基于随机森林的献血招募模型

答辩人：宋子星
指导教师：薛晖

东南大学计算机科学与工程学院

June 4, 2020

目录

SOUTHEAST UNIVERSITY

目录

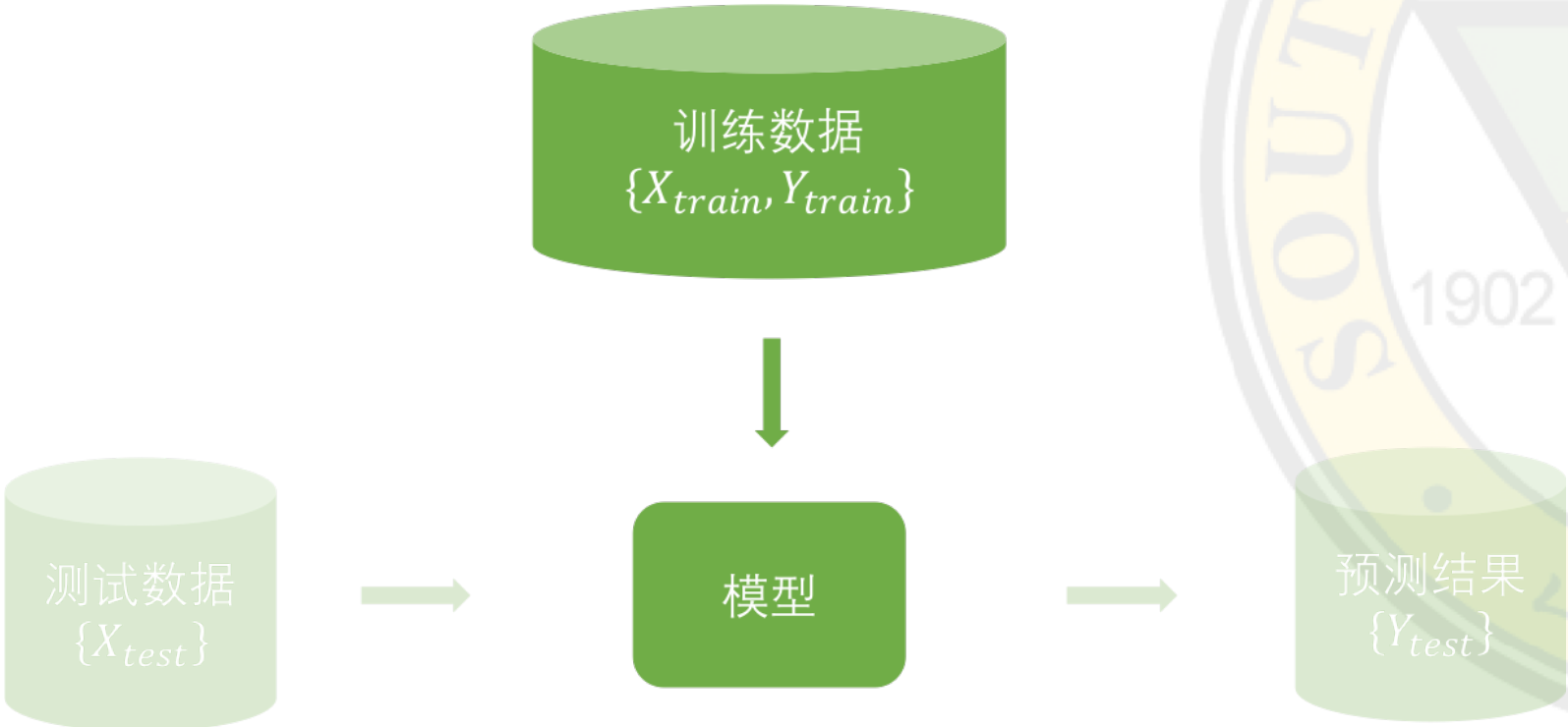
背景

现状与意义

- 医疗领域：当前献血招募主要采取人力招募方式
- 计算机领域：机器学习技术迅猛发展
- 献血招募 + 机器学习 $\Rightarrow$ 提升招募精度 + 降低招募成本?

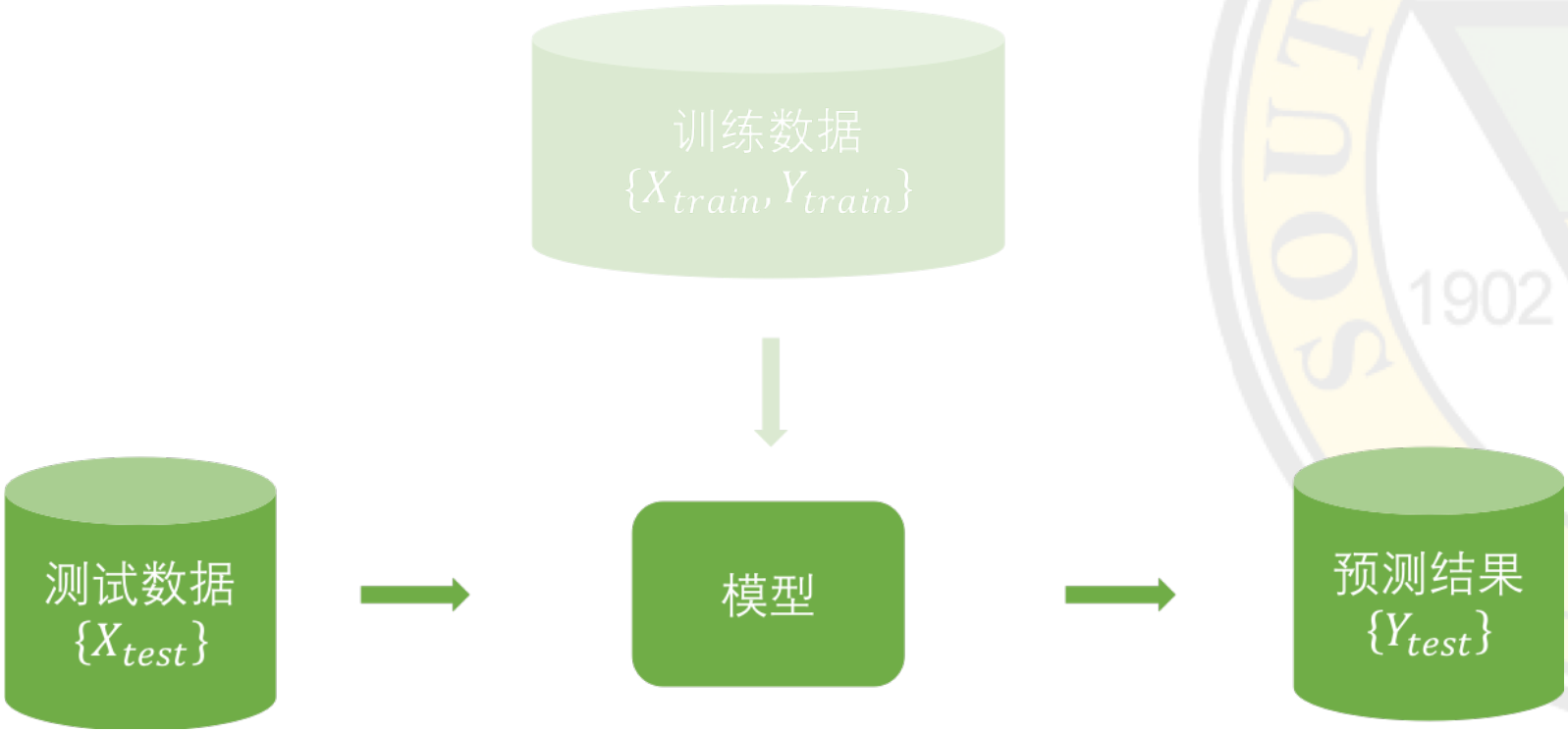
研究问题

年龄	性别	历史献血次数	历史献血总量	……	献血资格
38	女	2	200	…	有
17	男	0	0	…	无
…	…	…	…	…	…



研究问题

年龄	性别	历史献血次数	历史献血总量	……	献血资格
38	女	2	200	…	有
17	男	0	0	…	无
…	…	…	…	…	…



SOUTHEAST UNIVERSITY

目录

集成学习 (Ensemble Learning)

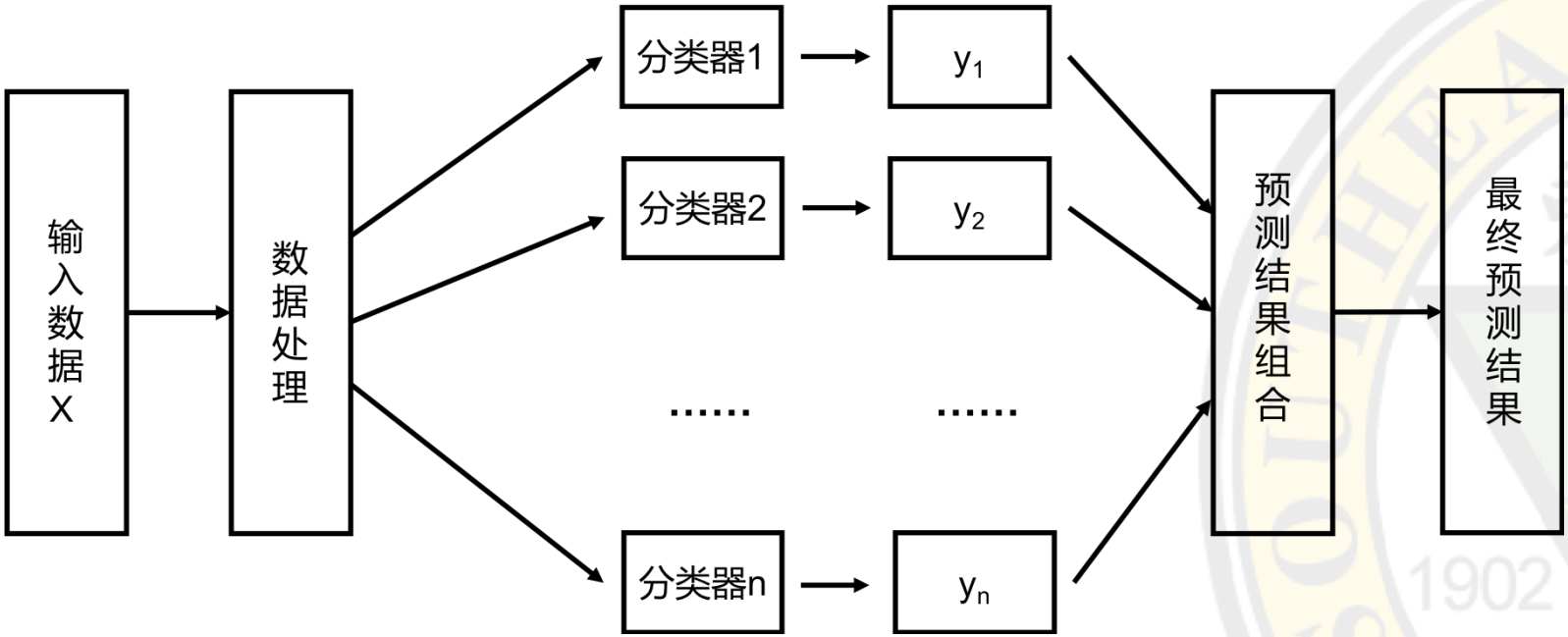
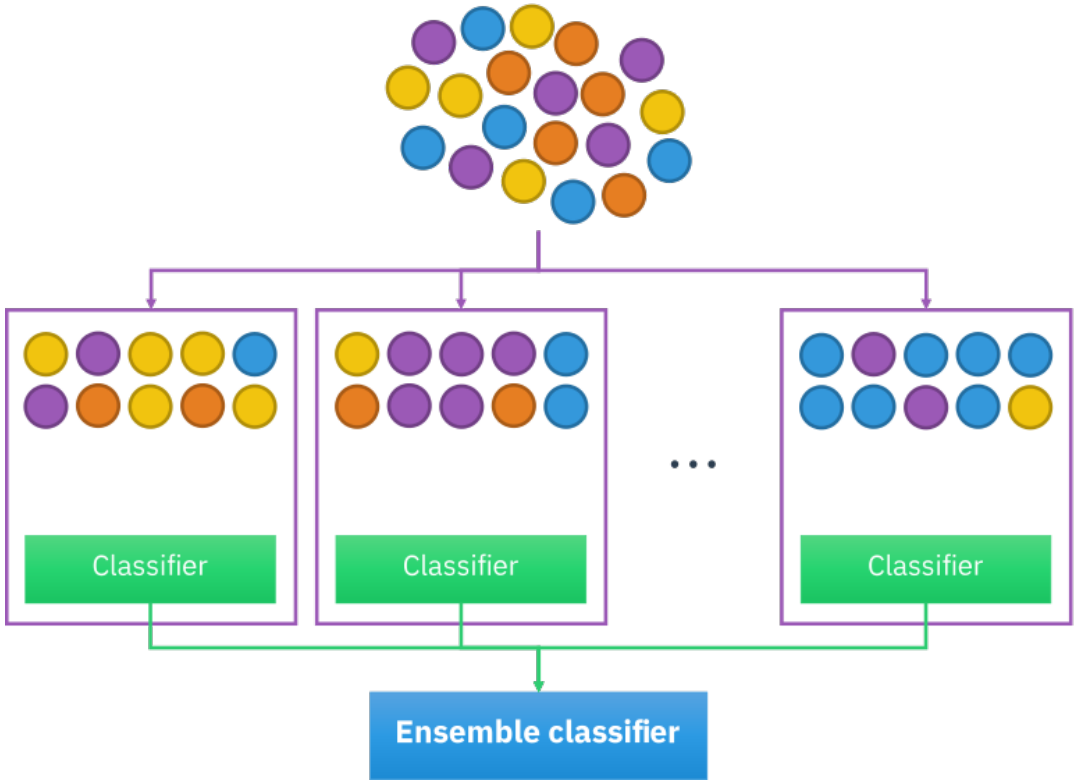


图 1: 集成学习流程框架

集成学习是指用于训练多个学习器并组合其输出，可以将其视为“决策委员会”的投票决策结果。

Bagging



Original Data

Bootstrapping

Aggregating

Bagging

图 2: Bagging 算法

Bagging 利用有放回抽样生成新的训练数据集 (Bootstrap samples)。

决策树 (Desicion Tree)

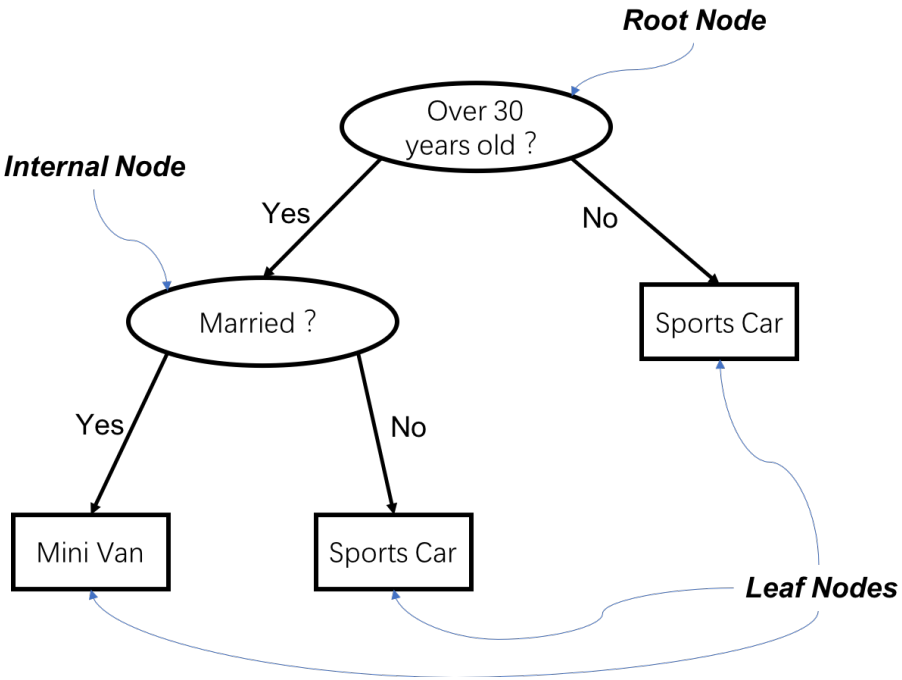


图 3: 决策树案例

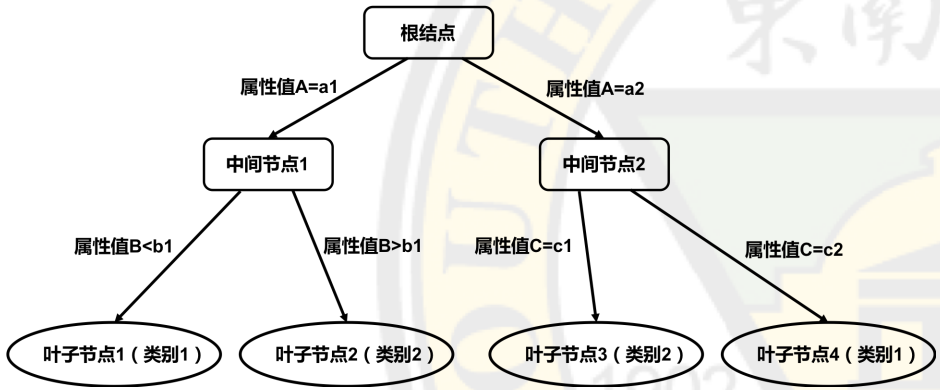
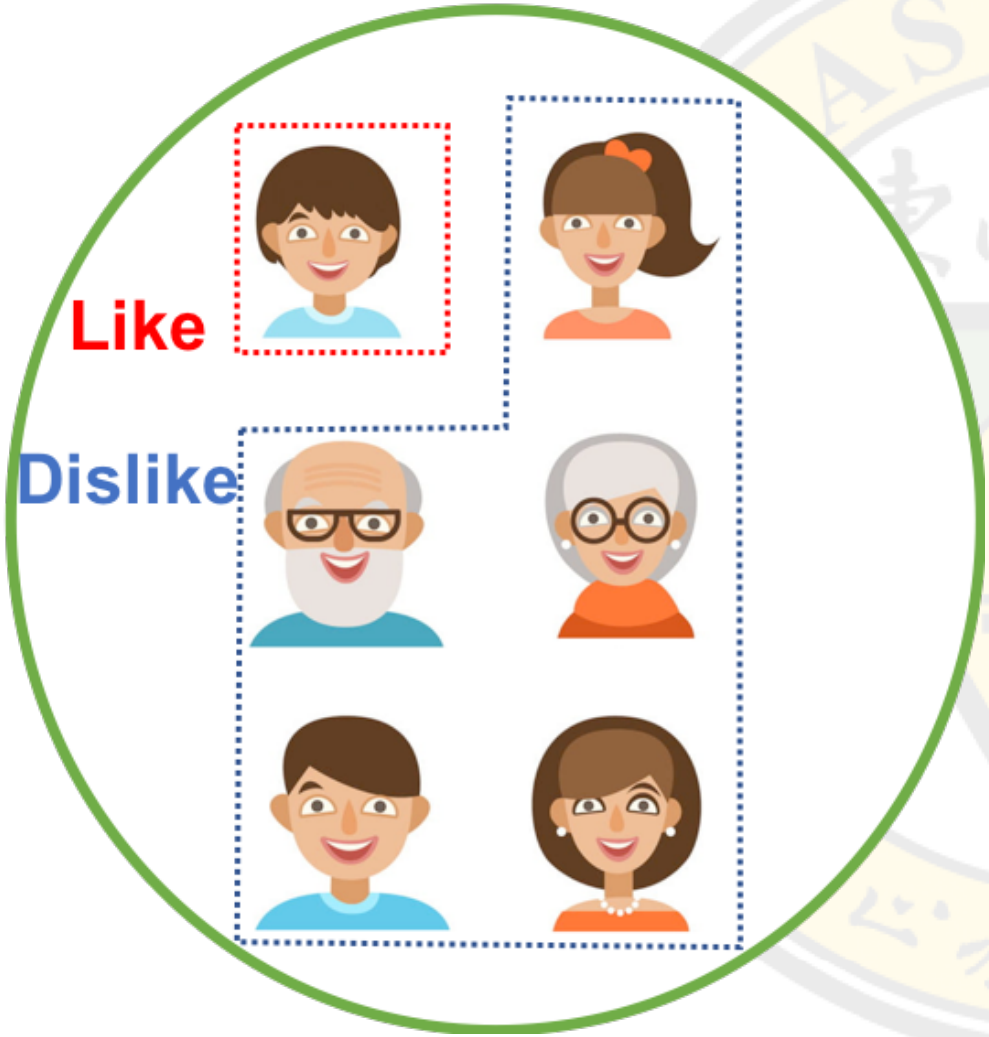


图 4: 决策树归纳案例

决策树的训练：构建一棵决策树（ID3，C4.5，CART）。
决策树的测试：自顶而下匹配一条路径。

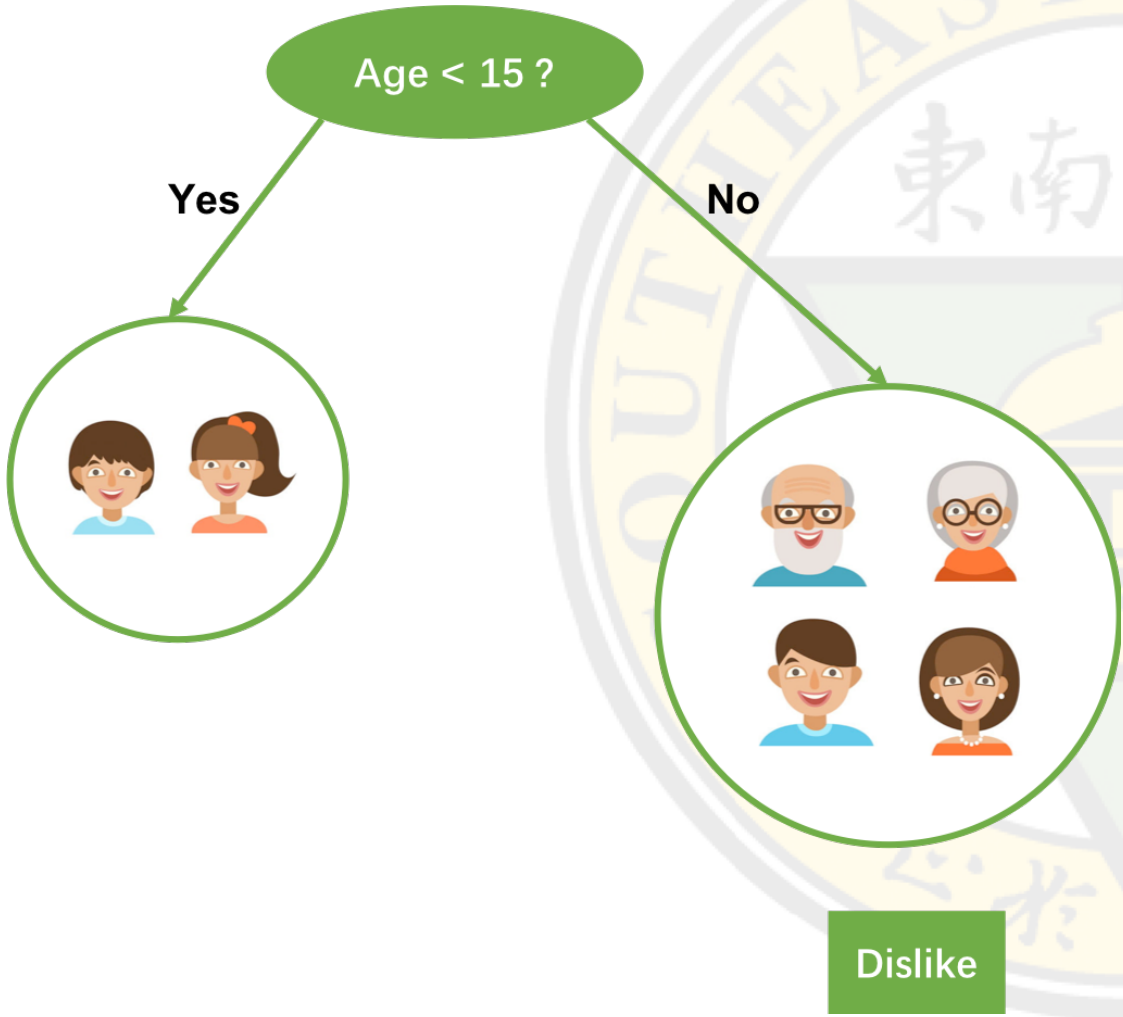
决策树生成算法

- 训练集样本特征：性别、年龄和职业。
- 预测目标：是否喜欢玩电脑游戏。
- 算法关键：每次寻找出当前最佳分割属性。



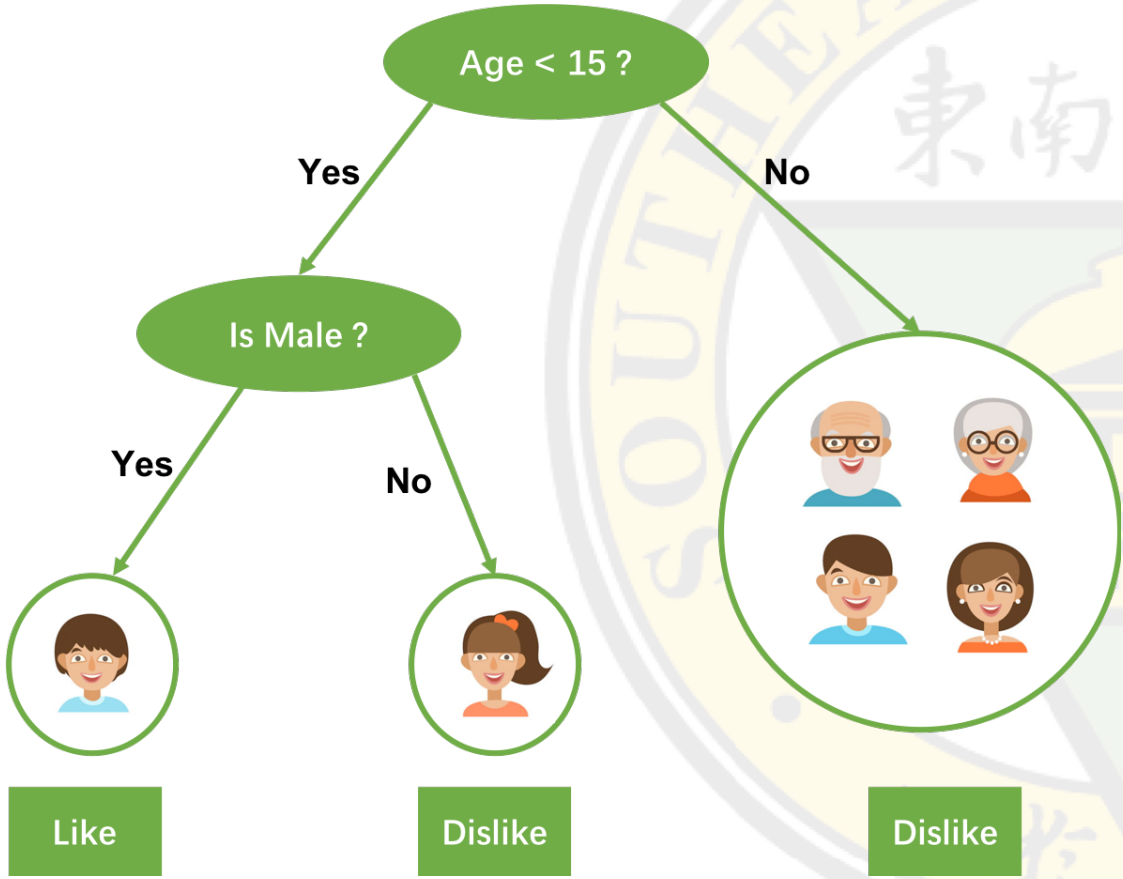
决策树生成算法

- 训练集样本特征：性别、年龄和职业。
- 预测目标：是否喜欢玩电脑游戏。
- 根据某分割指标，从 3 个属性中选择**年龄**作为分割属性，分裂节点。
- 无需分割，则停止分裂节点。



决策树生成算法

- 训练集样本特征：性别、年龄和职业。
- 预测目标：是否喜欢玩电脑游戏。
- 对需要再次分割的节点，根据某分割指标，从剩下的 2 个属性中选择**性别**作为分割属性，分裂节点。



决策树生成算法

表 1: 三种最常见的决策树生成算法比较

生成算法	分割指标	支持的属性	缺失值处理
ID3	信息增益	仅离散属性	不支持
C4.5	信息增益率	离散、连续属性	支持
CART	基尼指数	离散、连续属性	支持

随机森林 (Random Forest)

随机森林的随机性

- 随机森林 = Bagging + 决策树（CART）
- 训练集生成的随机性 $\Rightarrow$ Bagging (Bootstrap 样本)
- 特征变量选取的随机性 $\Rightarrow$ 决策树生成中，每次分裂节点时，随机选择一部分特征作为候选分割属性，常见 $M = \sqrt{N}$，$M = \log_2 N$，再从候选属性中，寻找出最佳分割属性。

随机森林 (Random Forest)

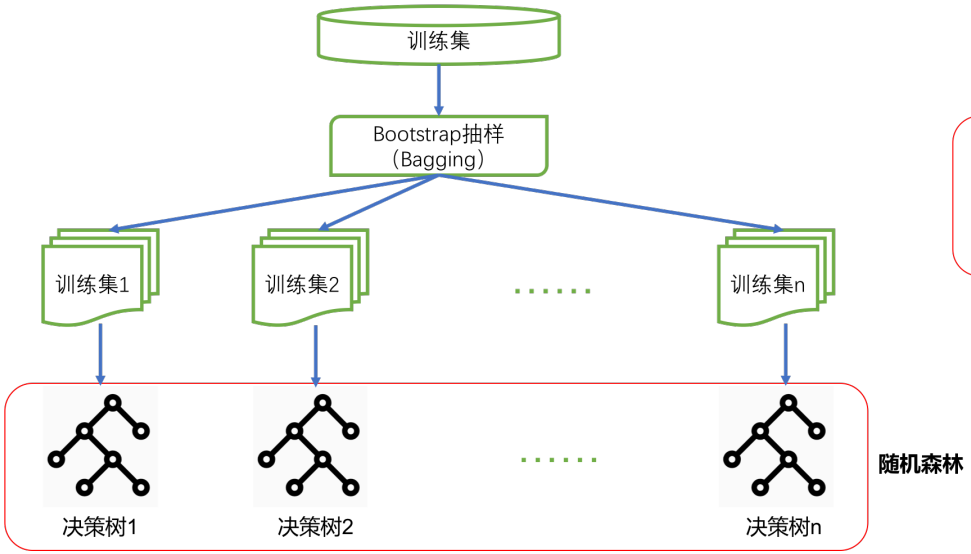


图 5: 随机森林训练过程

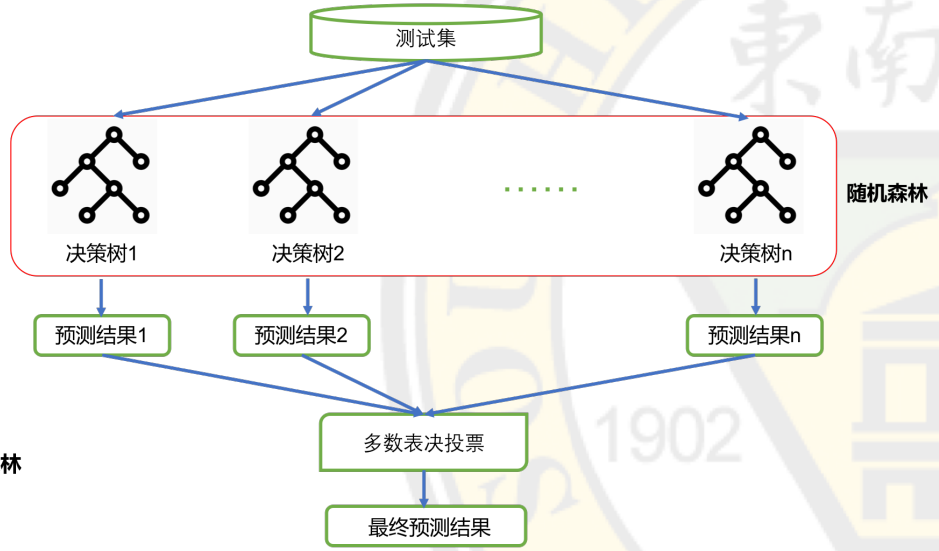


图 6: 随机森林测试过程

目录

献血者特征提取

献血者ID	登记时间	献血量	采样时间
00****008	2016-07-02	200	2016-07-02
献血地点	血型	Rh血型	性别
站内	B	**D**	女
出生日期	国籍	民族	居住类型
1978-10-02	中国	汉	
职业	文化程度	所属区县	工作组
其他	其他	扬州市	
实际采血量	献血方式	采血类型	有非标量
200	无偿	采血	标量
献血反应有无			
无			

图 7: 原始献血记录

年龄	性别	最近一次献血量
连续	离散	连续
总献血量	献血次数	上次献血合格与否
连续	连续	离散
职业	献血间隔	受教育程度
离散	连续	离散
居住情况	献血频率	献血反应有无
离散	连续	离散

图 8: 特征提取结果

决策树和随机森林的实现

主要贡献

- 利用层次遍历的思想，将主流的递归版本的决策树生成算法转化为非递归版本。
- 利用 graphviz，将决策树和随机森林可视化。
- 利用 UCI 公开数据集，初步评测了模型性能。

表 2: 3 种决策树时间性能比较 (UCI 数据集)

算法版本	ID3 (ms)	C4.5 (ms)	CART (ms)
递归	342.64±7.23	225.36±4.68	285.41±10.23
非递归	**213.45±6.13**	**121.39±8.09**	**135.67±8.79**
提升率	37.70%	46.13%	52.64%

决策树和随机森林的实现

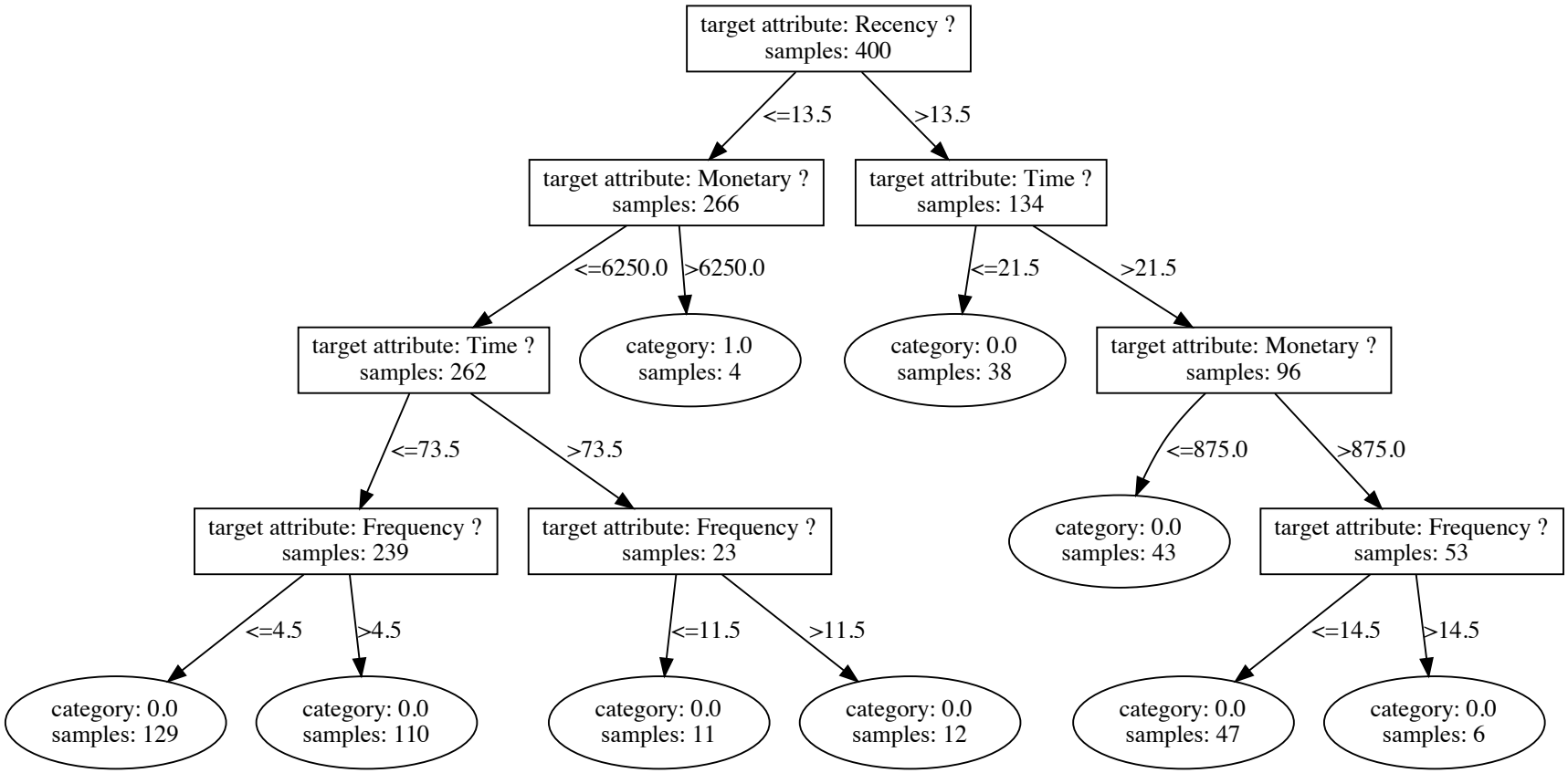


图 9: 训练的随机森林中的某棵决策树可视化（UCI 数据集）

目录

数据集特征分布分析

特征分布

- 绘制了箱形图（Box-whisker Plot）和直方图（Histogram）。
- 利用核密度估计（Kernel Density Estimation）绘制出了变量的概率密度函数。

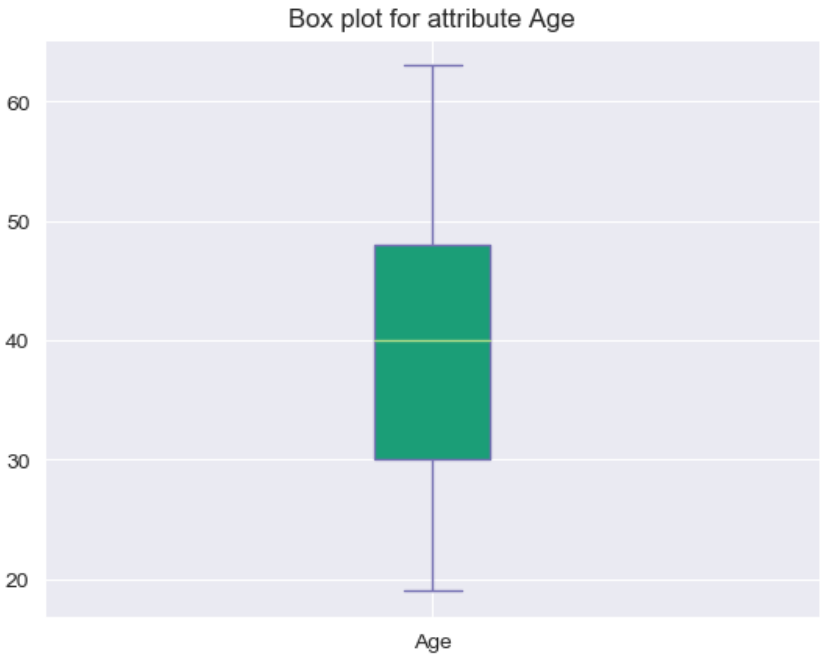


图 10: 特征**年龄**的箱形图（UCI 数据集）

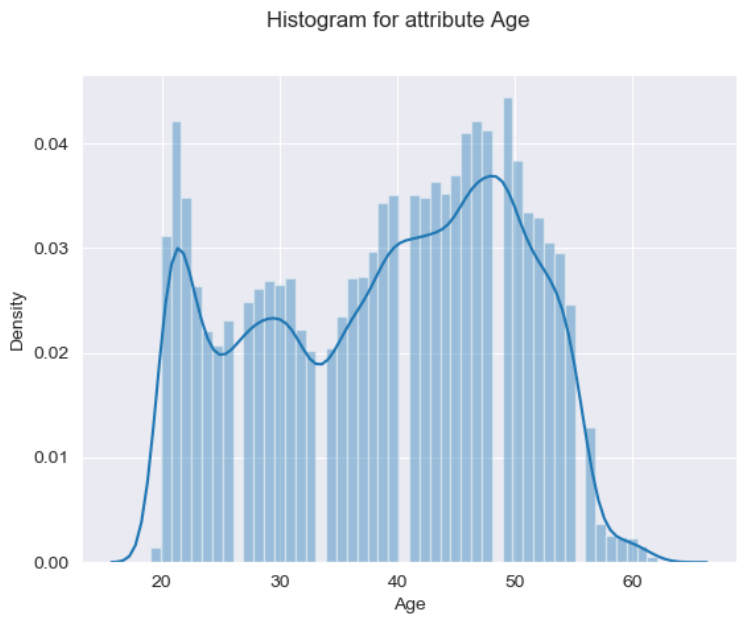


图 11: 特征**年龄**的直方图（UCI 数据集）

超参数分析

内部超参数

- 基分类器的数量。
- 决策树生成算法。(ID3，C4.5，CART 算法)
- 分裂节点时选择特征的比例公式。($M = \sqrt{N}$，$M = \log_2 N$ 和 $M = N$)

外部超参数

- 训练样本的数量。

内部超参数分析

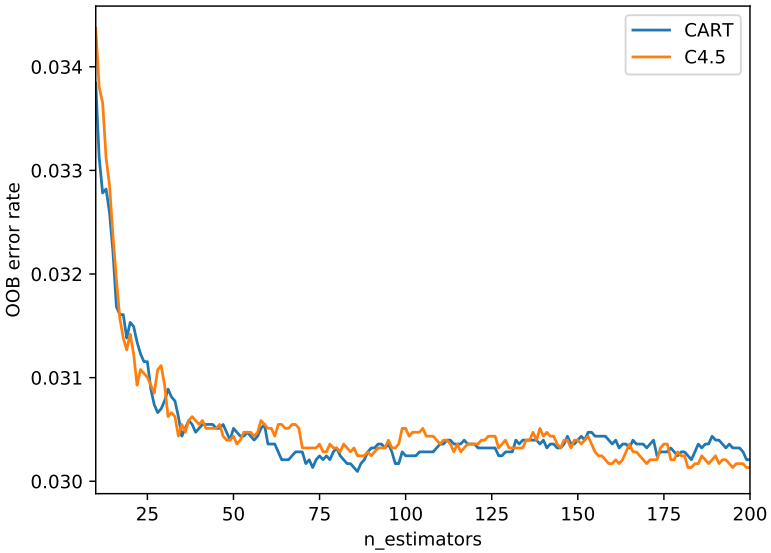


图 12: 基分类器数目和决策树生成算法对 OOB 错误率的影响

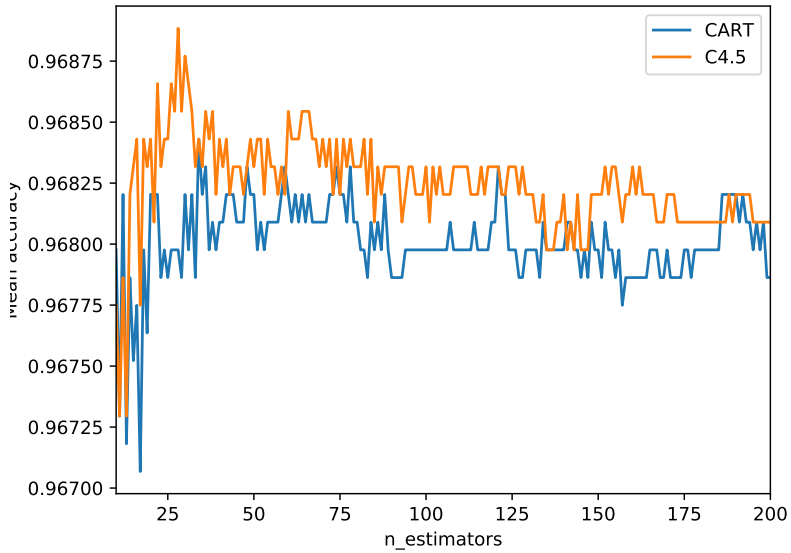


图 13: 基分类器数目和决策树生成算法对预测精度的影响

内部超参数分析

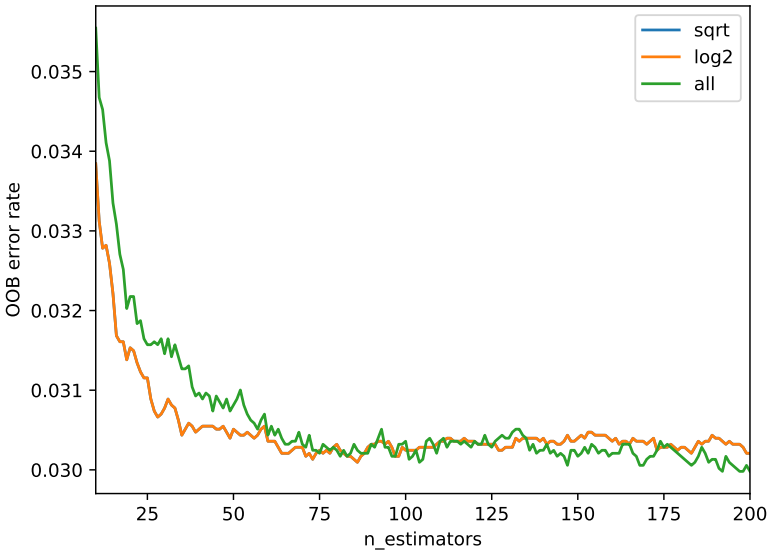


图 14: 基分类器数目和节点分裂时特征选择比例对 OOB 错误率的影响

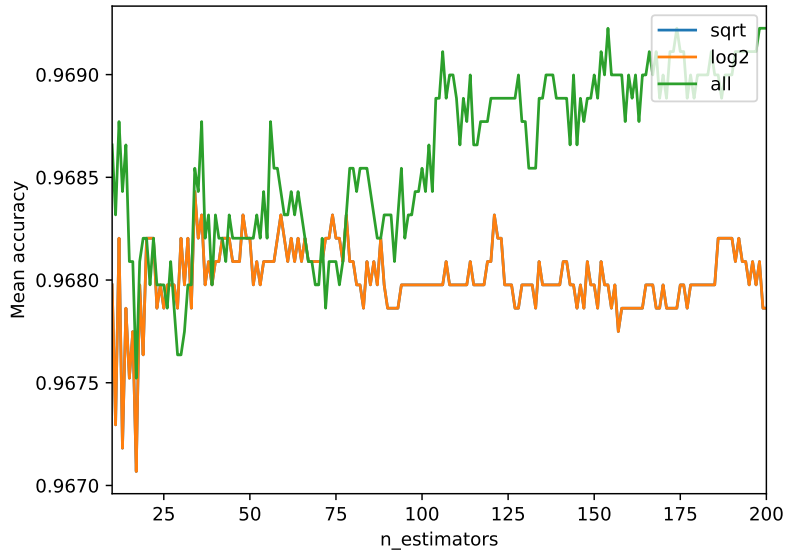


图 15: 基分类器数目和节点分裂时特征选择比例对预测精度的影响

模型性能比较分析

表 3: 随机森林在真实数据集上性能比较（与其他集成学习算法比较）

算法	准确率
Decision Tree	0.929±0.042
Bagging+Decision Tree	0.956±0.039
AdaBoost	0.950±0.032
Gradient Boosting	0.959±0.040
Random Forest	**0.969±0.039**

表 4: 随机森林在真实数据集上性能比较（与其他分类算法比较）

算法	准确率
SVM	0.952±0.042
KNN	0.943±0.004
Neural network	0.941±0.002
SGD	0.952±0.142
Random Forest	**0.969±0.039**

模型性能比较分析

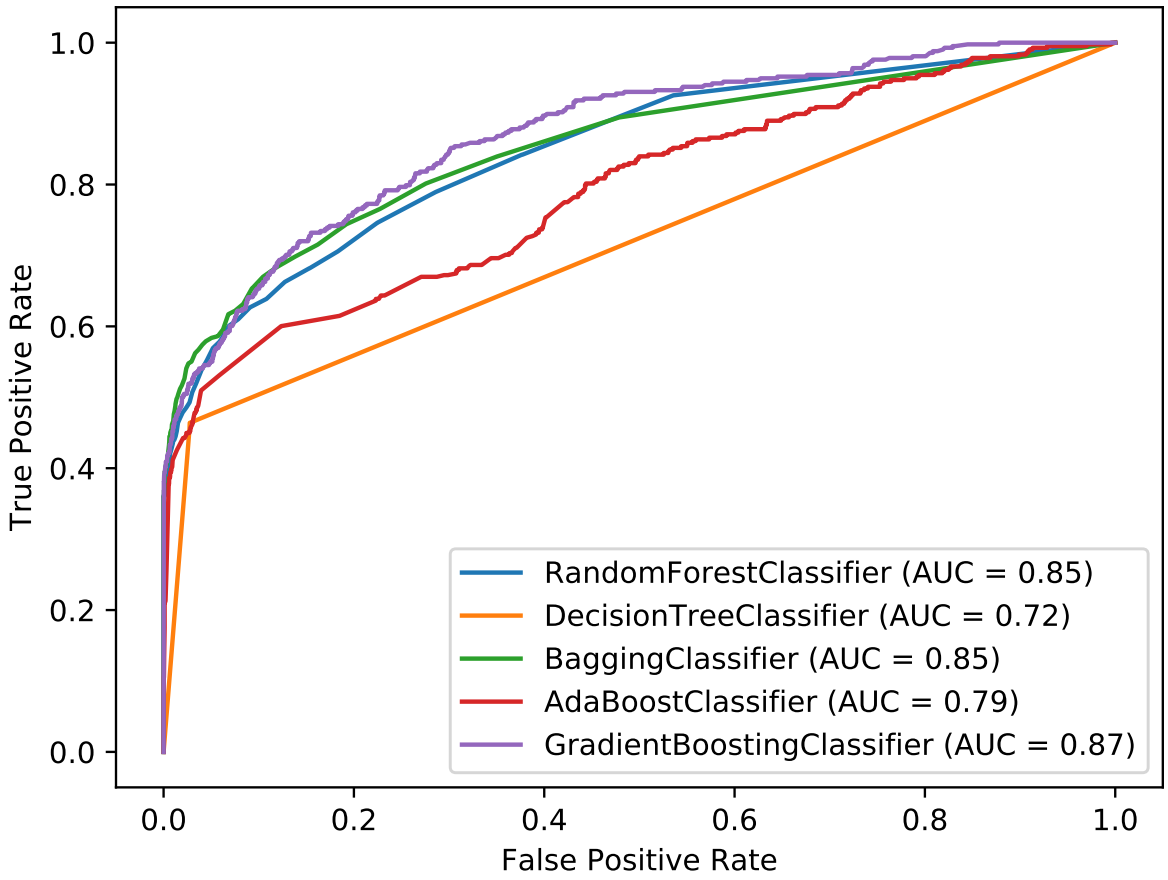


图 16: 表3中对应算法的 ROC 曲线比较

模型性能比较分析

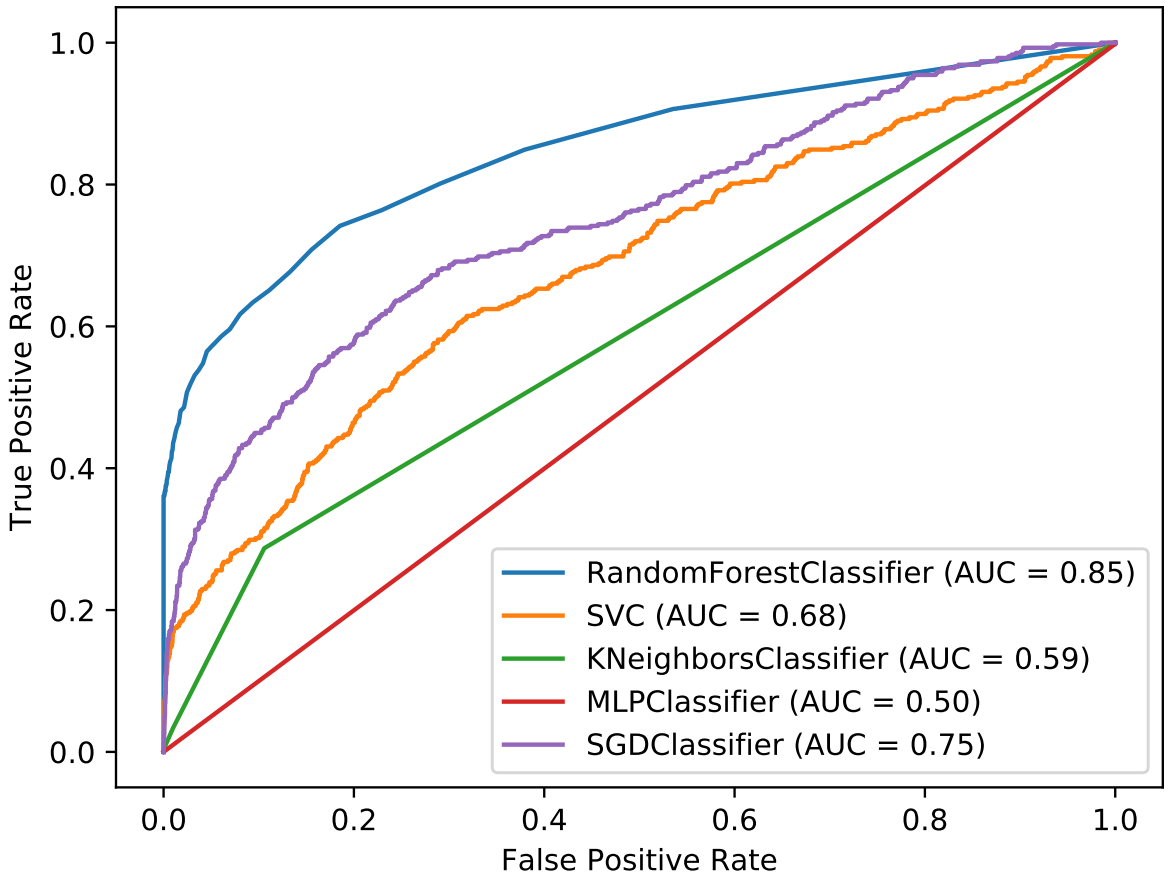


图 17: 表4中对应算法的 ROC 曲线比较

目录

主要贡献

1. 将随机森林算法引入到献血者招募问题当中，给出了一种利用机器学习技术辅助医护人员进行精准招募的方法，提升了招募精度，降低了招募成本。
2. 对当前学术界医疗和机器学习相关技术进行了全面而深入的文献调研工作，总结了近年来医疗 AI 的发展趋势。
3. 改进了三种主流的经典决策树生成算法：ID3、C4.5 和 CART，将递归版本转化成非递归版本，降低了时空开销，为随机森林的算法实现提供了基础。
4. 通过超参数调整，将模型的最终精度提升到了 95% 以上，并进行了详尽全面的模型比较分析。

谢谢！