新浪微博互动预测大赛答辩

2015天池大数据竞赛

TIANCHI天池

SeaSide

李邦鹏 钊魁

浙江大学

提纲

- 问题描述
- 模型建立
- 特征工程
- 总结展望

第1部分

问题描述

问题描述

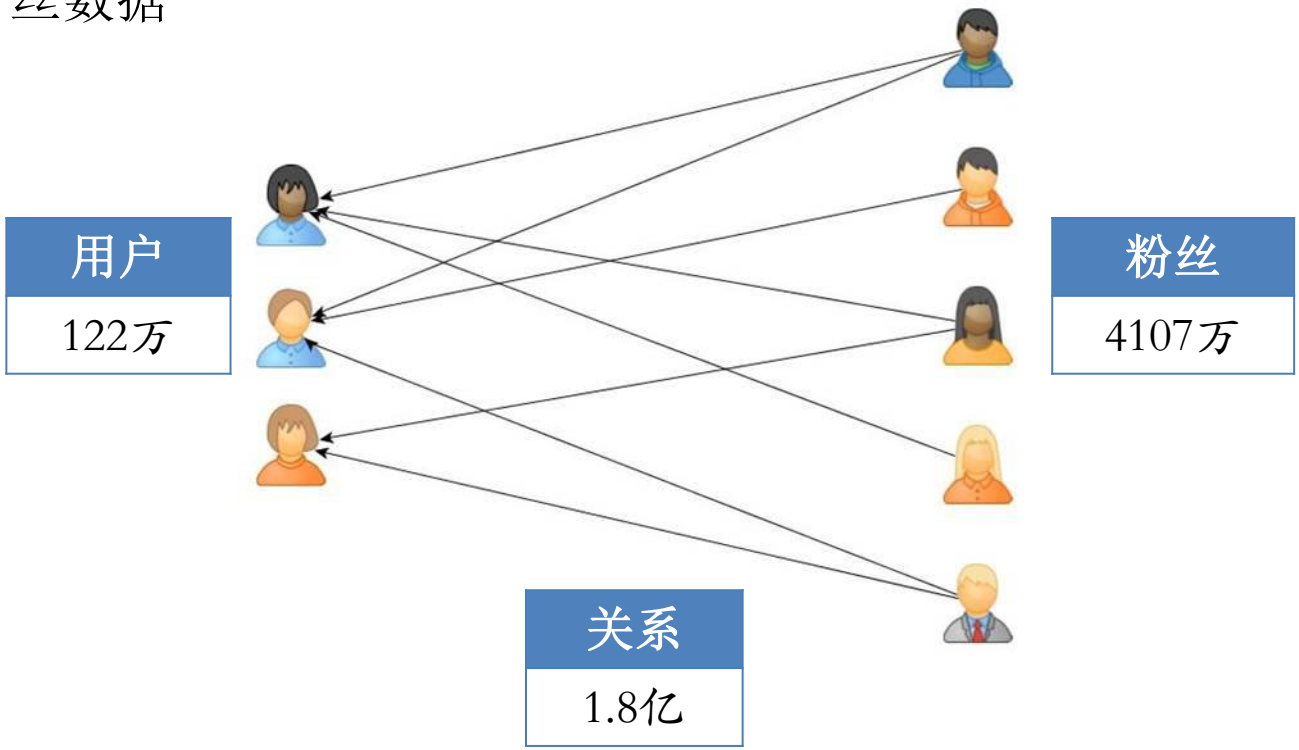
➢ 数据说明

- 博文数据

字段	训练	预测
用户ID	197万	106万
博文ID	1.1亿	1842万
发布时间	从2014-11-01 到2015-03-31	从2015-04-01 到2015-04-30
微博内容	中文明文数据 话题覆盖广	内容演化 有新话题出现

问题描述

- 数据说明
  - 粉丝数据

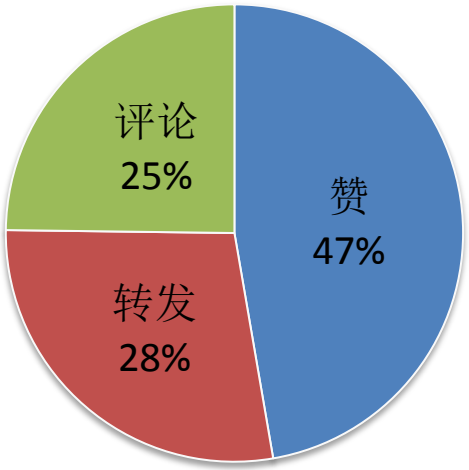


问题描述

➢ 数据说明

- 行为数据

字段	说明
行为用户ID	1780万（45%）
行为时间	发布24小时内
用户ID	59万（30%）
微博ID	1407万（13%）
行为标记	转发、评论、赞



问题描述

- 目标
  - 快速寻找有价值的微博，增加其曝光量
  - 评估指标

档位	互动数	权重
1	0-5	1
2	6-10	10
3	11-50	50
4	51-100	100
5	100+	200

$$p = \frac{\sum_{i=1}^{5}(w_i \times cr_i)}{\sum_{i=1}^{5}(w_i \times c_i)}$$

第2部分

模型建立

模型建立

- 方案1
  - 二分类：用户u与微博m产生交互的概率$p(u,m)$
  - 预测微博m的交互次数为 $\sum_{i=1}^{N} p(u_i,m)$
  - 训练数据 2.0×10^{15}，预测数据3.3×10^{14}
  - 正负样本比例 $1:7.5\times10^{6}$
  - 计算规模过大，模式稀疏
  - 不具可行性

模型建立

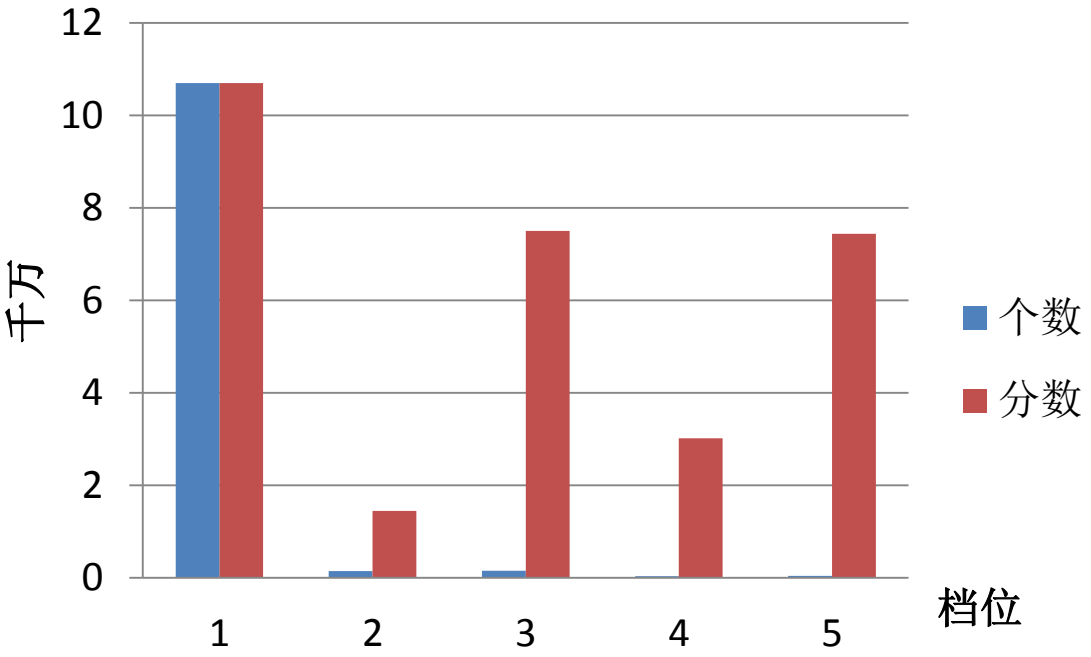
- 方案2
  - 四次二分类：微博m总交互是否大于5/10/50/100
  - Boosting方法组合四个分类结果
  - 训练数据 1.2×10^8，预测数据 2.0×10^7
  - 除1档外，数据分布较均匀
  - 计算规模大，执行成本高
  - 模型复杂，难以调优与部署

模型建立

- 方案3
  - 五分类：微博m总交互次数所在档位
  - 训练数据 1.1×10^8，预测数据 1.8×10^7
  - 除1档外，数据分布较均匀
  - 计算规模较大
  - 模型简单，有优化空间

模型建立

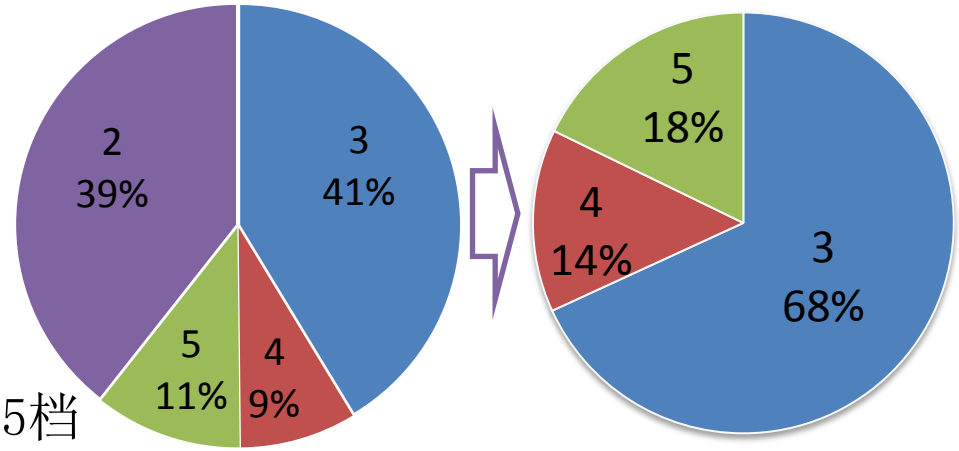
- 方案3.1
  - 训练数据中五档的分布情况



模型建立

➢ 方案3.1

- 简易方法分离第1档
  - ✓ 发表微博m的用户u历史平均被交互次数小于1.95
  - ✓ 极大地降低计算量
  - ✓ 消除数据不平衡问题
- 舍弃第2档
  - ✓ 难以区分，所占分数少
  - ✓ 进一步降低计算量
  - ✓ 提高对3/4/5档的识别
- 训练用历史数据中3/4/5档
- 训练数据217万（2.0%），需模型预测的数据155万（8.6%）



模型建立

- 方案3.2
  - 时间局部性
    - ✓ 用户、内容与社交网络等在演化
    - ✓ 与更近的历史更相似
    - ✓ 训练数据不是越多越好
  - 用最近3个月历史数据
    - ✓ 2015-01-01到2015-03-31
    - ✓ 计算量更低，效果更好
  - 训练数据136万（1.2%）
  - 最终方案
    - ✓ 简易分离第1档+舍弃第2档+三分类
    - ✓ 带权重的随机森林

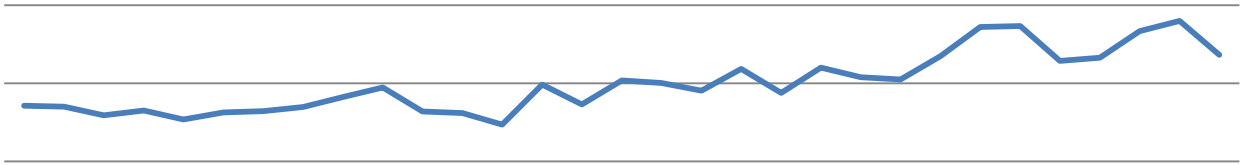
第3部分

特征工程

用户特征

➢ 互动特征

- 历史被转发、评论、赞与互动的均值、方差、变异系数
- 历史中不同互动级别微博所占比例
- 历史平均被交互所在等级
- 历史被交互变化趋势

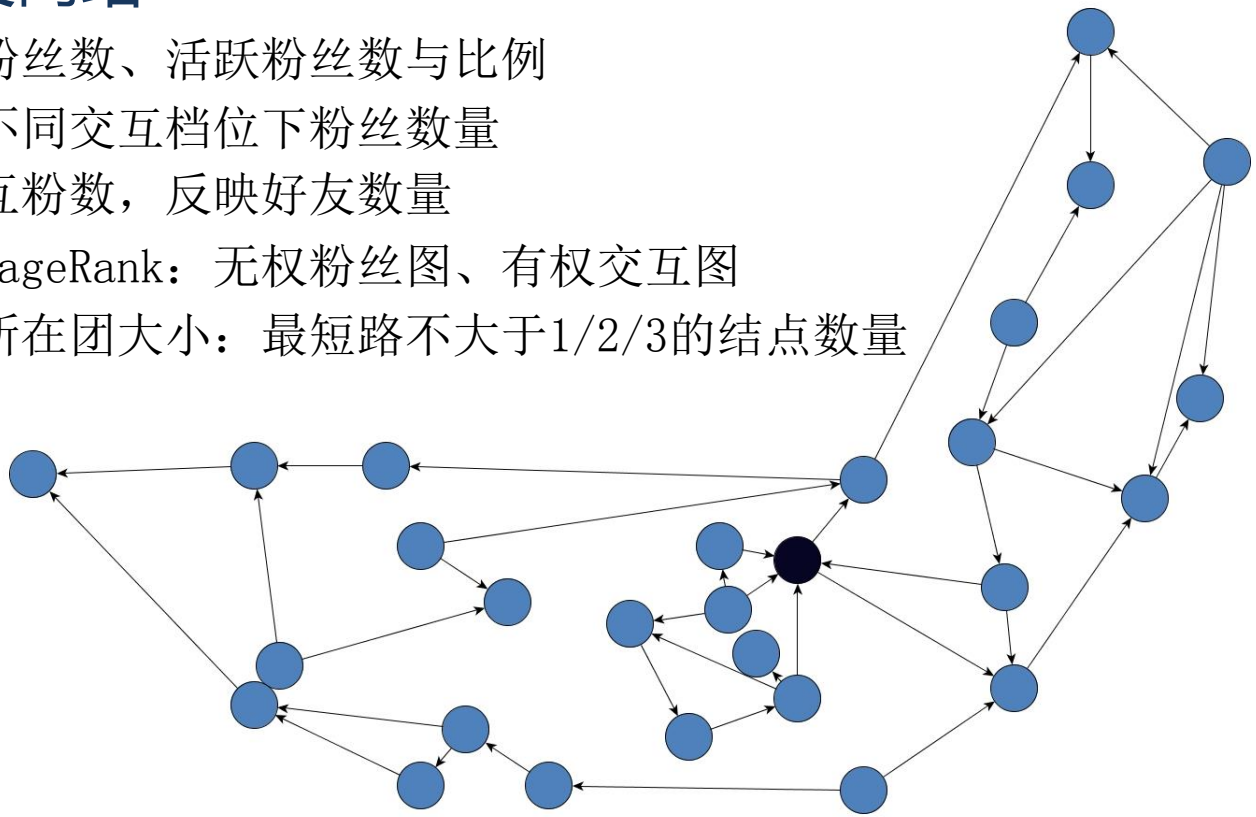


➢ 行为特征

- 转发、评论、赞与互动的均值、方差、变异系数
- 关注了多少人

用户特征

- 社交网络
  - 粉丝数、活跃粉丝数与比例
  - 不同交互档位下粉丝数量
  - 互粉数，反映好友数量
  - PageRank：无权粉丝图、有权交互图
  - 所在团大小：最短路不大于1/2/3的结点数量



微博特征

- 情感分析
  - 情感极性词典NTUSD，表情情感列表
  - 正、负面词汇与表情的个数、比例
- 热度分析
  - #话题#、 AT对象、URL
    - ✓ 历史平均出现次数
    - ✓ 截止微博m时出现次数
  - Topic Model
    - ✓ 句子由隐式主题构成
    - ✓ 历史平均出现次数
    - ✓ 截止微博m时出现次数
  - 不要引入未来信息

微博特征

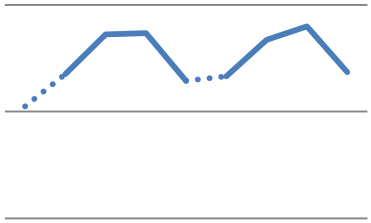
- 内容分析
  - 是否包含@、#、？、！、【、转发、赞等
  - 词性比例：名称、动词、形容词与标点符号等
  - 微博Topic有多少潜在粉丝感兴趣
- 反垃圾
  - #话题#、AT、URL、Topic的黑名单指数
    - ✓ 使用次数、交互次数、两者比值
  - 微博长度异常
    - ✓ 用户历史微博平均长度，这条微博长度
  - 以AT开头是对话
    - ✓ 交互规模有限

用户微博特征

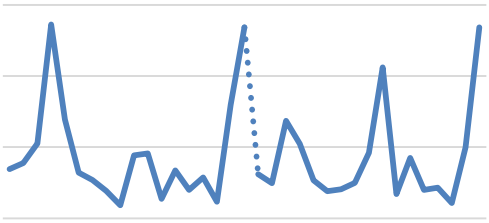
➢ 时间特征

- 微博m发表时间与该用户第一条微博时间差
- 微博m发表的时间点有多少潜在交互粉丝
- 微博m发表时是星期几，是几点

星期



小时



➢ 反异常

- 用户平均每天发微博数量，截止微博m用户发表了多少微博
- 用户发布微博m后5秒内是否发布新微博

特征选择

- 皮尔森系数
  - 反映两个变量间的相关性
  - 计算每个特征与档位之间的相关系数
  - 去除相关系数绝对值很小和符号与预期不符的特征
- 筛选
  - 特征筛选前95维，筛选后72维
  - 成绩提升0.3%
  - 特征不是越多越好
  - 举例：用户u粉了多少人

第4部分

总结展望

总结展望

- 模型
  - 简易分离第1档+舍弃第2档+三分类随机森林
  - 历史数据1.2%作为训练集，实时数据8.6%需要RF预测
  - 训练：500棵树的RF需15分钟
  - 预测：16000条/CPU/分钟
- 特征
  - 用户特征：相对稳定
  - 微博特征：O(M)
  - 用户微博特征：O(1)

总结展望

- 特色
  - 单模型，部署难度与执行成本低
  - 数据预处理，执行效率高
  - 避免未来信息，实用可行
- 展望
  - 生产环境中用滑动窗口产生训练数据
  - 其他分类模型，如GBRT
  - 更好的简易方法提取第1档
  - 尝试区分第2档

致谢

- 感谢新浪微博提供宝贵的数据
- 感谢阿里提供功能强大的平台
- 感谢天池团队辛勤地付出
- 感谢比赛中共同成长的小伙伴
- 感谢今天在场聆听的老师和同学