基于随机森林的献血招募模型

宋子星

东南大学计算机科学与工程学院

May 29, 2020

目录

SOUTHEAST UNIVERSITY

目录

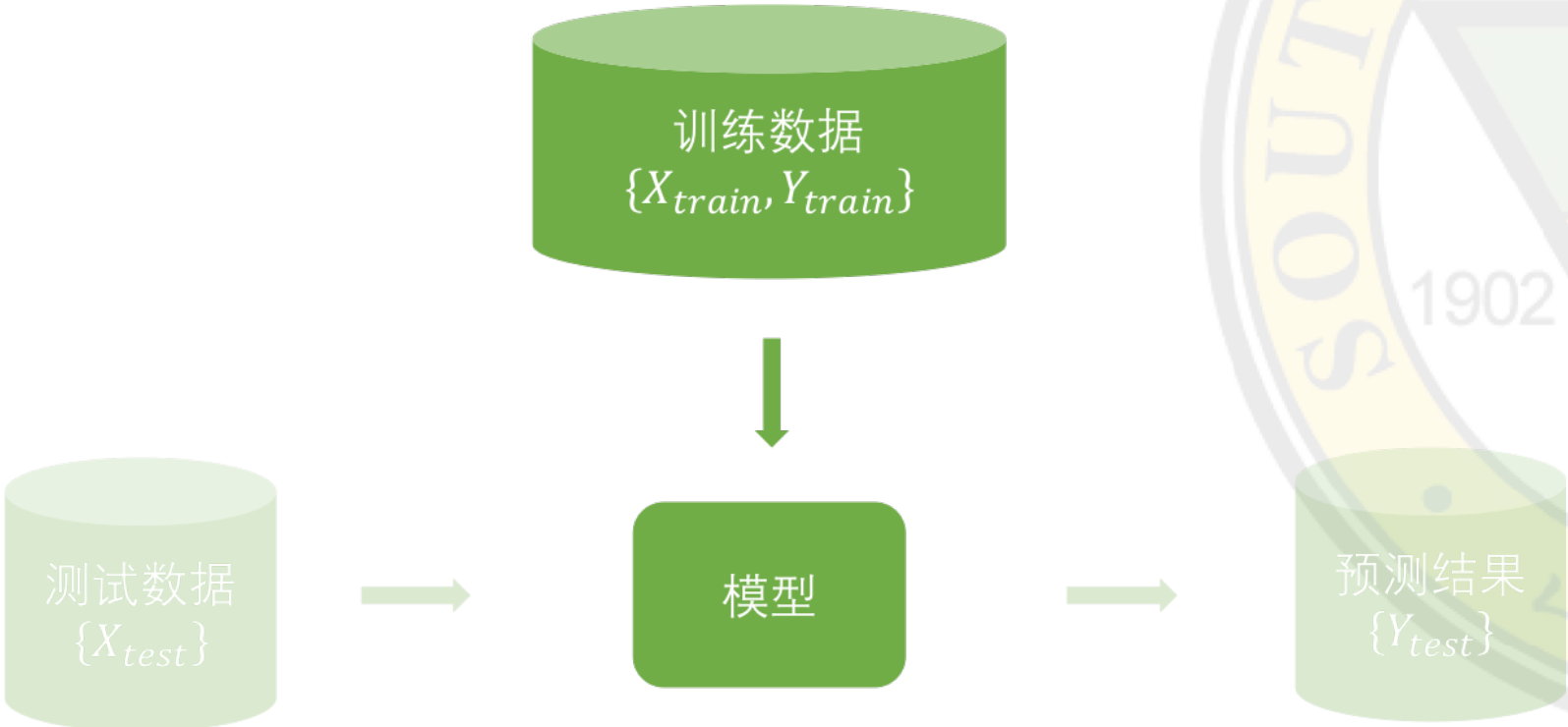
背景

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?

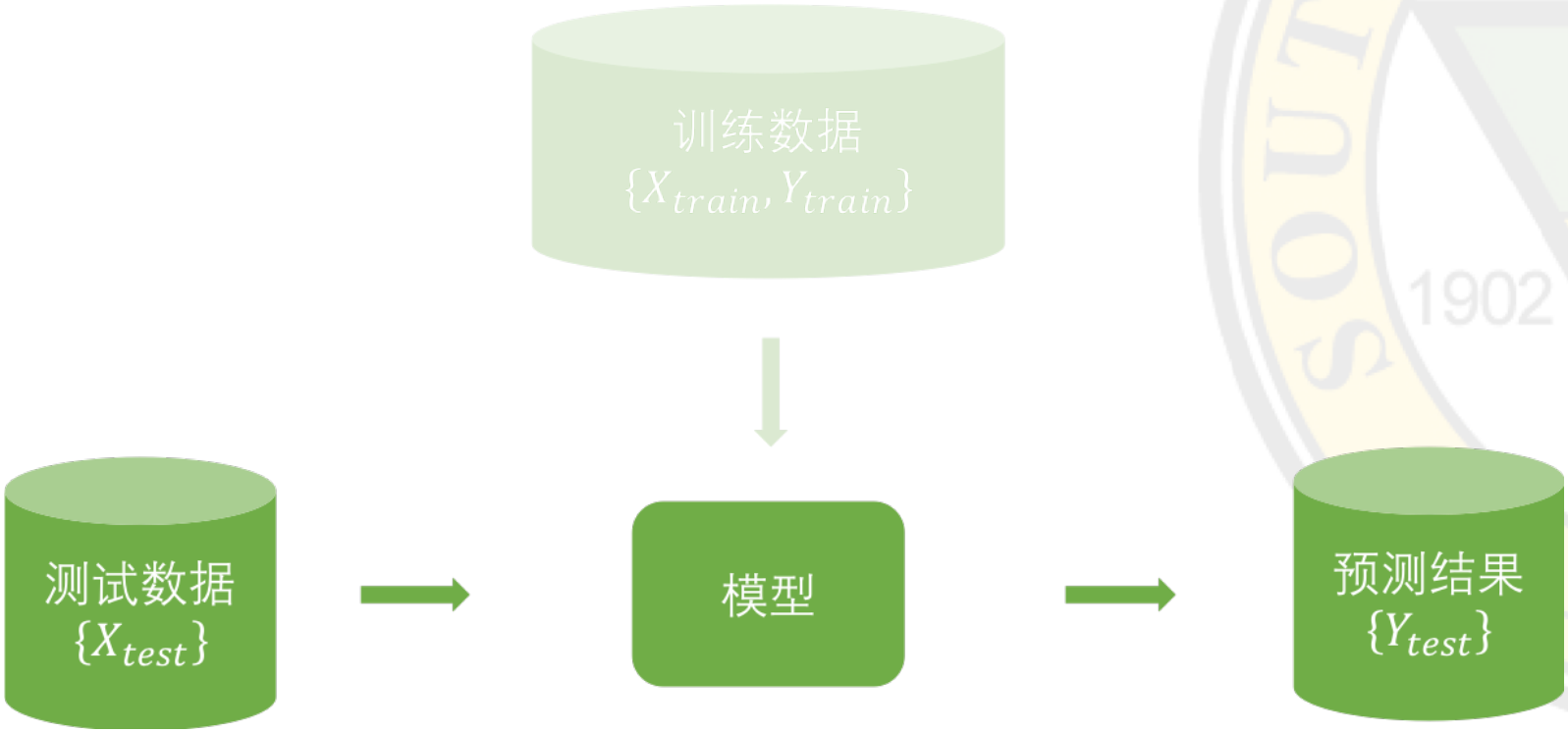
研究问题

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)

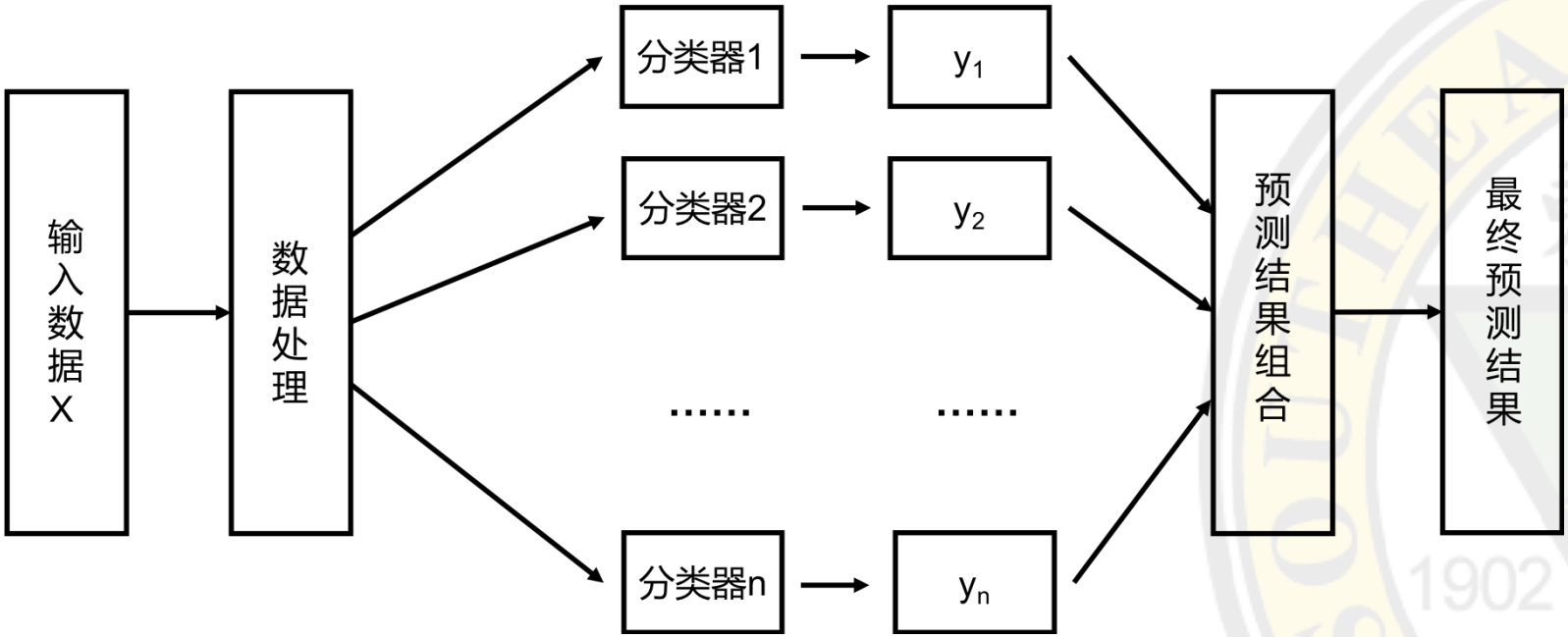


图 1: 集成学习流程框架

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

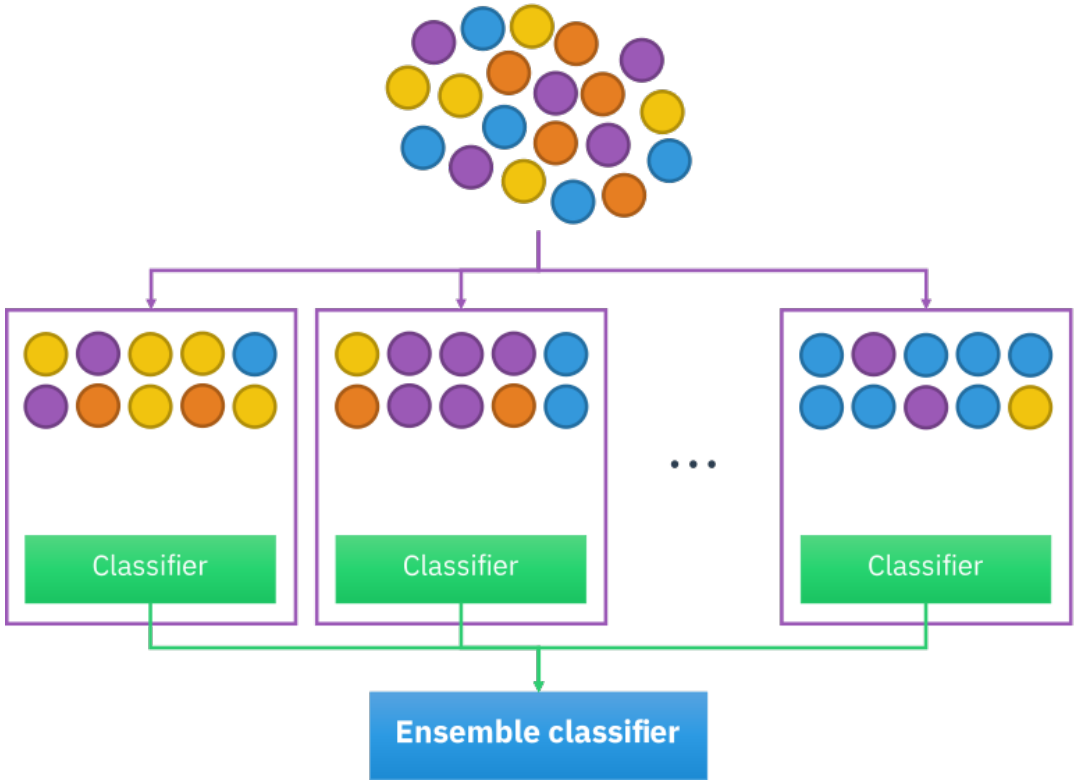


图 2: Bagging 算法

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

决策树 (Desicion Tree)

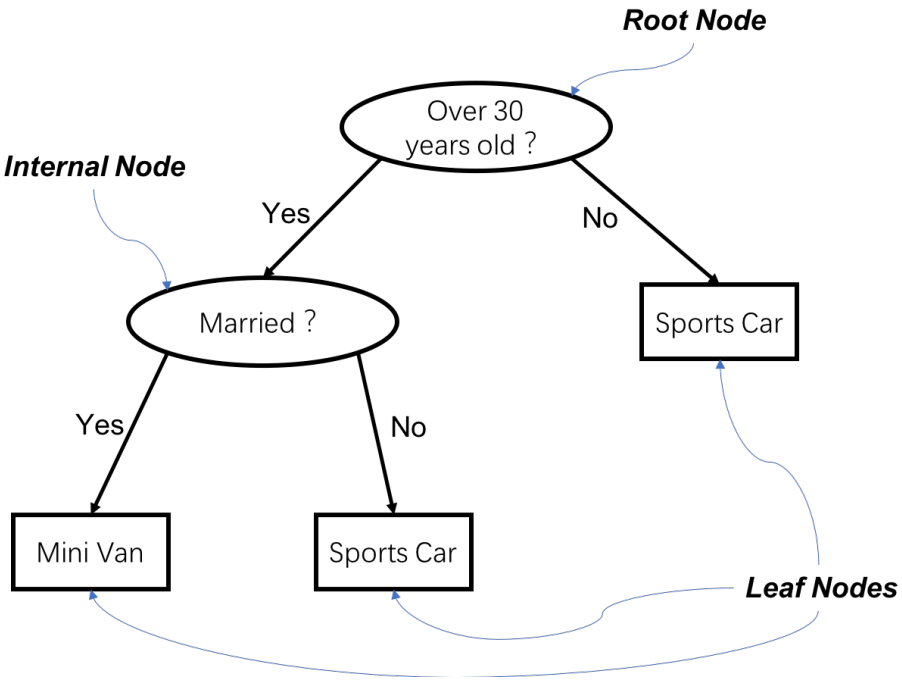


图 3: 决策树案例

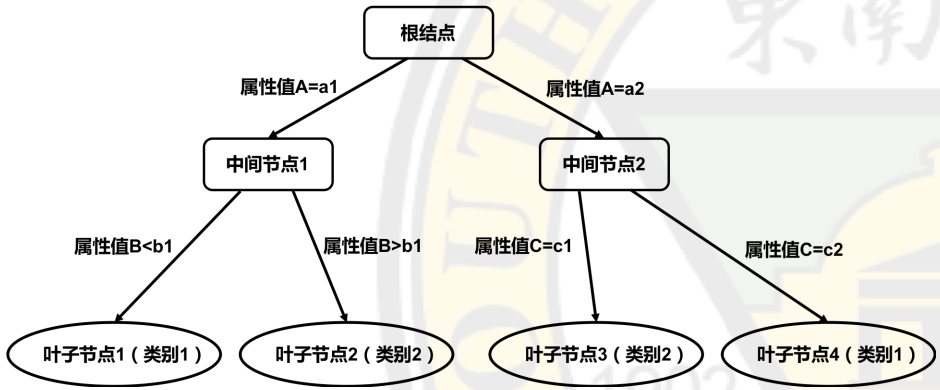
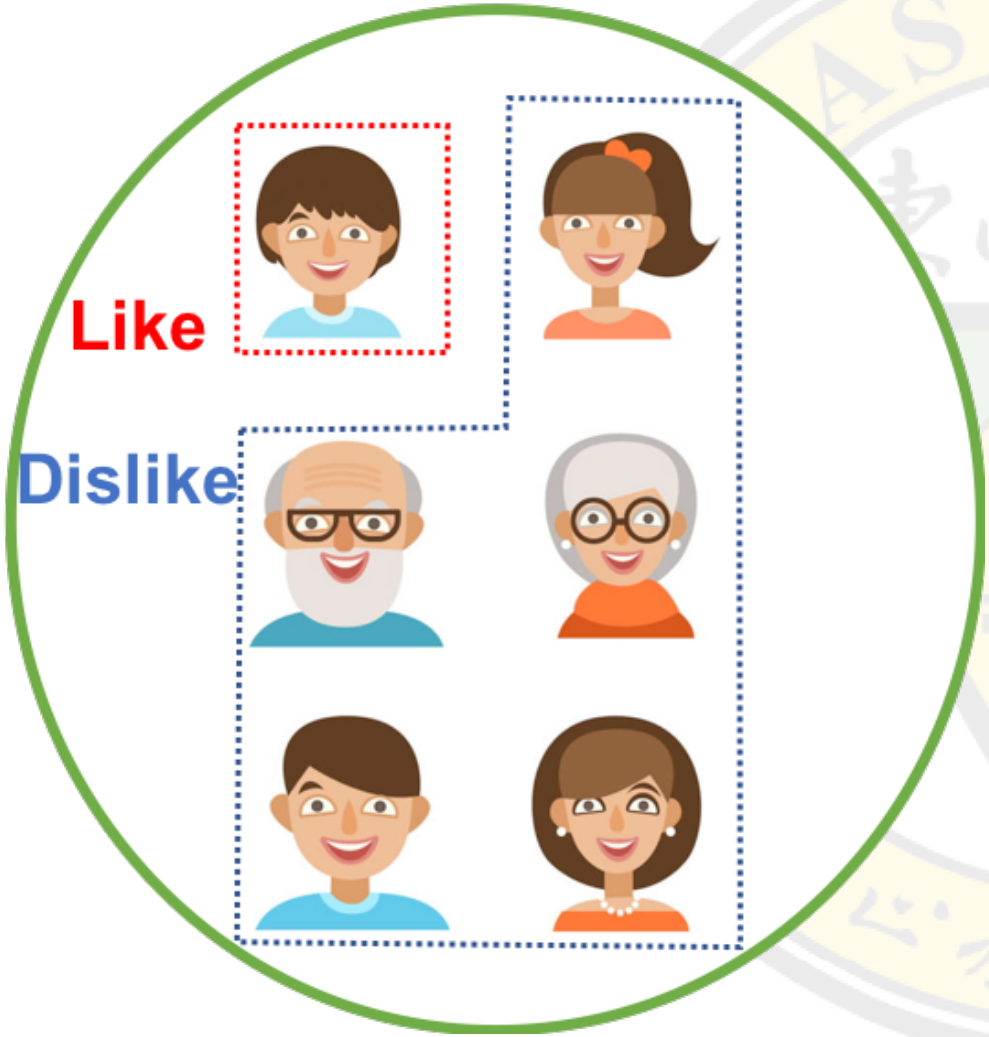


图 4: 决策树归纳案例

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。

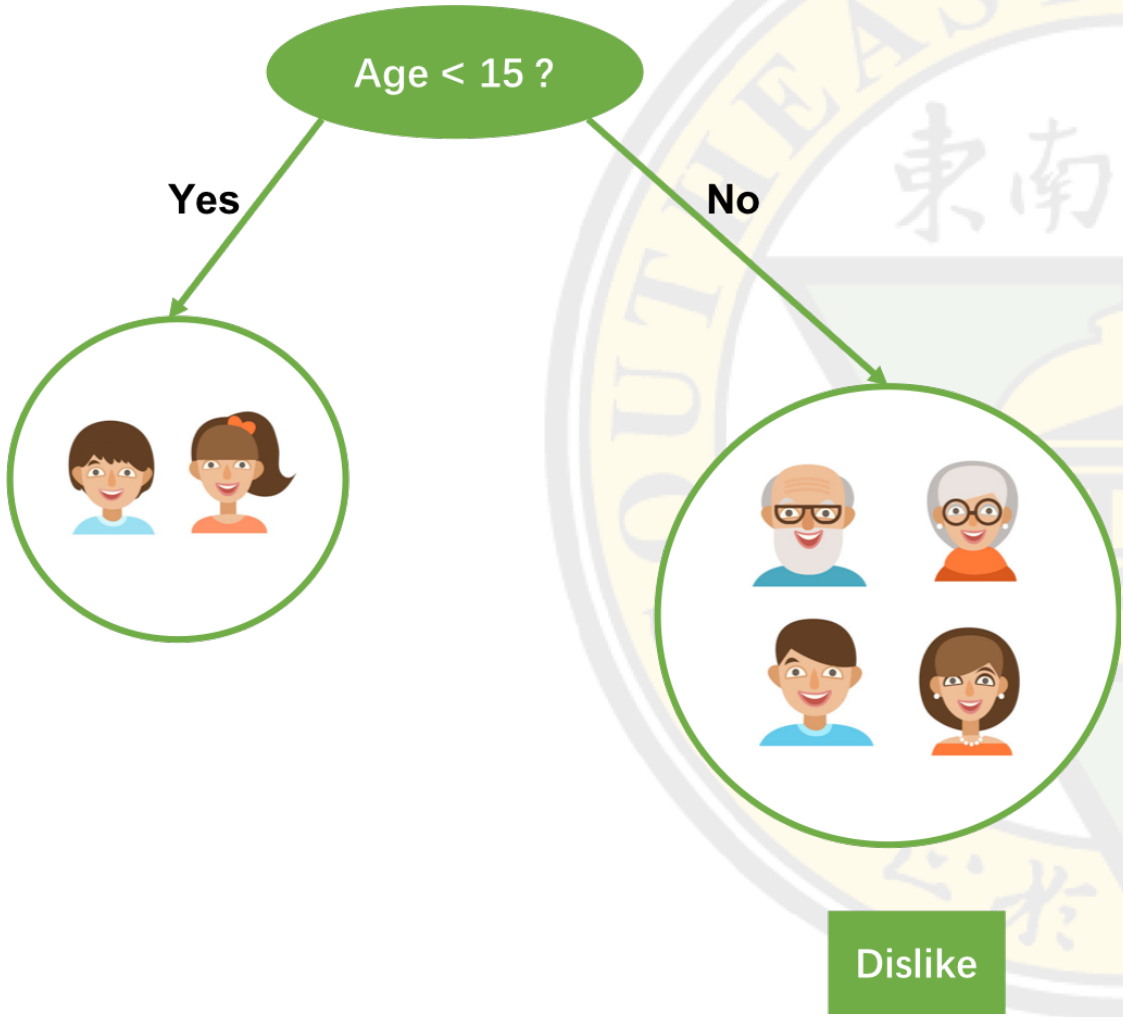
决策树生成算法

- 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。



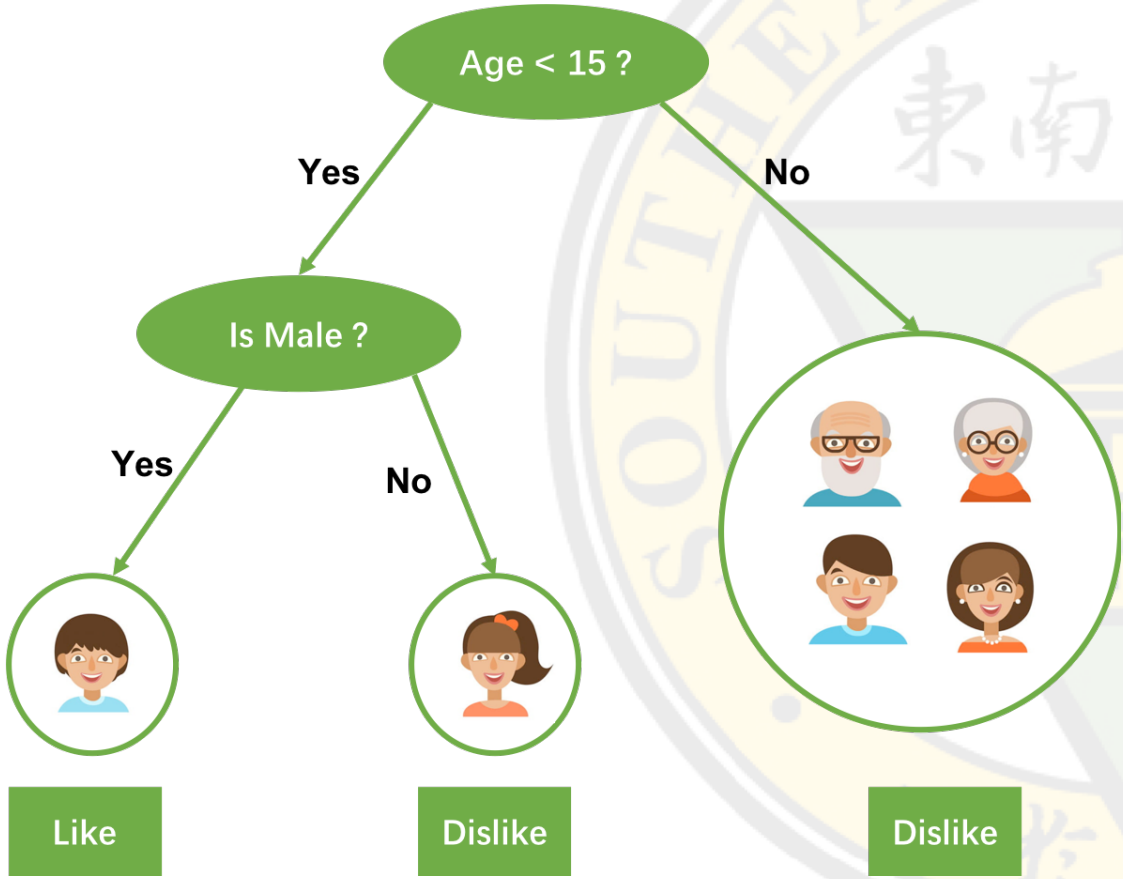
决策树生成算法

- 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)

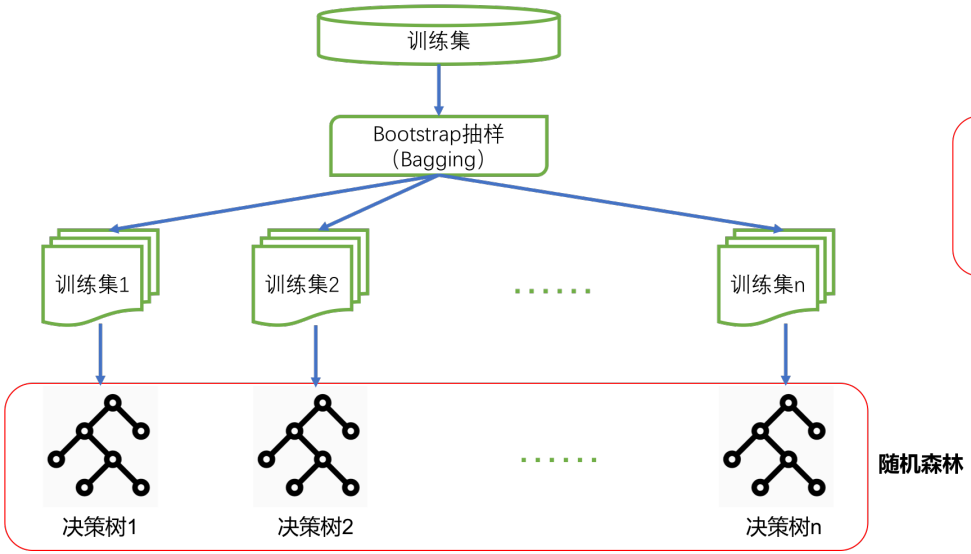


图 5: 随机森林训练过程

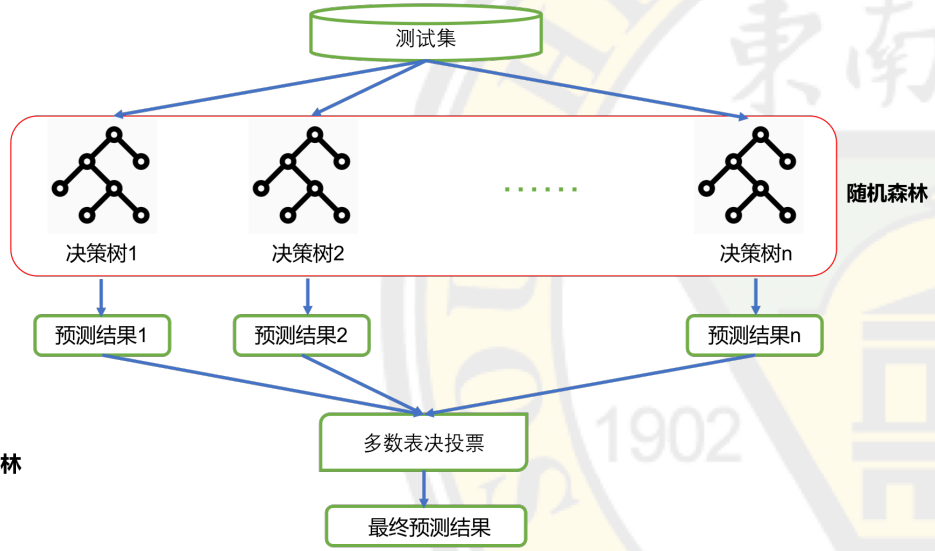


图 6: 随机森林测试过程

目录

爬取公共短消息网关

- 使用 Scrapy 框架爬取公共网关
- 收集 8 个公共短信网关在 14 个月的数据
- 共抓取 386,327 条数据

表 2: 公共网关及抓取的信息数

Site	Messages
receivesmsonline.net	81313
receive-sms-online.info	69389
receive-sms-now.com	63797
hs3x.com	55499
receivesmsonline.com	44640
receivefreesms.com	37485
receive-sms-online.com	27094
e-receivesms.com	7107

消息聚类分析

基本思路

- 使用编辑距离矩阵将类似的消息归于一张连通图中。
- 使用固定值替换感兴趣的消息，如代码、email 地址。
- 查找归一化距离小于阈值的消息，并确定聚类边界。

实现步骤

1. 加载所有消息。
2. 用固定的字符串替换数字、电子邮件和 URL 以预处理消息。
3. 将预处理后的信息按字母排序。
4. 通过使用编辑距离阈值 (0.9) 来确定聚类边界。
5. 手动标记各个聚类，以确定服务提供者、消息类别等。

消息分类结果

- **账户创建确认信息**：向来自服务提供者的用户提供了一个代码，该服务提供者需要在新帐户创建期间进行 SMS 验证。
- **活动确认信息**：向来自服务提供者的用户提供了请求授权进行活动的代码 (例如，付款确认)。
- **一次性密码**：包含用户登录的代码的短信息。
- **用于绑定不同设备的一次性口令**：将消息发送给用户，以绑定一个新的电话号码或启用相应的移动应用程序。
- **重置密码口令**：包含密码重置密码的短信息。
- **其他**：其他未被指定为某种特定功能的消息。

消息分类结果

- 账户创建和移动设备绑定占比最大，占 51.6%
- 一次性密码信息占 7.6%
- 密码重置消息占 1.3%
- 包含“测试”关键词的消息占 0.8%

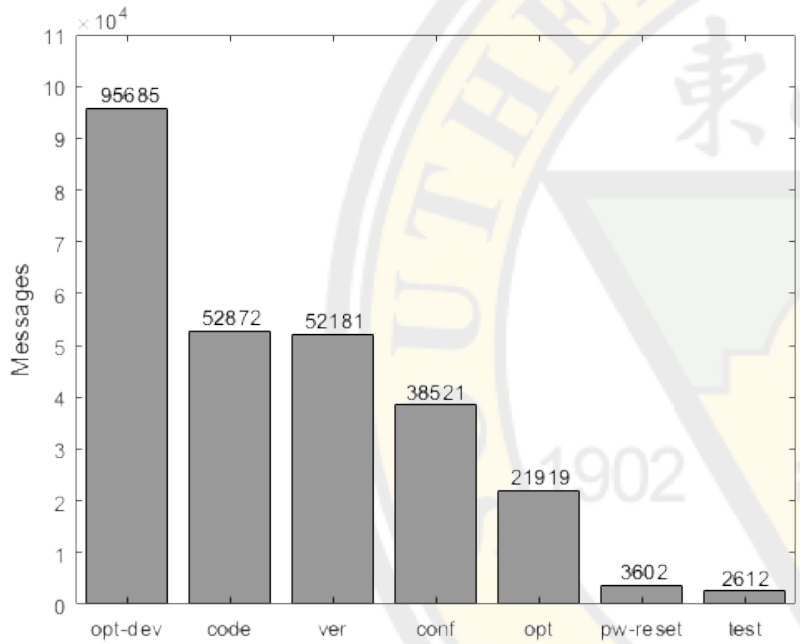


图 7: 消息的聚类

目录

使用 SMS 作为安全信道

PII 和其他敏感信息

- 财务信息
- 用户名和密码
- 重置密码口令
- 其他个人识别信息 (PII)
- 敏感程序的 SMS 活动

使用 SMS 作为安全信道

SMS 编码熵

使用 χ 方检验测试每组编码的熵。χ 方检验是一个零假设的显著性检验，用于测试 SMS 服务的编码是否是从低位到高位均匀分布的。若 p 值小于 0.01，则表明观测值和理想均匀分布之间存在统计学上的显著差异。检验结果表明，65% 的 SMS 服务的编码熵较低，容易被预测和攻击。

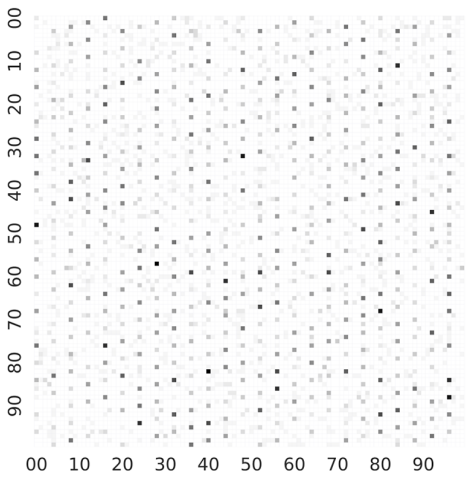


图 8: WeChat

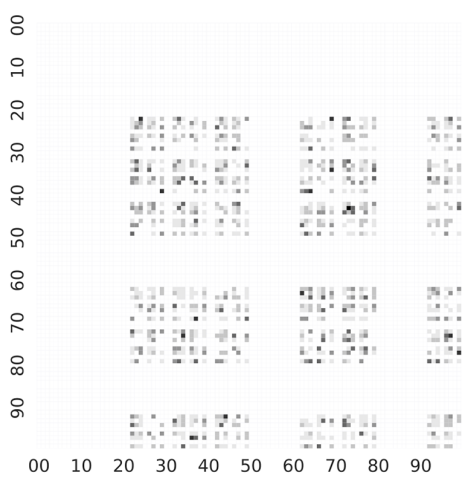


图 9: Talk2

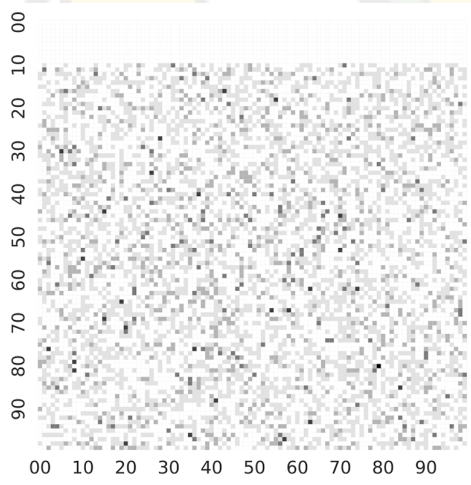


图 10: Google

SMS 的恶意应用

公共网关检测到的恶意信息

- **泄露用户位置信息**：短 URL 可以用于确定消息的源和目的地，即会泄漏用户的位置信息。
- **垃圾邮件宣传广告**：在公共网关服务中比例较低，约为 1.0%。
- **网络钓鱼活动**：试图欺骗用户，使其相信自己正与合法网站通信。

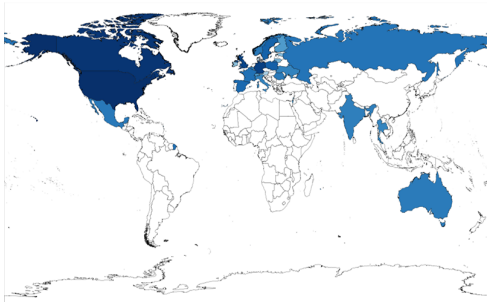


图 11: SMS 地址分布

```
Apple Customer,
Your lost iPhone has been found \
    and temporarily switched ON.
To view iPhone map location
lostandfounds-icloud*com
Apple
```

图 12: 钓鱼短信实例

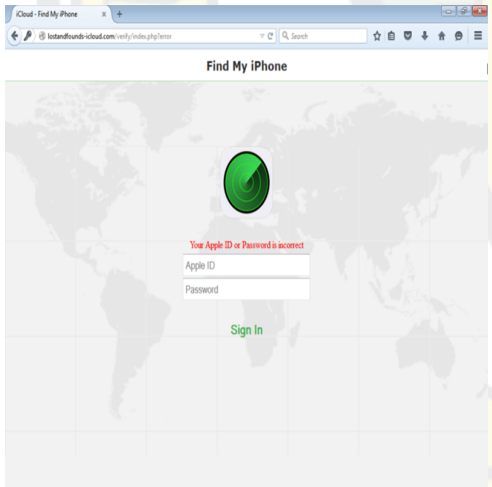


图 13: 钓鱼网站

目录

结论

- SMS 生态系统在智能手机时代出现了新的发展，加入了更多新的设备和参与者。
- 公共网关为用户提供了基于 SMS 的各种安全解决方案。
- 根据该研究，将 SMS 作为安全信道传递敏感信息存在一定的危险性。一些一次性的消息传递机制亟待改进。
- 至于短信滥用，公共网关可以用于规避一些安全性较差的认证机制，或进行 PVA 欺诈行为。

Thanks for Listening.

LaTeX Beamer template opensource on Github now!

https://github.com/wurahara/SEU-Beamer-Slide

Welcome Star and Fork.