D^v . 我们可根据式(4.1) 计算出 D^v 的信息熵, 再考虑到不同的分支结点所包含的样本数不同, 给分支结点赋予权重 $|D^v|/|D|$, 即样本数越多的分支结点的影响越大, 于是可计算出用属性 a 对样本集 D 进行划分所获得的“信息增益” (information gain)

$$\text{Gain}(D, a) = \text{Ent}(D) - \sum_{v=1}^V \frac{|D^v|}{|D|} \text{Ent}(D^v) . \quad (4.2)$$



如何生成决策树-特征选择

信息增益, 信息增益比, 基尼系数, 残差平方和

信息增益 (information gain)

a	b	c	class
0	0	0	X
0	0	0	X
0	0	0	X
0	0	1	X
1	0	1	Y
1	0	1	Y
1	0	1	Y
0	1	1	Y
0	1	1	Y
0	1	1	Y

a, b, c为三个特征变量, class为类别, 分X, Y两种。在第一次对数据集进行划分时, 根据

$$\text{Gain}(D, a) = \text{Ent}(D) - \sum_{v=1}^V \frac{|D^v|}{|D|} \text{Ent}(D^v).$$

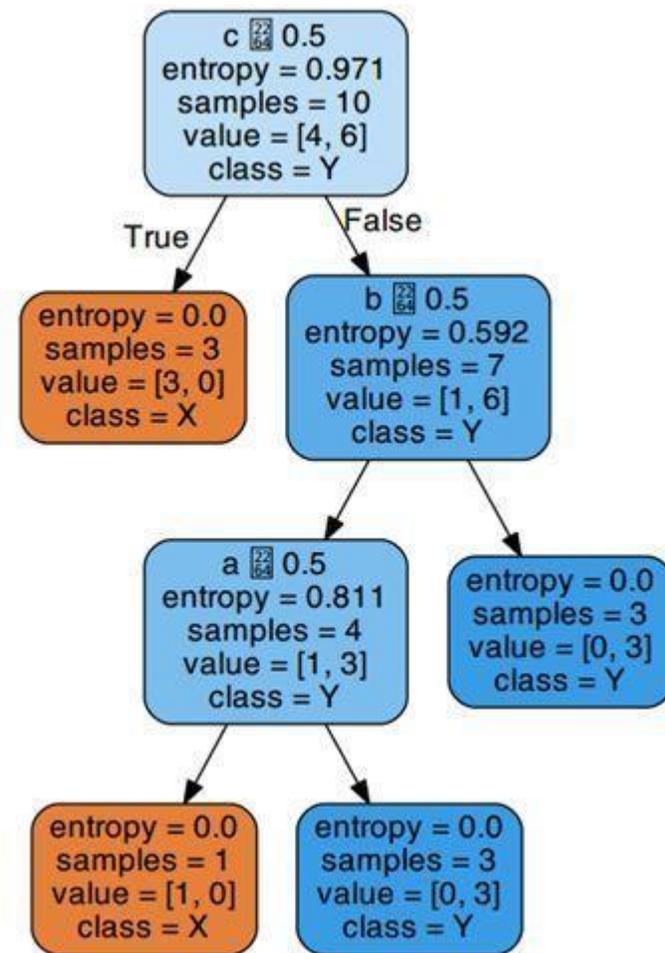
计算属性的信息增益, 得

$$\text{Ent}(D) = -\left(\frac{4}{10} \log_2 \frac{4}{10} + \frac{6}{10} \log_2 \frac{6}{10}\right) = 0.971$$

$$\text{Gain}(D, c) = \text{Ent}(D) - \text{Ent}(D, c) = 0.971 - \left[\frac{3}{10} * 0 - \right]$$

$$\frac{7}{10} \left(\frac{1}{7} \log_2 \frac{1}{7} + \frac{6}{7} \log_2 \frac{6}{7} \right) = 0.557,$$

$$\text{Gain}(D, a) = \text{Gain}(D, b) = 0.2816$$





如何生成决策树-特征选择

信息增益, **信息增益比**, 基尼系数, 残差平方和

信息增益的局限性

信息增益准则对可取值数目较多的属性有所偏好。因为信息增益反映的是给定一个条件以后不确定性减少的程度, 必然是分得越细的数据集确定性更高, 信息增益越大。

信息增益比

$$\text{Gain_ratio}(D, a) = \frac{\text{Gain}(D, a)}{\text{IV}(a)}, \quad (4.3)$$

其中

$$\text{IV}(a) = - \sum_{v=1}^V \frac{|D^v|}{|D|} \log_2 \frac{|D^v|}{|D|} \quad (4.4)$$

特征a的取值数目越多, 即v越大, IV(a)越大



如何生成决策树-特征选择

信息增益, 信息增益比, 基尼系数, 残差平方和

基尼系数

$$\begin{aligned} \text{Gini}(D) &= \sum_{k=1}^{|\mathcal{Y}|} \sum_{k' \neq k} p_k p_{k'} \\ &= 1 - \sum_{k=1}^{|\mathcal{Y}|} p_k^2 . \end{aligned}$$

Gini(D)反映了从数据集D中随机抽取两个样本, 其类别标记不一致的概率。因此, Gini(D)越小, 数据集D的纯度越高。

特征a的基尼系数

$$\text{Gini_index}(D, a) = \sum_{v=1}^V \frac{|D^v|}{|D|} \text{Gini}(D^v) .$$

在候选特征集合中, 选择那个使得划分后基尼系数最小的特征作为最优化分特征。



如何生成决策树-特征选择

信息增益, 信息增益比, 基尼系数, 残差平方和

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right]$$

y_i : 样本目标变量的真实值。
 R_1 & R_2 : 被划分的两个子集
 c_1 & c_2 : R_1 & R_2 子集的样本均值。
 j : 当前的样本特征
 s : 划分点

上面公式的含义是：计算所有的特征以及相应所有切分点下的残差平方和，找到一组(特征 j ，切分点 s)，以满足：分别最小化左子树和右子树的残差平方和，并在此基础上再次最小化二者之和。



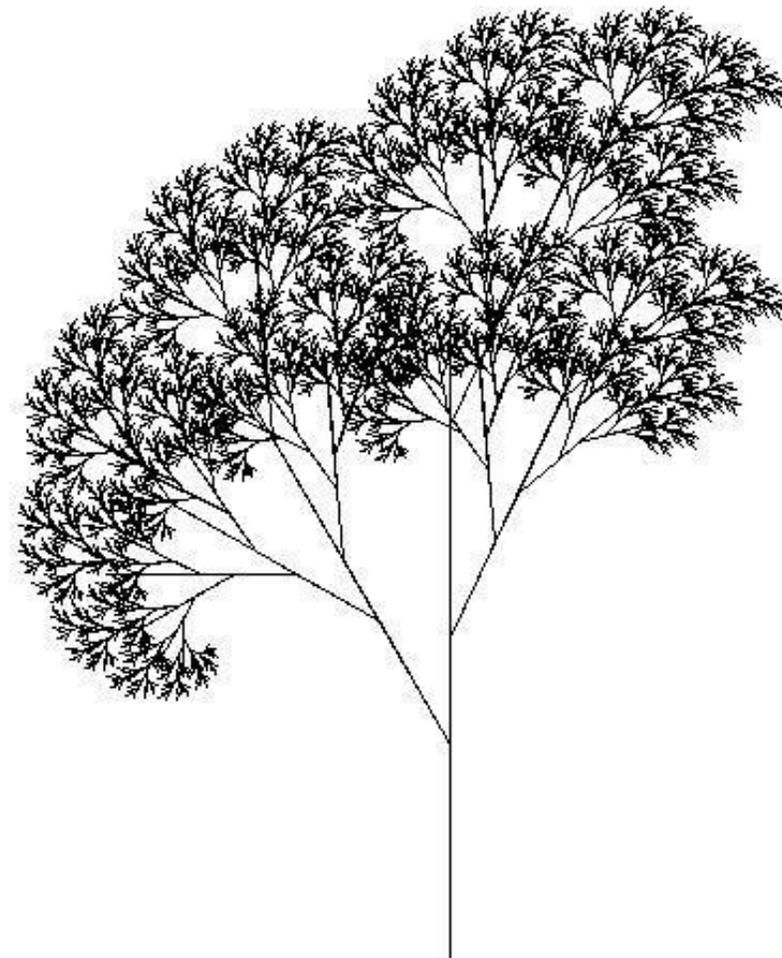
如何生成决策树-生成

递归

选择特征不断地生成子树，直到达到
停止条件

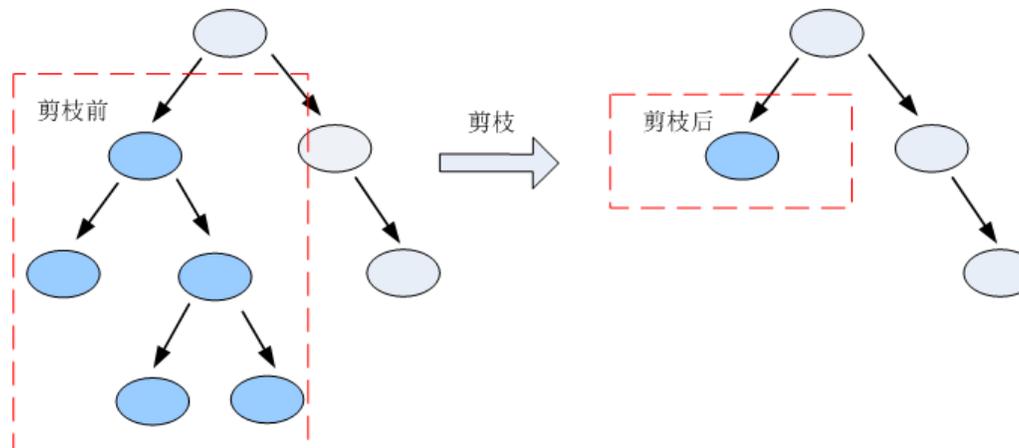
递归的停止条件

- 没有更多特征供选择
- 子集已经是纯数据集



如何生成决策树-剪枝

- 防止“过拟合”
- 分为“预剪枝”和“后剪枝”
- “预剪枝”指在完全正确分类之前，决策树会较早地停止树的生长
 - 如果树到达一定高度
 - 如果节点下包含的样本点小于指定的阈值
 - 如果样本的类分布是独立于可用特征的（使用卡方检验）
 - 如果扩展当前节点不会改善信息增益，即信息增益小于指定的阈值
- “后剪枝”即首先通过完全分裂构造完整的决策树，允许过拟合，然后采取一定的策略来进行剪枝。
 - 错误率降低剪枝REP、悲观剪枝法 PEP、代价-复杂度剪枝 CCP、最小错误剪枝 MEP
 - 相比预剪枝，后剪枝的优点是：通常会保留了更多的分支，欠拟合风险更小。后剪枝的缺点是：决策树训练时间花费比未剪枝决策树和预剪枝决策树都要大的多。





如何生成决策树-剪枝

比较项目和枝剪方法	CCP	REP	PEP	MEP
独立剪枝集	CV 方式：不需要	需要	不需要	不需要
剪枝方式	自顶向上	自底向上	自顶向下	自底向上
误差估计	使用 CV 或标准误差	利用剪枝集	使用连续性校正	基于 m^2 概率估计
计算复杂性	$O(n^2)$	$O(n)$	$O(n)$	$O(n)$

- 错误率降低剪枝 REP——《机器学习》，周志华，决策树章节
- 代价-复杂度剪枝 CCP——《统计学习方法》，李航，决策树章节
- 最小错误剪枝 MEP——<https://zhuanlan.zhihu.com/p/30296061>
- 悲观剪枝法 PEP



如何生成决策树-剪枝

悲观剪枝法 PEP

- 根据剪枝前后的错误率来判定子树的修剪。
- 假定错误率为二项式分布，并计算它的标准差。对误差估计增加了连续性校正即加 0.5。

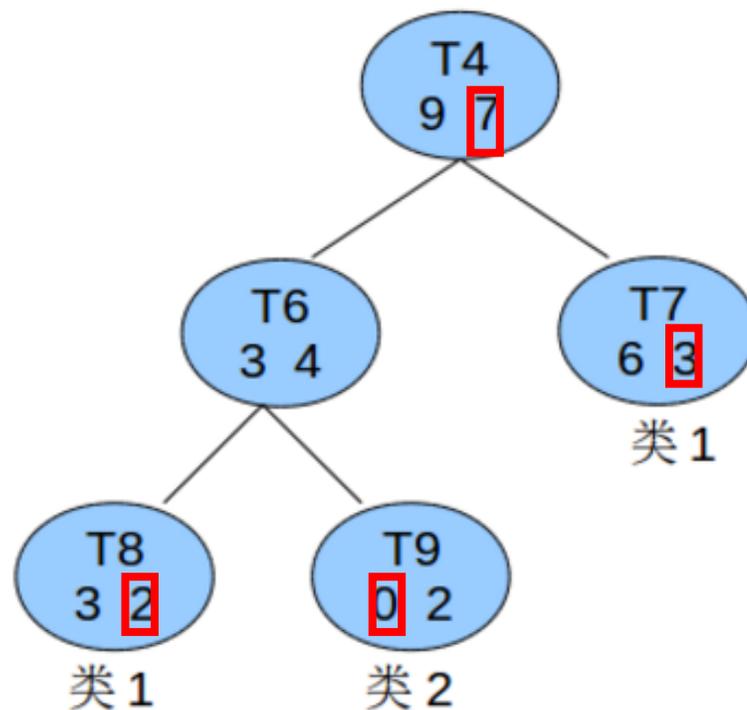
$$e'(t) = \left[e(t) + \frac{1}{2} \right];$$

$$e'(T_i) = \sum e(i) + \frac{N_i}{2};$$

t节点下叶子节点的错误率

$$\text{var}(\text{subtree_err_count}) = \sqrt{N * e * (1 - e)}$$

$$e'(t) \leq e'(T_i) + S_e(e'(T_i))$$



比如T4这棵子树的误差率：

$$\frac{(3 + 2) + 0.5 * 3}{16} = \frac{6.5}{16} = 0.40625$$

子树误差率的标准误差：

$$\sqrt{16 * 0.40625 * (1 - 0.40625)} = 1.96$$

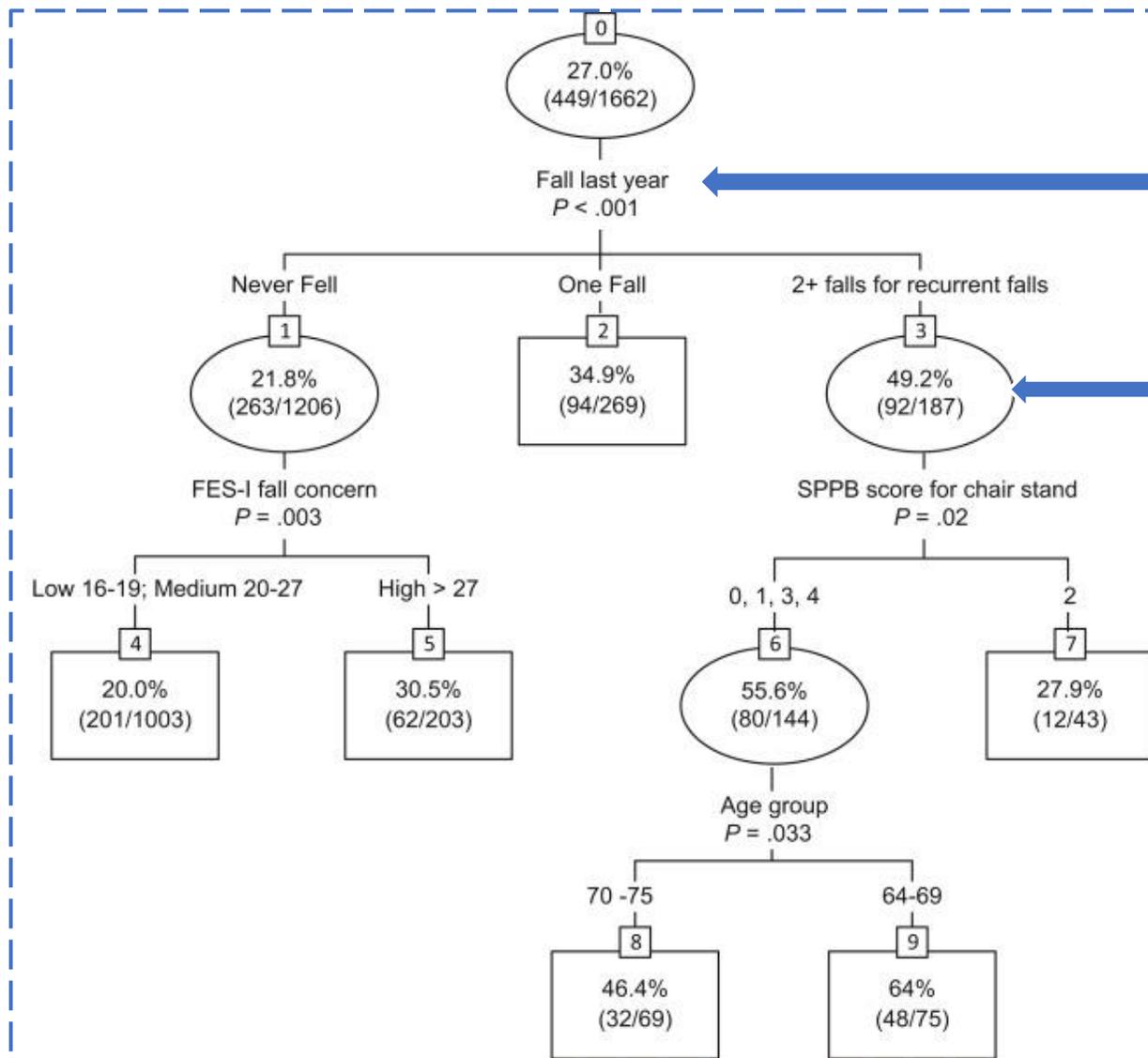
子树替换为一个叶节点后，其误差率为：

$$\frac{7 + 0.5}{16} = 0.46875$$

因为 $6.5 + 1.96 > 7.5$ ，所以决定将子树T4替换这一个叶子节点。



如何生成决策树-算法构成



选取哪个特征?

特征选择

继续分枝还是停止?

决策树的生成

过拟合?

决策树的剪枝



如何生成决策树-算法

↪ **ID3 (Iterative Dichotomiser 3)**
↪ **C4.5**

↪ **CART (Classification and Regression Trees)**

↪ AID (Automatic Interaction Detection)

↪ **CHAID (Chi-Squared Automatic Interaction Detection)**

↪ FACT (Fast Algorithm for Classification Trees)

↪ QUEST (Quick, Unbiased, Efficient, Statistical Tree)

↪ CRUISE (Classification Rule with Unbiased Interaction Selection and Estimation)

↪ GUIDE (Generalized, Unbiased, Interaction Detection and Estimation)



如何生成决策树-ID3算法

- 根据**信息增益**来选取特征，容易偏向于取值较多的特征。

$$\text{Gain}(D, a) = \text{Ent}(D) - \sum_{v=1}^V \frac{|D^v|}{|D|} \text{Ent}(D^v) .$$

- 不能处理为连续性变量的特征，不能处理缺失值。
- ID3算法**只有树的生成**，所以该算法生成的树容易产生过拟合。



如何生成决策树-C4.5算法

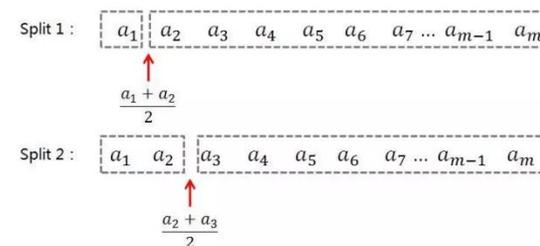
- 根据**信息增益比**来选取特征作为决策树分裂的节点。

$$\text{Gain_ratio}(D, a) = \frac{\text{Gain}(D, a)}{\text{IV}(a)}, \quad (4.3)$$

其中

$$\text{IV}(a) = - \sum_{v=1}^V \frac{|D^v|}{|D|} \log_2 \frac{|D^v|}{|D|} \quad (4.4)$$

- 可以处理连续的特征。
- 将连续特征离散化进行处理。将特征a的连续值从小打大进行排列，生成n-1个切分选择，遍历每个切分，根据信息增益比选择最优的切分。
- 可以处理缺失值。
 - 若样本x在特征a取值缺失，则将x划入所有子节点，但要调整样本的权重值
 - 在选取最优特征时，根据无缺失样本集计算信息增益比，但同时要给与计算出的值一个权重
- 《机器学习》，周志华，决策树章节
- 引入**悲观剪枝法 PEP**进行剪枝





如何生成决策树-CART算法

分类树算法

- 根据**基尼系数**来选取特征作为决策树分裂的节点。
- 可以处理连续的特征，可以处理缺失值。
- 与C4.5算法一样，将连续特征离散化进行处理。
- **CART永远只有两个分支**
- 每一次将其中一个特征分为一类，其它非该特征分为另外一类。依照这个标准遍历所有的分类情况，计算每种分类下的基尼指数，最后选择值最小的一个作为最终的特征划分。

Split 1 : a_1 a_2 a_3 a_4 a_5 a_6 $a_7 \dots a_{m-1}$ a_m

Split 2 : a_1 a_2 a_3 a_4 a_5 a_6 $a_7 \dots a_{m-1}$ a_m

...

...

Split m-1 : a_1 a_2 a_3 a_4 a_5 a_6 $a_7 \dots a_{m-1}$ a_m

- 引入**代价-复杂度剪枝 CCP**进行剪枝
- 《统计学习方法》，李航，决策树章节

$$\begin{aligned} \text{Gini}(D) &= \sum_{k=1}^{|\mathcal{Y}|} \sum_{k' \neq k} p_k p_{k'} \\ &= 1 - \sum_{k=1}^{|\mathcal{Y}|} p_k^2. \end{aligned}$$

$$\text{Gini_index}(D, a) = \sum_{v=1}^V \frac{|D^v|}{|D|} \text{Gini}(D^v).$$



如何生成决策树-CART算法

回归树算法

- 回归树的预测方法
- 回归树的预测方式是：将叶子节点中样本的y均值作为回归的预测值。而分类树的预测方式则是：叶子节点中概率最大的类别作为当前节点的预测类别。
- 根据**残差平方和**来选取特征作为决策树分裂的节点。

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right]$$

yi: 样本目标变量的真实值。
R1&R2: 被划分的两个子集
c1&c2: R1&R2子集的样本均值。
j: 当前的样本特征
s: 划分点

- 计算所有的特征以及相应所有切分点下的残差平方和，找到一组(特征j，切分点s)，以满足：分别最小化左子树和右子树的残差平方和，并在此基础上再次最小化二者之和。
- 可以处理连续的特征，可以处理缺失值。
- 与分类树算法一样，将连续特征离散化进行处理。
- 引入**代价-复杂度剪枝 CCP**进行剪枝
- 《统计学习方法》，李航，决策树章节



如何生成决策树-CHAID算法

- 根据**卡方检验的P值**来选取特征作为决策树分裂的节点。
- P值越小，说明输入变量与输出变量的关系越紧密，应当作为当前最佳分组变量。当P值相同时，应该选择检验统计量观测值最大的输入变量，也就是卡方最大的输入变量。
- 特征可以是分类型，也可以是连续型。
- 与C4.5算法一样，将连续特征离散化进行处理。
- 目标变量可以是分类型，也可以是连续型。
- 如果目标变量是数值型，那么在数据预处理的时候，把采用的卡方值计算改为方差计算，选择最佳分割点的时候，使用的是方差分析计算得到F统计量的P值，而不是卡方的P值。



如何生成决策树-算法总结

算法	支持模型	树结构	特征选择	连续值处理	缺失值处理	剪枝
ID3	分类	多叉树	信息增益	不支持	不支持	不支持
C4.5	分类	多叉树	信息增益比	支持	支持	支持
CART	分类, 回归	二叉树	基尼系数, 残差平方和	支持	支持	支持
CHAID	分类, 回归	多叉树	独立性检验和相关性	支持	支持	支持

决策树在流行病学研究中的应用

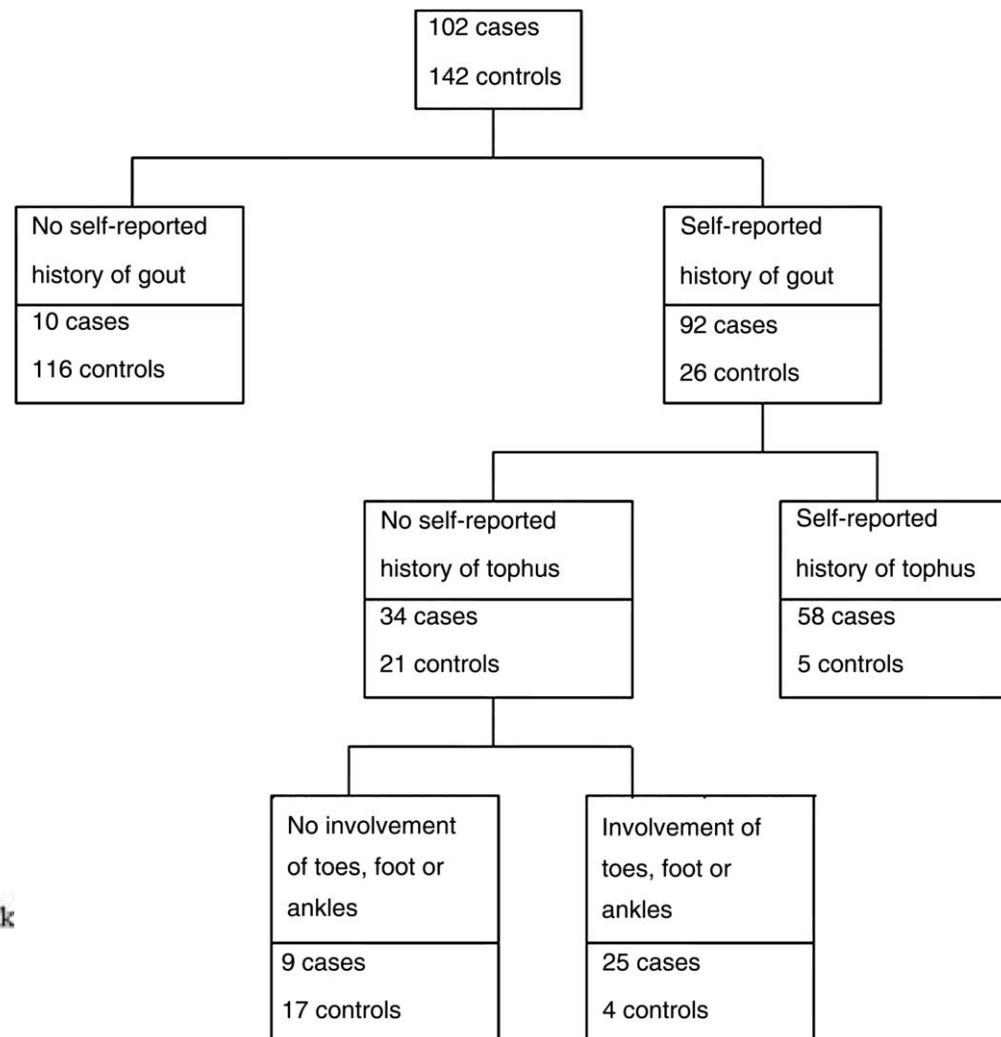
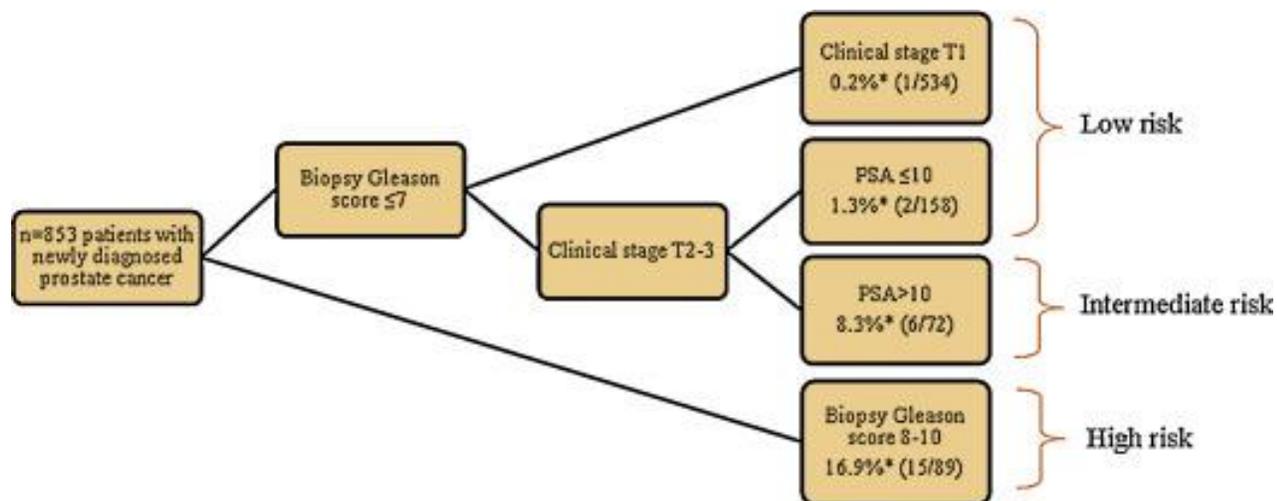
- 建立预测 (分类) 模型
- 建立危险度分层工具, 辅助医学决策

<https://doi.org/10.1016/j.jamda.2017.03.021>

<https://doi.org/10.1136/annrheumdis-2013-204976>

<https://doi.org/10.1016/j.eururo.2009.12.023>

e.g. 预测65岁以上的老年人两年内是否会跌倒
 根据问卷调查推断潜在的痛风患者
 前列腺癌患者何种情况下进行骨密度检查?





决策树在流行病学研究中的应用

- 寻找连续性变量的cut-off 值，便于后续分组分析

<https://doi.org/10.1093/eurheartj/ehw448>

<https://doi.org/10.1016/j.exger.2019.01.010>

e.g.传统分位数分组不理想，又怕被说人为操纵分组以获得有意义结果？

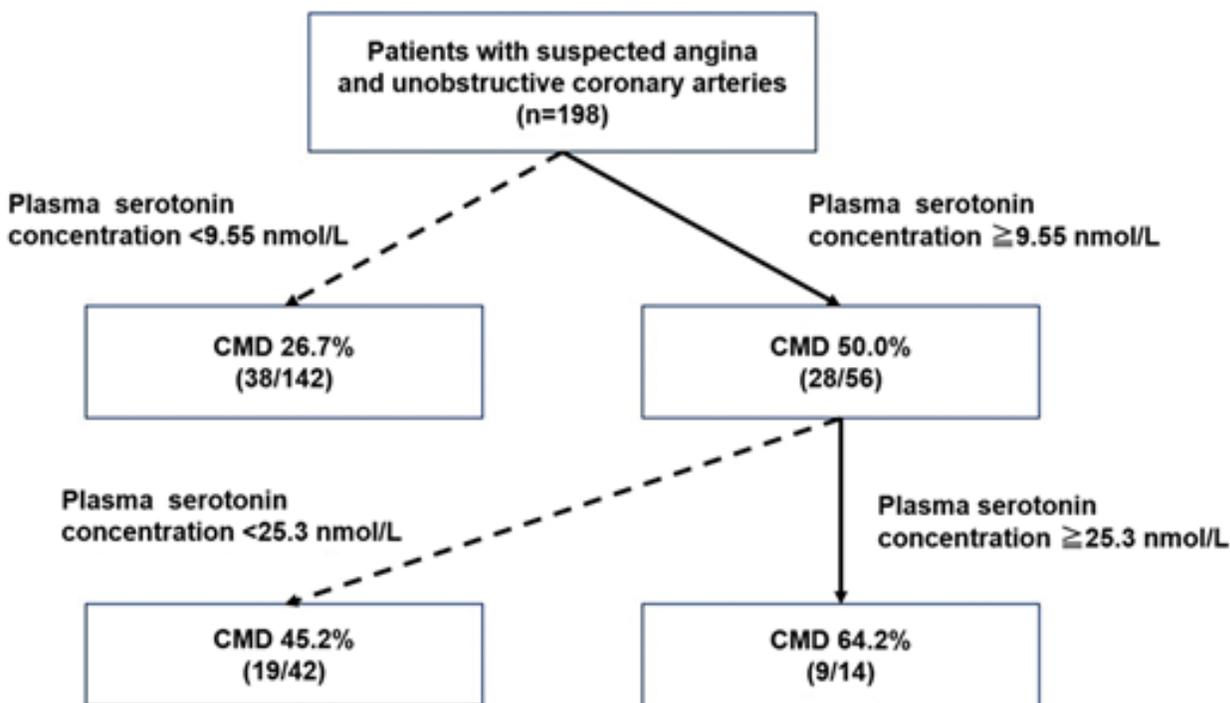


Table 2 Logistic regression analysis for the presence of coronary microvascular dysfunction in patients with suspected angina and unobstructive coronary arteries

	Univariate analysis					Stepwise analysis				
	Coefficient	Standardized coefficient	OR	95% CI	P-value	Coefficient	Standardized coefficient	OR	95% CI	P-value
Age	0.009	0.012	1.009	0.986–1.032	0.451					
Sex	−0.434	0.305	0.648	0.357–1.177	0.153					
LVEF	0.011	0.016	1.011	0.98–1.044	0.484					
Haemoglobin	−0.080	0.085	0.923	0.782–1.091	0.348					
Platelet	0.001	0.003	1.001	0.996–1.006	0.811					
CKD	−0.183	0.363	0.833	0.409–1.698	0.615					
BNP	−0.003	0.003	0.997	0.991–1.003	0.326					
hsCRP	0.473	0.786	1.604	0.344–7.481	0.548					
Hypertension	0.246	0.305	1.279	0.704–2.327	0.419	0.573	0.366	1.774	0.875–3.703	0.118
Dyslipidaemia	−0.091	0.302	0.913	0.505–1.65	0.763					
Diabetes mellitus	0.159	0.343	1.172	0.598–2.295	0.644					
Smoking history	−0.153	0.305	0.858	0.472–1.559	0.614					
Plasma serotonin concentration > 9.55 nmol/L	1.007	0.328	2.737	1.44–5.202	0.002	1.031	0.383	2.803	1.330–6.004	0.007

The ORs are for a 1 unit difference.

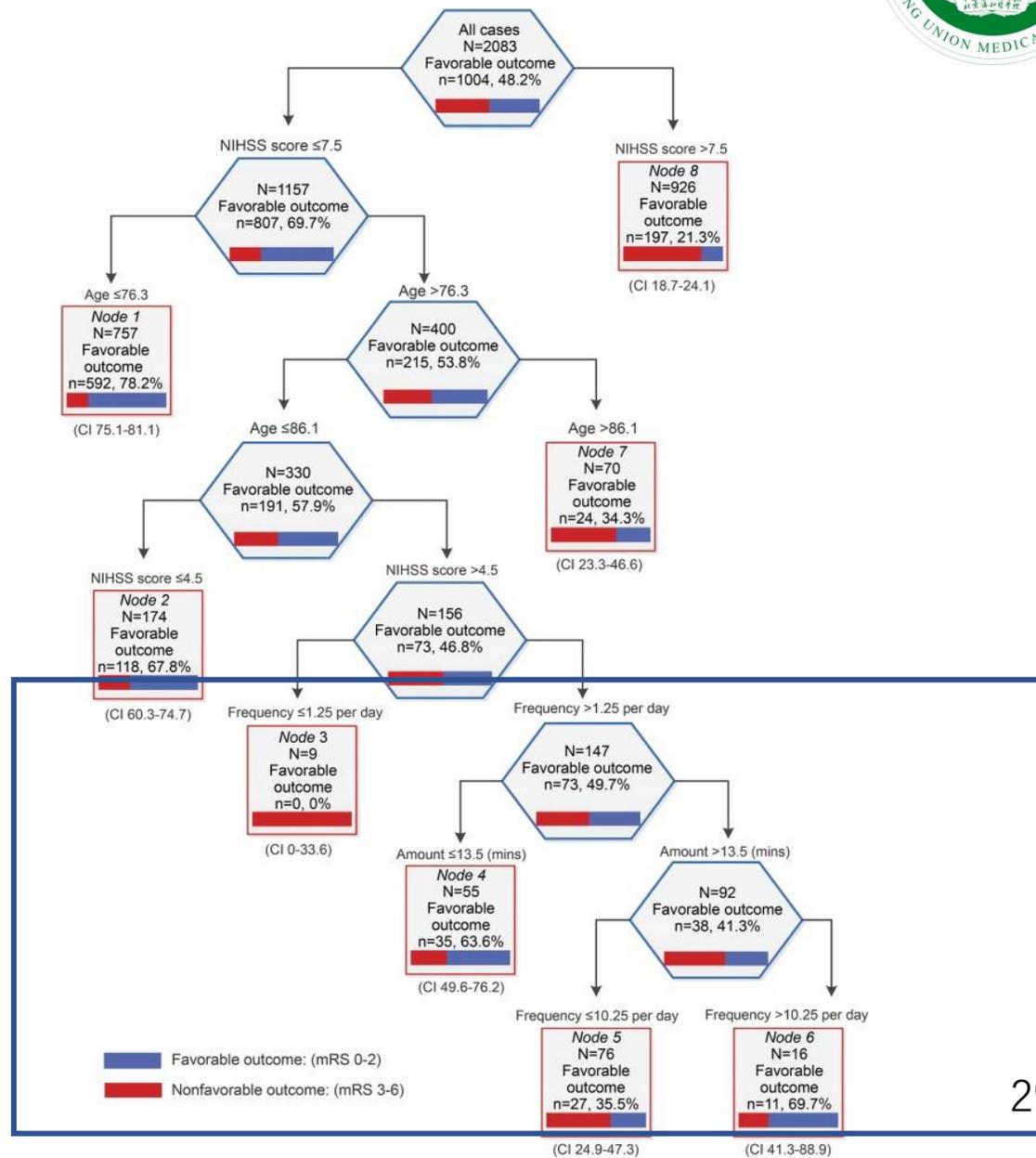
BNP, B-type natriuretic peptide; CI, confidence interval; CKD, chronic kidney disease; CMD, coronary microvascular dysfunction; hsCRP, high sensitivity C-reactive protein; LVEF, left ventricular ejection fraction; OR, odds ratio.

决策树在流行病学研究中的应用

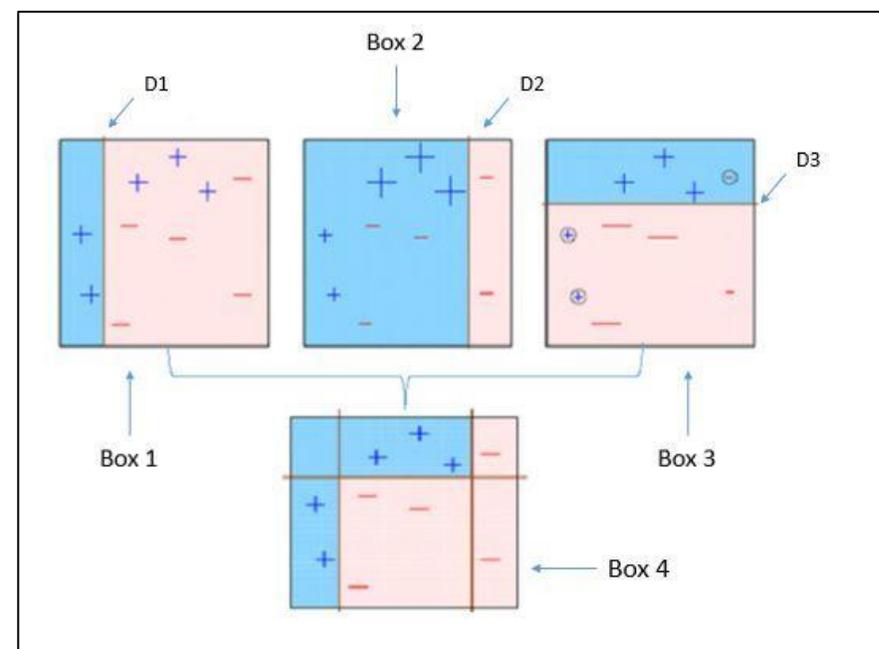
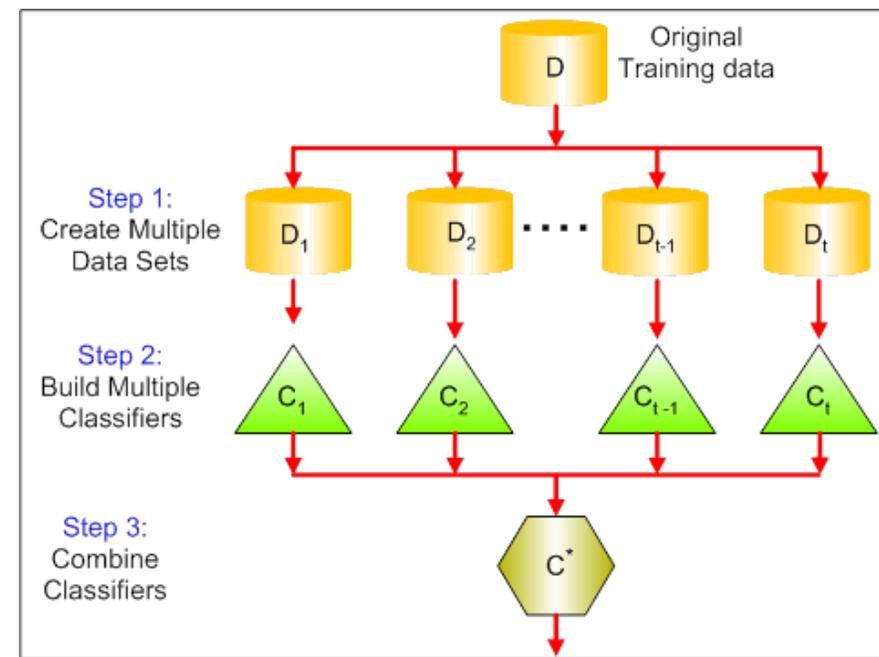
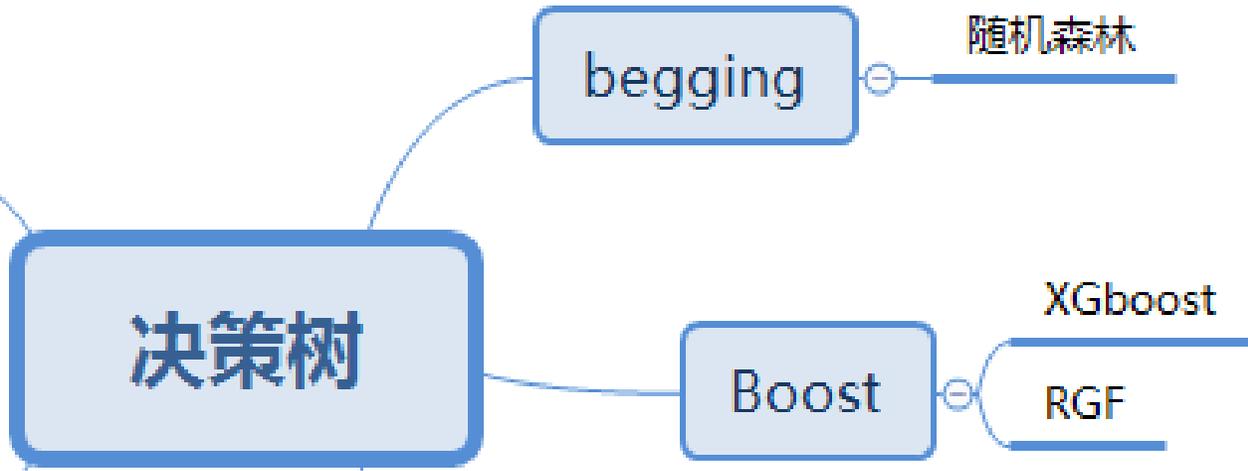
- 探寻变量间的交互作用，指导亚组分析

<https://doi.org/10.1212/WNL.0000000000002459>

e.g. 针对不同特征的患者，如何确定给药剂量和频率



树模型





随机森林：青出于蓝而胜于蓝

森林

使用多颗树进行单独预测，最后的结论由这些树预测结果的组合共同来决定。

每个基分类器可以很弱，但最后组合的结果通常能很强。

“三个臭皮匠顶个诸葛亮”

随机

保证各树之间的独立性，通常会采用两到三层的随机性。

数据抽样：随机有放回的抽取数据，数量可以和原数据相同，也可以略小。

构建分枝：随机选取N个特征，选择最好的特征进行分裂 OR 在N个最好的分裂特征中，随机选择一个进行分裂。

Moving Beyond the Forest Plot and Into the Forest



Home About Articles Submission Guidelines

We'd love your feedback. Please complete this 3 question [survey](#).

Abstract
Background
Methods
Results
Discussion
Conclusions
Declarations
References
Comments

Methodology | Open Access | Open Peer Review

Preventing false discovery of heterogeneous treatment effect subgroups in randomized trials

Joseph Rigdon , Michael Baiocchi and Sanjay Basu

Trials 2018 19:382

<https://doi.org/10.1186/s13063-018-2774-5> | © The Author(s). 2018



Journal
Journal of Computational and Graphical Statistics >

Volume 27, 2018 - Issue 1

Enter keywords, authors, D

458

Views

Data Mining

Estimating Individual Treatment Effect in Observational Data Using Random Forest Methods

Min Lu, Saad Sadiq, Daniel J. Feaster & Hemant Ishwaran

Pages 209-219 | Received 01 Oct 2016, Accepted author version posted online: 19 Jul 2017, Published online: 01 Feb 2018

Enter keywords, authors, DOI etc

Journal
Journal of the American Statistical Association >

Volume 113, 2018 - Issue 523

Theory and Methods

Estimation and Inference of Heterogeneous Treatment Effects using Random Forests

Stefan Wager & Susan Athey

Pages 1228-1242 | Received 01 Dec 2015, Accepted author version posted online: 21 Apr 2017, Published online: 06 Jun 2018

9,888

Views

33

CrossRef citations to date

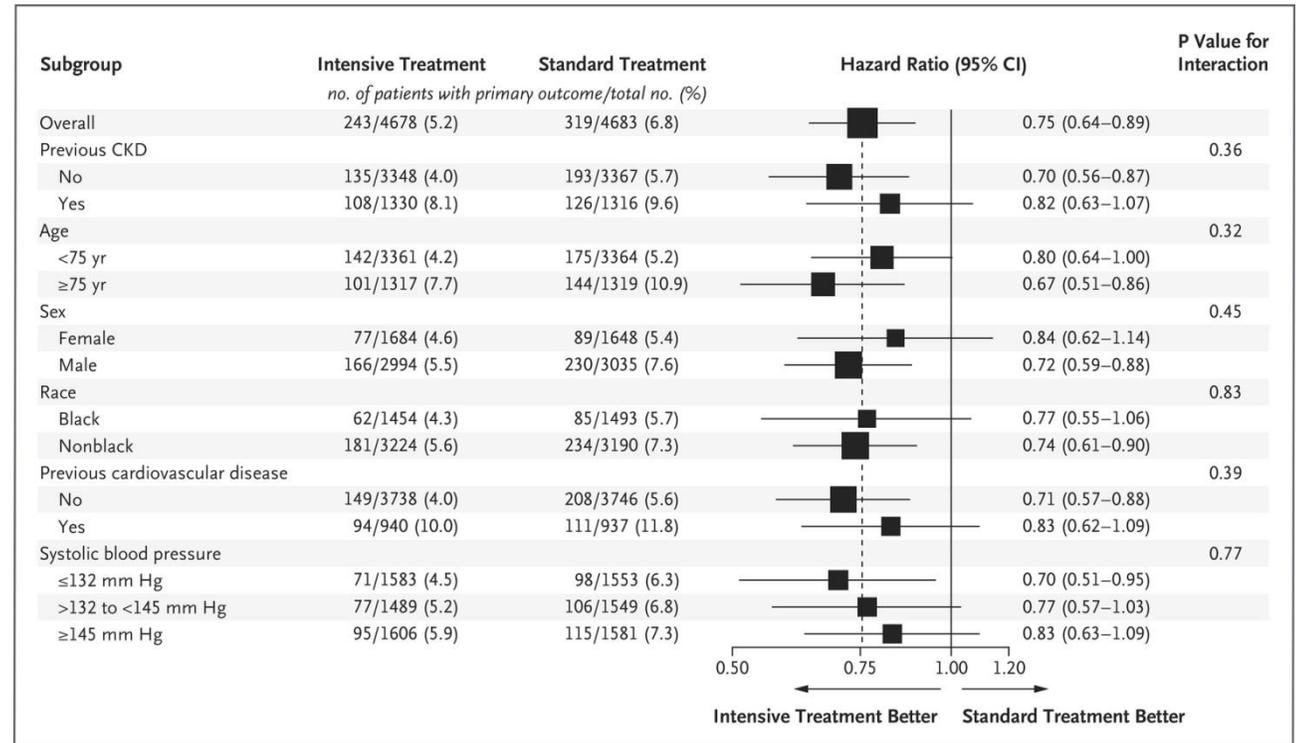
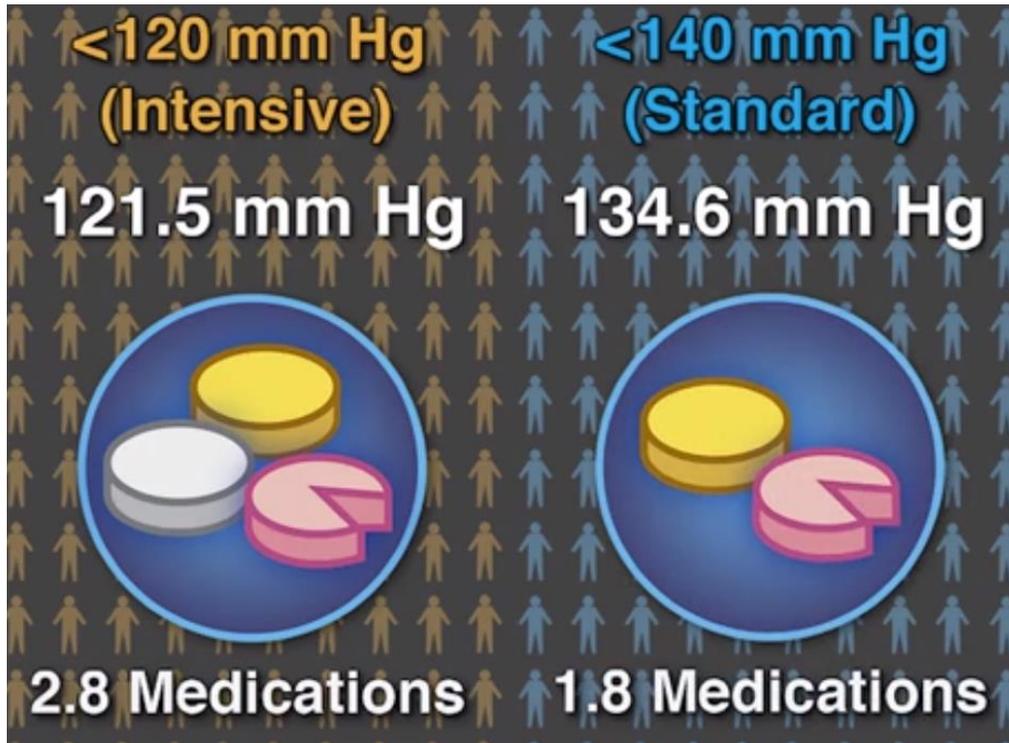
8

<https://doi.org/10.1056/NEJMoa1511939>

Moving Beyond the Forest Plot and Into the Forest



The SPRINT Research



Moving Beyond the Forest Plot and Into the Forest



Table 2. Observed Outcomes by Treatment Group Using Testing Data, Stratified by Subgroup Identified Using Training Data

Primary Composite Outcome ^a	No. of Patients (No. of Events)		HR (95% CI)	P Value
	Treated	Control		
Overall	2354 (121)	2326 (160)	0.7 (0.6-0.9)	.02
Subgroups				.02 ^b
Current smokers with baseline systolic blood pressure >144 mm Hg	110 (12)	126 (6)	10.6 (1.3-86.1)	.03
Remainder of trial population	2244 (109)	2200 (154)	0.7 (0.5-0.9)	.003

Abbreviation: HR, hazard ratio.

^a Includes myocardial infarction, other acute coronary syndromes, stroke, heart failure, or death due to cardiovascular causes.

^b P value for the interaction term; false discovery rate for the interaction term is 0.007.

参考文献I



- Gomez, F., Wu, Y. Y., Auais, M., Vafaei, A., & Zunzunegui, M. V. (2017). A Simple Algorithm to Predict Falls in Primary Care Patients Aged 65 to 74 Years: The International Mobility in Aging Study. J Am Med Dir Assoc, 18(9), 774-779. doi:10.1016/j.jamda.2017.03.021
- 《机器学习》，周志华，决策树章节
- 《统计学习方法》，李航，决策树章节
- <https://blog.csdn.net/am290333566/article/details/81187562>
- <https://blog.csdn.net/u012328159/article/details/79413610>
- <http://leijun00.github.io/2014/10/decision-tree-2/>
- [https://datajobs.com/data-science-repo/Decision-Trees-\[Rokach-and-Maimon\].pdf](https://datajobs.com/data-science-repo/Decision-Trees-[Rokach-and-Maimon].pdf)
- <https://www.zhihu.com/question/22178202/answer/49929786>

参考文献II



- Richette, P., Clerson, P., Bouée, S., Chalès, G., Doherty, M., Flipo, R. M., . Bardin, T. (2015). Identification of patients with gout: elaboration of a questionnaire for epidemiological studies. *Annals of the Rheumatic Diseases*, 74(9), 1684. doi:10.1136/annrheumdis-2013-204976
- Takahashi, J., Nishimiya, K., Ito, K., Hao, K., Tsuburaya, R., Miyata, S., . Hirowatari, Y. (2016). Plasma concentration of serotonin is a novel biomarker for coronary microvascular dysfunction in patients with suspected angina and unobstructive coronary arteries. *European Heart Journal*, 38(7), 489-496. doi:10.1093/eurheartj/ehw448 %J European Heart Journal
- The SPRINT Research Group (2015). A Randomized Trial of Intensive versus Standard Blood-Pressure Control. *NEJM*. 373(22), 2103-2116. doi:10.1056/NEJMoa1511939
- Scarpa J, Bruzelius E, Doupe P, Le M, Faghmous J, Baum, A. (2019). Assessment of risk of harm associated with intensive blood pressure management among patients with hypertension who smoke: a secondary analysis of the Systolic Blood Pressure Intervention Trial. *JAMA Netw Open*, 2(2):e190005. doi:10.1001/jamanetworkopen.2019.0005



You can reach me via

e-mail: dianqinsun@student.pumc.edu.cn

wechat: SDQ19960601