集成分类树及其在脑梗塞预后分析中的应用

谢益辉

中国人民大学统计学院

2007年12月8日

目录

统计的两大任务

描述统计

描述：刻画过去的事实

- 所谓的“描述统计”本质上还是推断
- 描述的做法往往都是扔掉原始数据提炼参数信息，这是一种权衡
- 意识到描述统计会损失原始信息的特点有助于对统计更好的理解

统计的两大任务（续）

推断统计

推断：探索发展的规律

- 推断统计和算命先生都在作预测，算命先生也会用 k-NN 算法
- 推断最大的前提是：我们假定样本是具有代表性的
- 代表性意味着代表了规律：若事件本无规律，再好的统计也是没辙

本文的内容

给可能发生脑梗塞的病人算算命

缘起

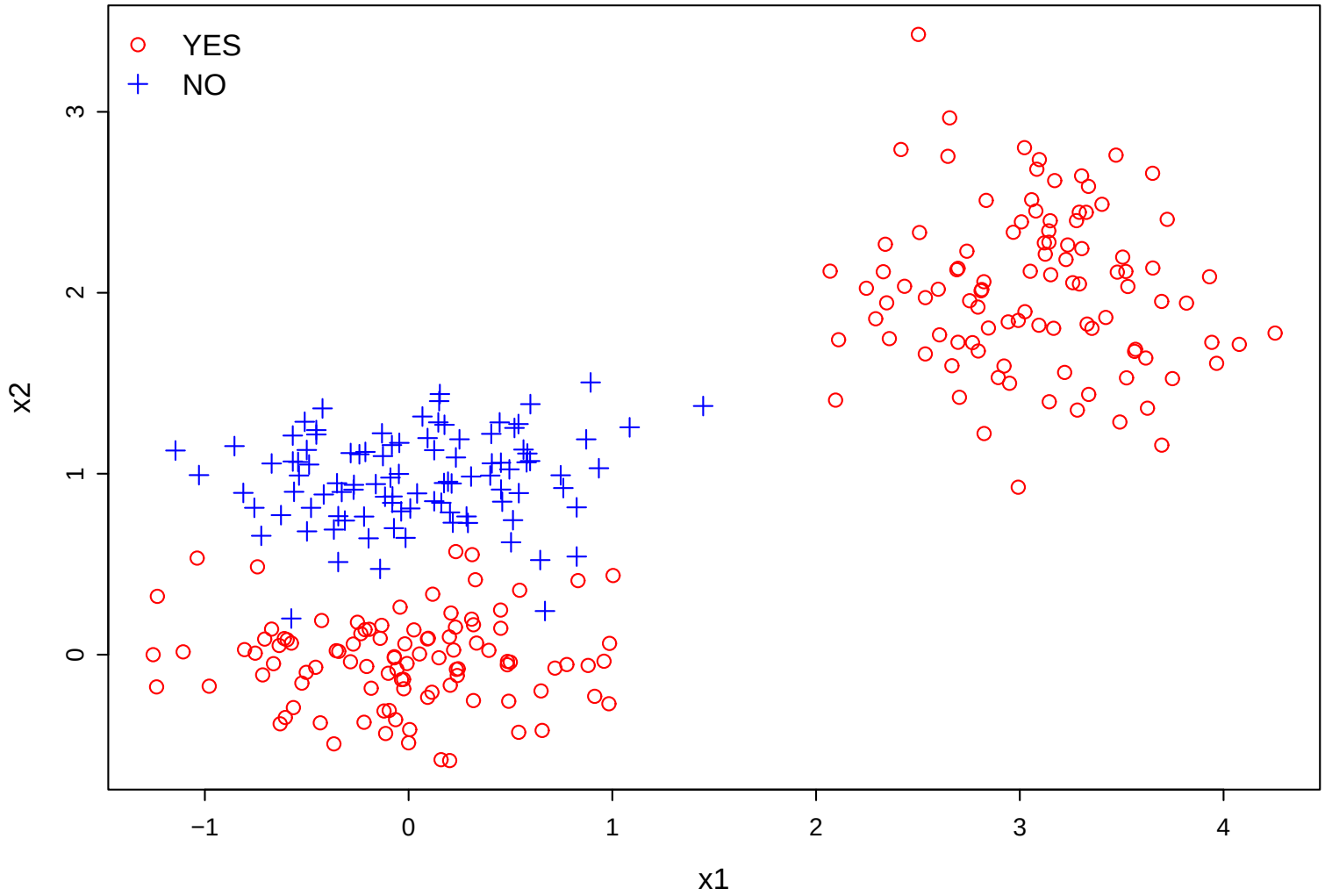
- 很多老年人患脑血管疾病，但我们可以收集一些前兆
- 比如短暂性脑缺血（Transient Ischemic Attack，TIA）的发作等
- 希望通过一些观测的自变量来预测将来脑梗塞是否会发作
- 数据来源：北京友谊医院神经内科，选取自2005年1月至2007年1月在北京友谊医院神经内科住院治疗的 TIA 患者
- 106 例病人；1 个因变量（是否发展为脑梗塞）；23 个自变量，包括是否患高血压、胆固醇浓度、病因等

模型

- 抓到老鼠的就是好猫，关键是预测准确性
- 当然若模型也能够反映一定的逻辑思路更好（避免黑箱）

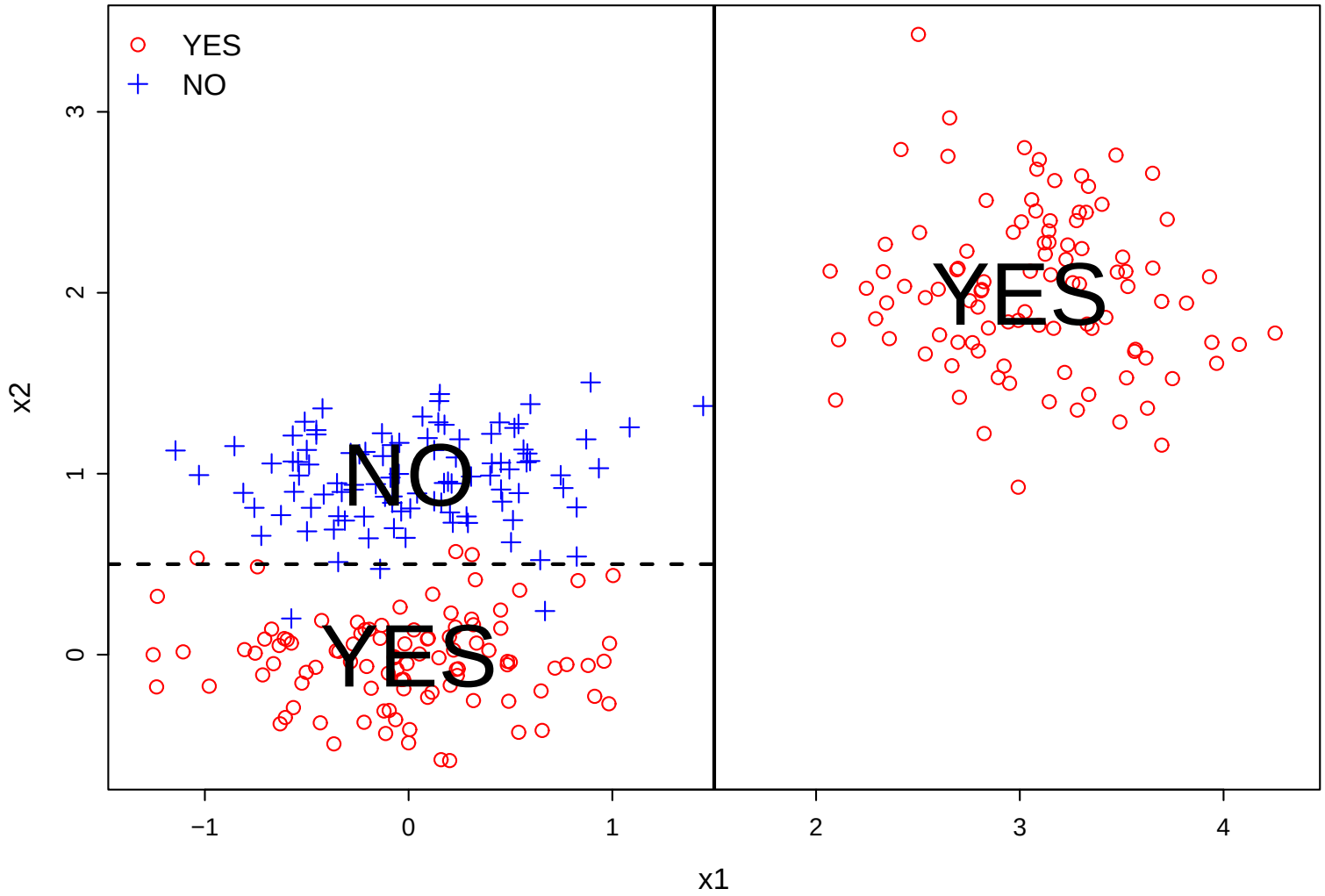
统计方法是高深莫测的吗?

未必！看看这批数据，直觉告诉你该怎样预测“+”和“o”?



统计方法是高深莫测的吗？（续）

为什么我们会这么想？

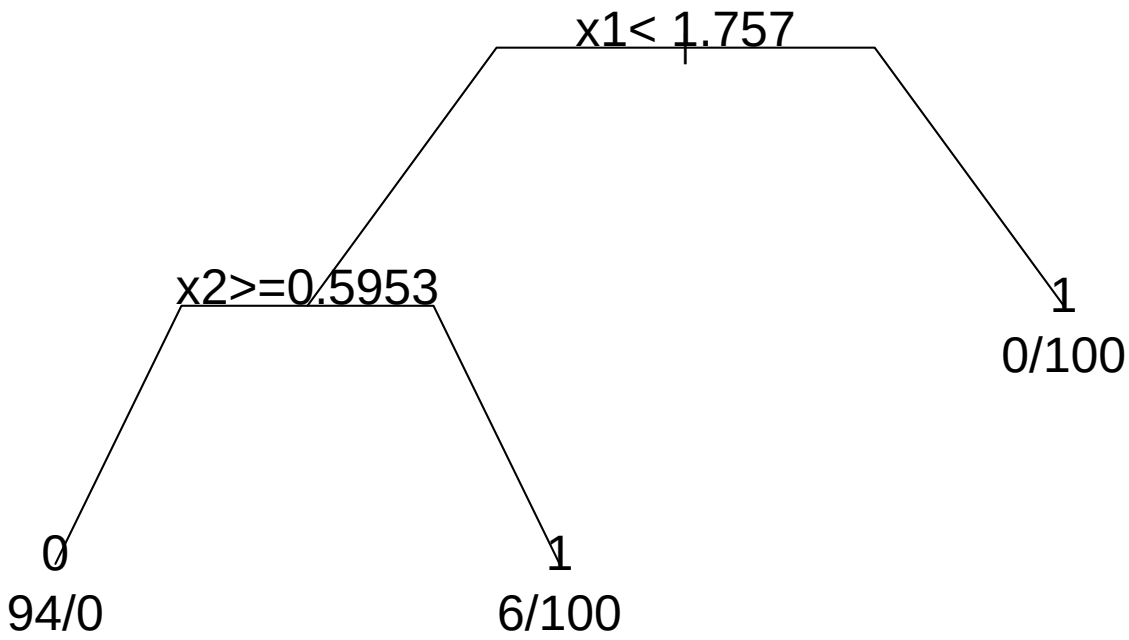


只有直觉是不够的

比如我们可以凭直觉想象，每一组样本均值的数值差异越大，用统计检验方法（比如 t 检验、方差分析等）检验出来的结果也应该是倾向于显著。但若没有下面这样的式子，统计就不是科学了：

$$
\begin{aligned}
SS_{\text{total}} &= SS_{\text{error}} + SS_{\text{treatments}} \\
\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\bar{X})^2 &= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2 + \sum_{i=1}^{r}(\bar{X}_i-\bar{X})^2 \\
F &= \frac{SS_{\text{treatments}}/(r-1)}{SS_{\text{error}}/(n-r)} \sim F(r-1, n-r)
\end{aligned}
$$

分类树的理论



- （倒）树形结构，由上一个节点生长到下一个节点，最后到叶子
- 反映出来的是一系列判断规则，而叶子给出了最终分类
- 构建树的基本要点：生长（分裂）和剪枝

分类树的理论（续）

生长

- 扫描所有变量，找到合适的分裂点
- 规则：每一次分裂都可以使分类的纯度更高
- 纯度的表达：误分类比例、Gini系数、熵

剪枝

- 原因：树的无限增长会导致每片叶子里面只有一个样本点，完全拟合的树在预测时一般都是极不稳定的
- 规则：生长带来的纯度提高小于一个阈值，那么就砍掉这根树枝

我一棵树太孤单

单个分类器的问题（参读论文 III 节 2 小节）

- 经不起风浪（独生子女？）数据的扰动可能会让它的稳定性变差
- 扶不起的阿斗：无论怎样训练模型，精度就是没法提高

怎么办？团结就是力量

- 实现人民民主，大家来进行民主投票预测，化解独裁者一意孤行带来的风险
- 回到幼儿园时老师的教导：知错就改就是好孩子；不断指导“坏样本”走正道

本文的集成方法

方法论

1. Bagging
2. Boosting

好方法的特征

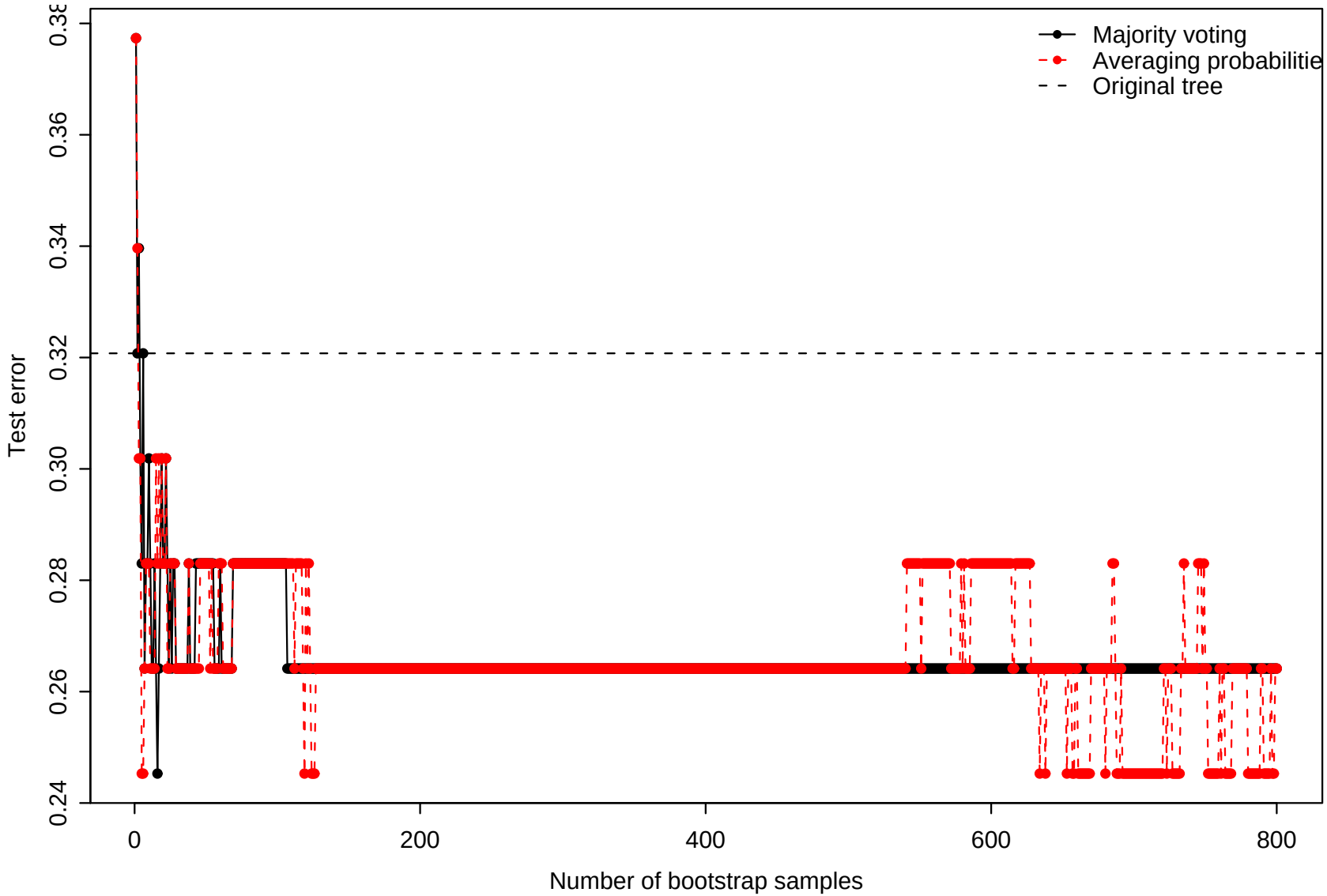
1. 对于历史数据拟合得好
2. 对于未来数据预测得准

本论文采用了数据挖掘和机器学习的一贯做法：将所有数据拆分为两部分，一部分用来训练模型，剩下的一部分作测试用。从训练数据得到的误差反映了模型对数据的拟合程度，从测试数据得到的误差反映了模型的预测能力。（什么是Cross-Validation？）

Bagging 方法：民主的好处

- 做法：用 *Bootstrap* 方法（什么是 Bootstrap）从原样本中不断重新抽样得到一批批新样本，根据新样本重新建立一棵棵树，最后在预测的时候就不只是拿一棵树来预测，而是用这一批树一起预测，并将预测值作民主投票，哪个分类票数多就预测为哪一类
- 好处：避免了单棵树受样本扰动而不稳定的缺点，并且也可以提高预测精度

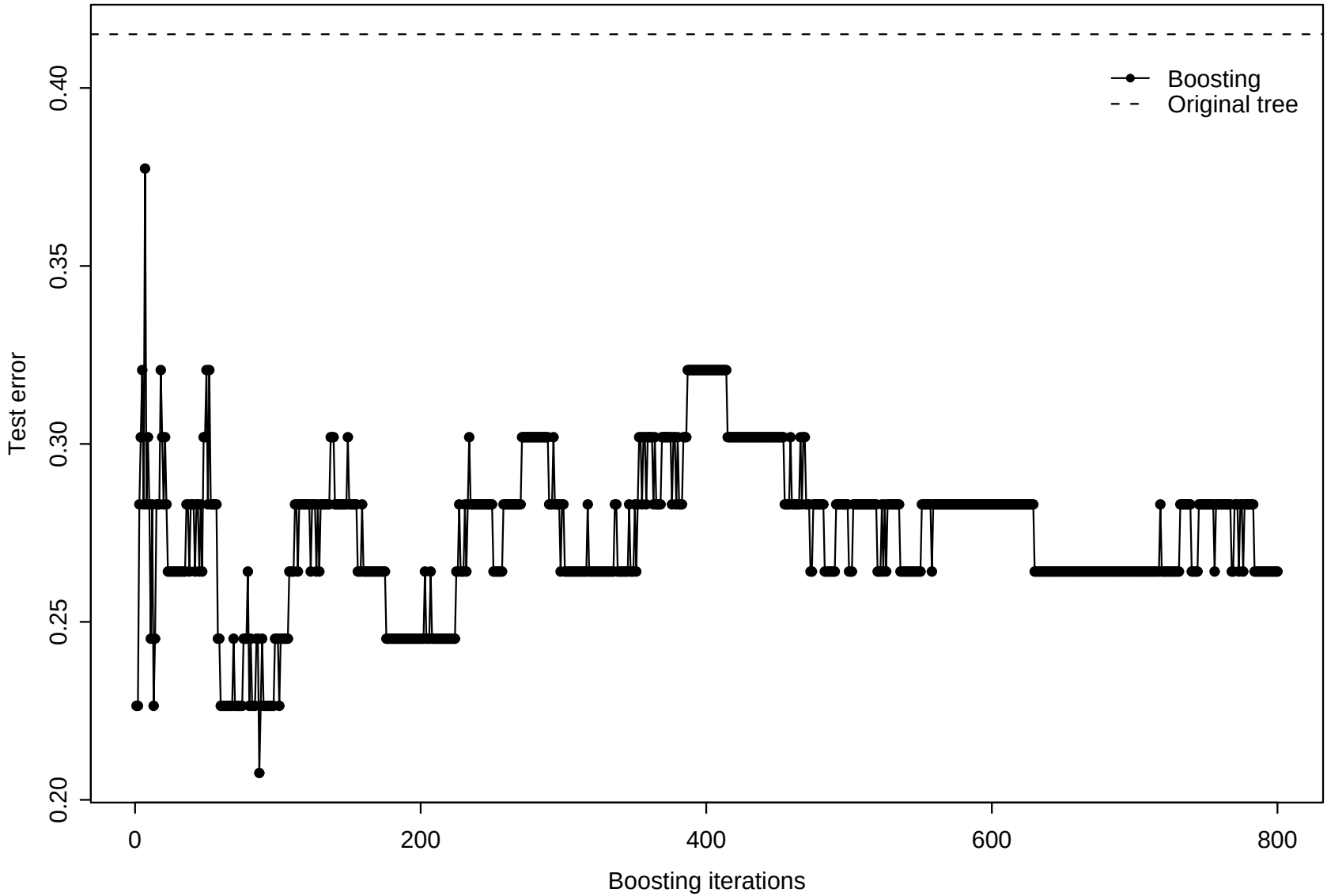
别看广告，看疗效（Bagging）



Boosting 方法：帮助误分类样本改邪归正

- 训练模型不是像 Bagging 那样并行的，而是有顺序的一串模型
- 每一次训练模型几乎都会有数据点的分类出现错误，那么下一次的模型就应该更关注这些分错的样本点
- 具体方法是通过*抽样权重*来实现：在重抽样的时候加大对误分点的权重，这样它们在下一次训练模型时就更容易出现，模型也会更专注于训练误分样本
- 最后，把这些模型根据各自的预测能力综合起来（好模型给大权重，差模型反之），用于新的预测

别看广告，看疗效（Boosting）



两点讨论（参读论文 IV 节）

- Bagging 树打乱了树的结构，从而不容易看出预后因素是如何影响脑梗塞的发作，怎么办？
- 本文的小小措施：仅分析树桩
- Boosting 对于弱分类器有好的效果，但对于好的分类器可能会有反作用
- 具体问题具体分析，理论结合实践

关于医学数据的零碎感触

- 总体感觉：很有挖掘价值
- 总是感觉：收集和处理数据的方式不合理
- 具体症状：将连续数据离散化、数据库的建设显得较随意、欠缺数据库理论
- 关于本论文的数据：样本量显得有点少，只有 106 例，是否具有代表性和推广性值得怀疑（预测误差总是很高）

关于医学模型的开发应用

- 处处都是 Logistic 回归和卡方检验的身影，在“中国期刊网数据库”中检索“分类与回归树”，结果只有寥寥数十篇论文
- P-value 能代表什么？（代表了统计先进生产力的发展要求？统计先进文化的前进方向？最广大统计群众的根本利益？……）
- 回到算命的主题，谁不希望算得准一些？谁不希望急性疾病能早点得到预防？
- 统计与医学，怎样更好地沟通和结合？……

谢谢大家！

这是一张有点费力的名片……