阿里巴巴大数据竞赛

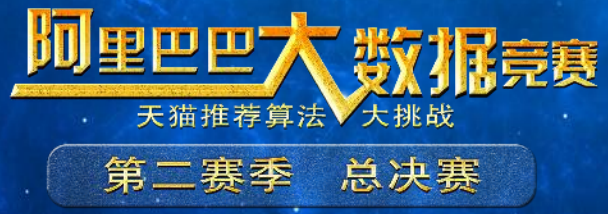
天猫推荐算法大挑战

第二赛季 总决赛

大数据，小细节

Fly402 团队

参赛历程



思路

预测购买

未交互 → 协同过滤、关联规则、隐语义，矩阵分解……

交互 → 逻辑回归、贝叶斯、SVM、随机森林，GBRT……

交互

核心要点

阿里巴巴大数据竞赛

天猫推荐算法大挑战

第二赛季　总决赛

- 特征选择
- 模型选择
- 训练
- 评价
- 融合

特征选择

样本形式：

$User_i$	$Brand_x$			

总分类 \ 时间维度	7天	15天	30天	60天	90天
$User_i$					
$Brand_x$					
$User_i$-$Brand_x$					

类型维度：点击　购买　收藏　购物车

特征选择

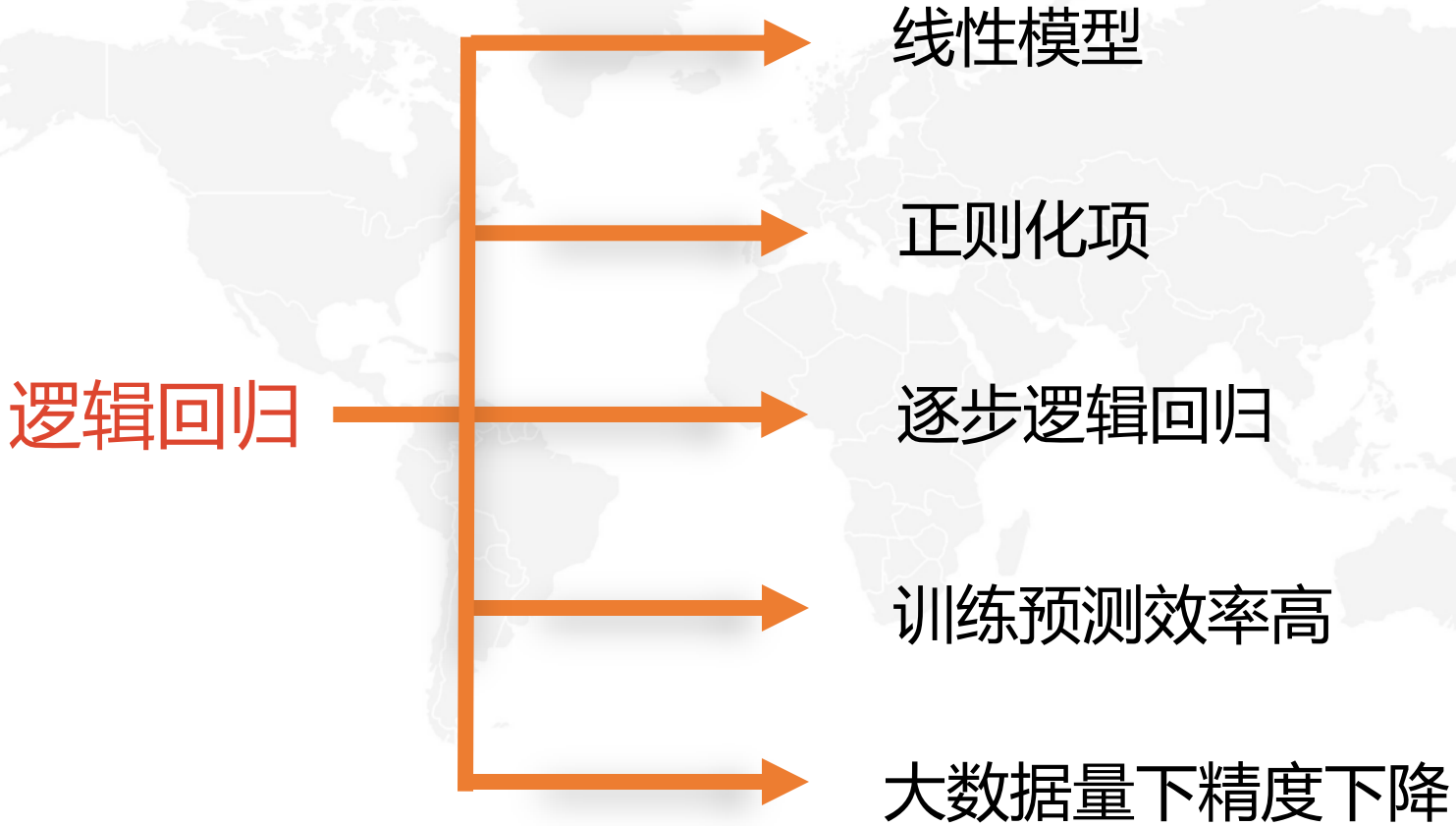

业务指标
数据分析
特征提取
品牌购买价值
购买该品牌的用户数
b_bought_u_count
品牌被购买次数
b_bought_times
回头客的数量
b_rebought_u_count
平均每个回头客的回头次数
b_avg_rebought_days
回头客数量/购买的用户数
b_rebought_ratio

特征选择

- 比值类特征
  - 部分基本特征相除
- 分类训练预测的启发
  - u买过b跟u没买过b的特征有区别
- 均值、方差等
  - 体现特征值的平均水平与波动情况
- 策略
  - Season1的策略转换为特征

- 特征选择
- 模型选择
- 训练
- 评价
- 融合

模型选择

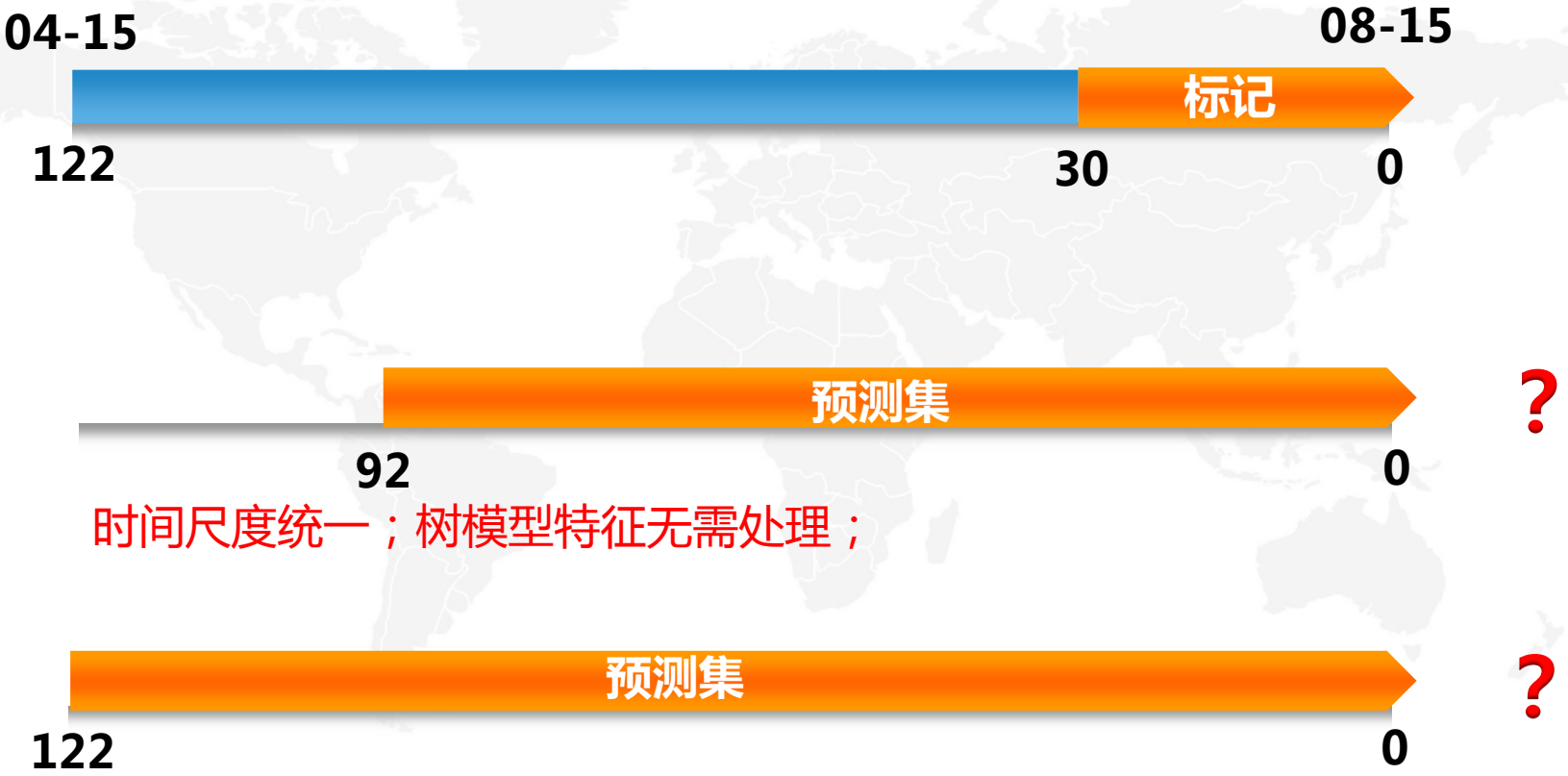


模型选择

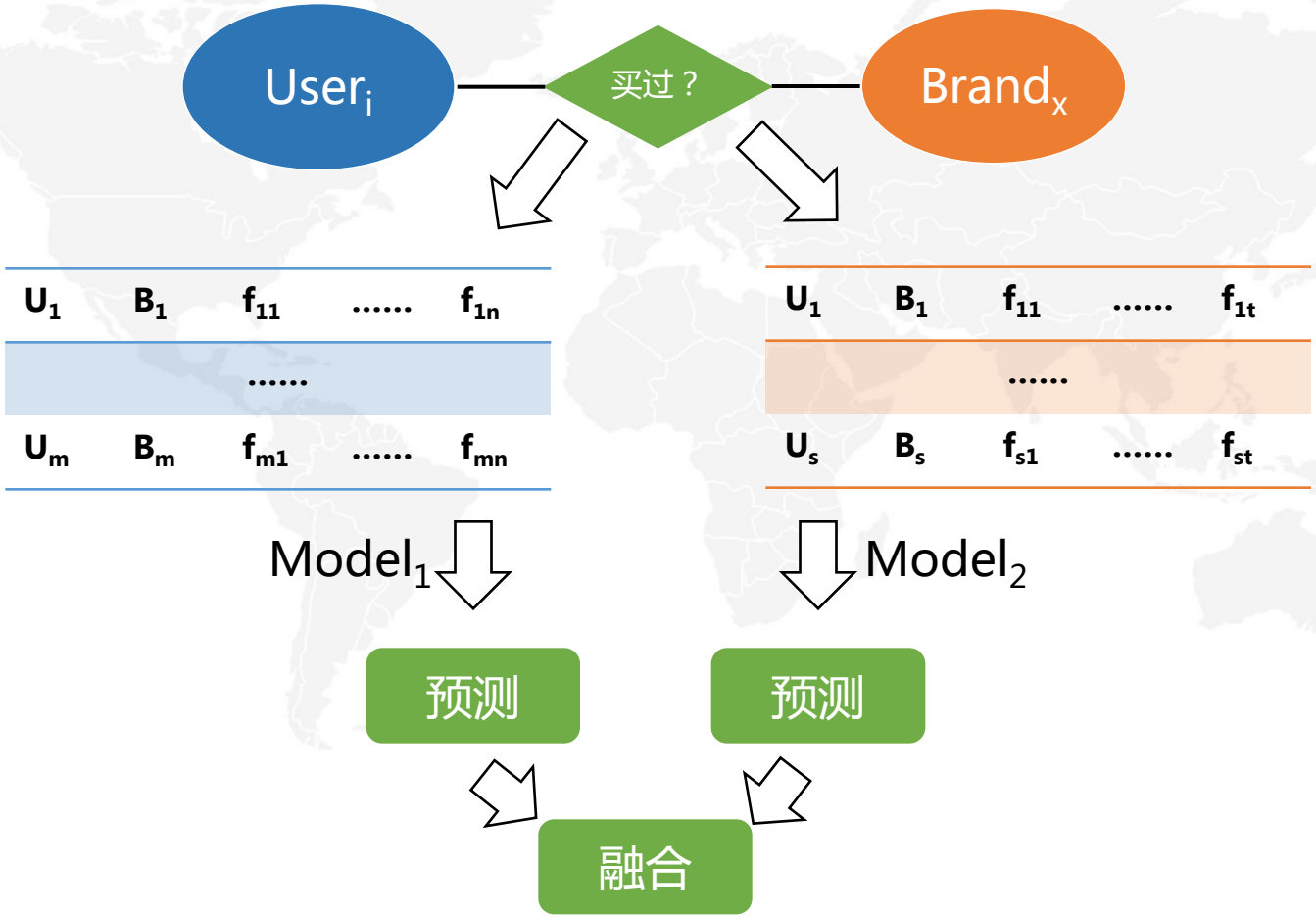
模型选择

- 特征选择
- 模型选择
- 训练
- 评价
- 融合

训练集构造

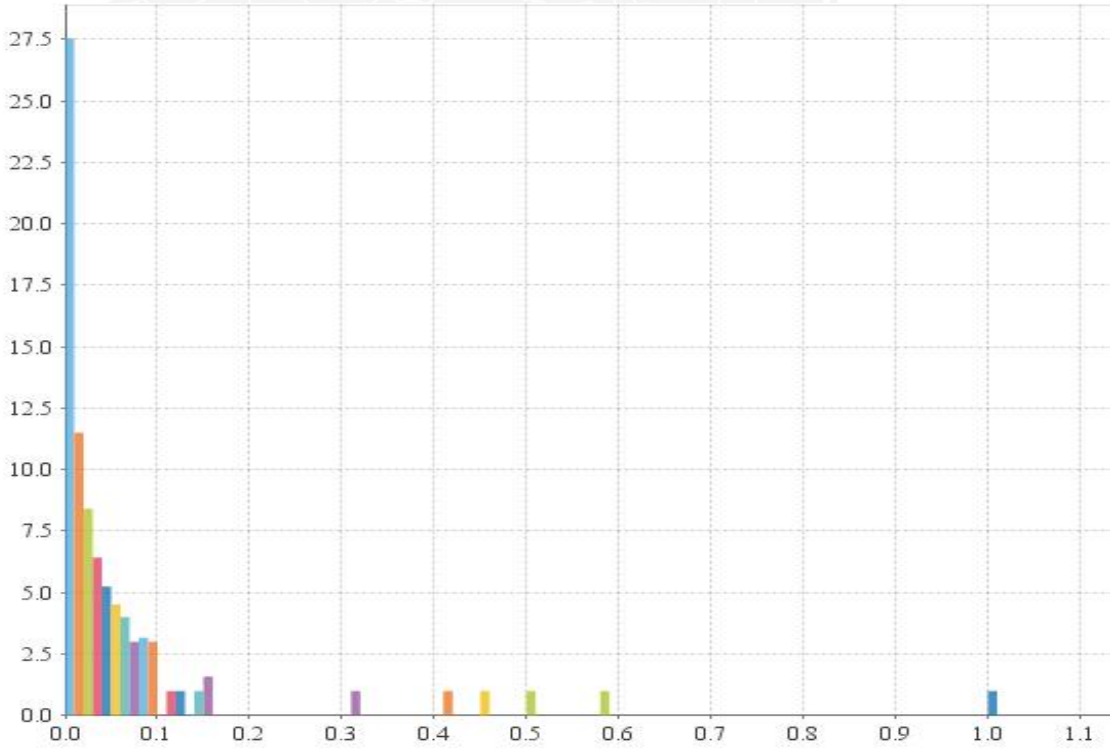

04-15
08-15
标记
122
30
0
预测集
?
92
0
时间尺度统一；树模型特征无需处理；
预测集
?
122
0
信息量更足；更多的用户品牌对；时间尺度不统一，需要特殊处理

分类训练预测



数据处理

- 归一化
  - $(X-\min)/(\max-\min)$
- 分箱
  - 等量、等距
- 去噪



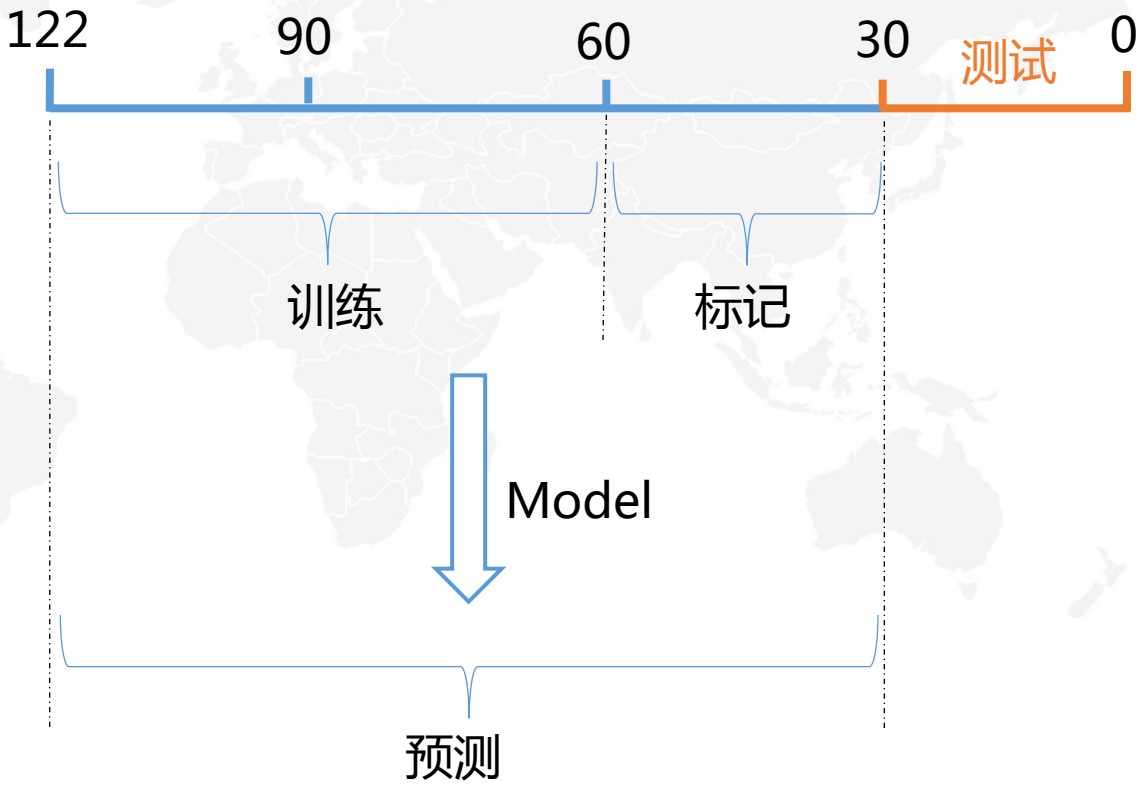
- 特征选择
- 模型选择
- 训练
- 评价
- 融合

评价

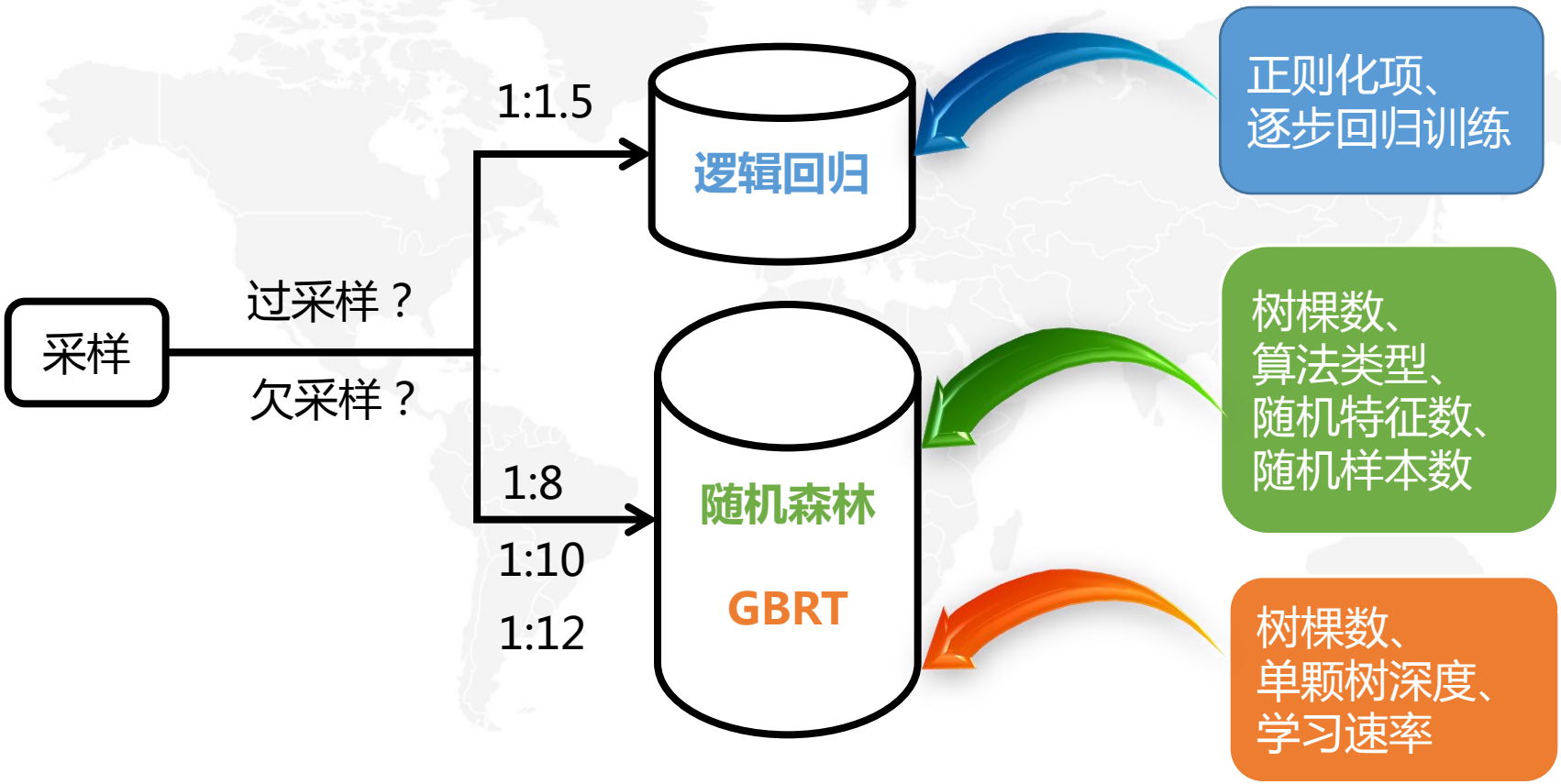
1

特征分布一致

2



采样与调参


采样
过采样？
欠采样？
1:1.5
逻辑回归
1:8
1:10
1:12
随机森林
GBRT
正则化项、
逐步回归训练
树棵数、
算法类型、
随机特征数、
随机样本数
树棵数、
单颗树深度、
学习速率

特征选择

- 逐步回归
- 相关矩阵
- 属性选择
  - 信息增益、基尼指数、信息增益率
- 编写脚本
  - 模型反馈
- 分批迭代测试

过拟合

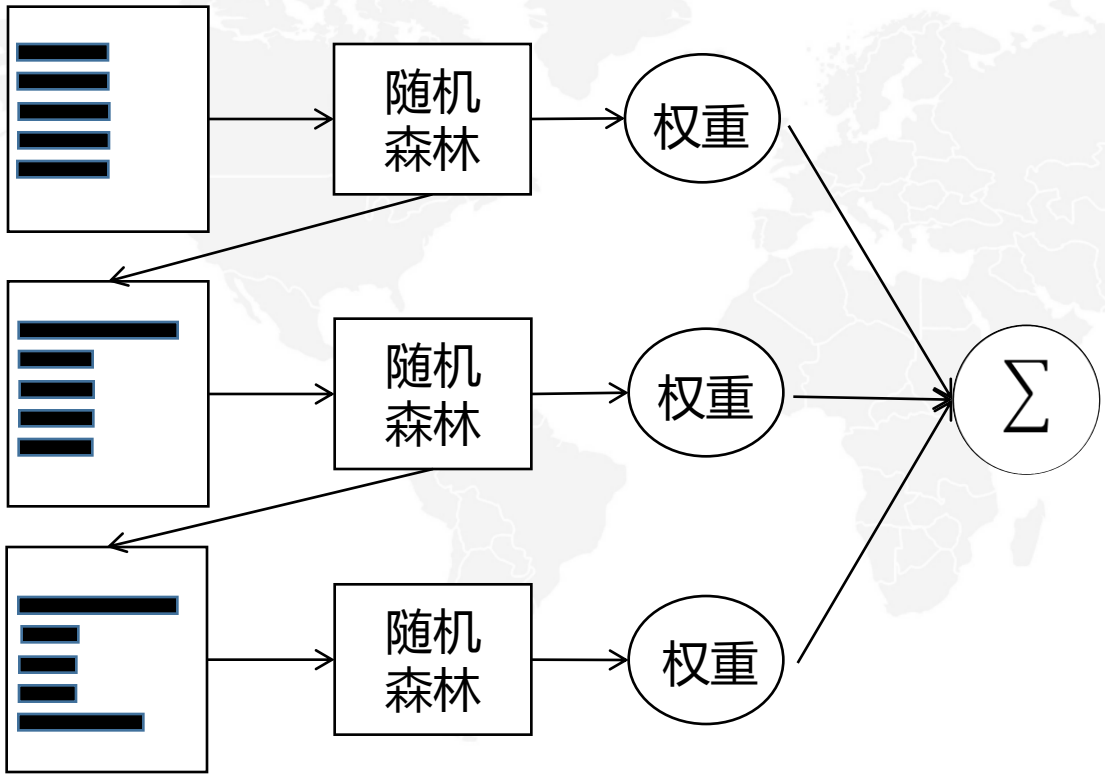
- 健壮的本地测试
- 鲁棒性强的特征体系
- 数据分布
- 模型参数

- 特征选择
- 模型选择
- 训练
- 评价
- 融合

简单融合

- 不同时间尺度融合
  - 4个月+3个月+2个月
- 同模型融合
  - 不同采样比例、不同参数
- 不同模型融合
  - 逻辑回归、随机森林、GBRT
- 分开与全集

复杂融合



没效果原因：

- 随机森林是强分类器
- 迭代次数不够
- 错误率、权重调整不当

改进

- 使用弱分类器：LR或者参数简单的GBRT
- 不用权重列，使用有放回加权采样

谢谢！

阿里巴巴大数据竞赛

天猫推荐算法大挑战

第二赛季　总决赛