
新浪微博互动预测大赛答辩
2015天池大数据竞赛
TIANCHI天池
给女朋友赢旅游经费
2015.12.22

团队介绍

周兴

吕福煜

中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

网络数据科学与技术重点实验室
Key Laboratory of Network Data Science & Technology, CAS

提纲

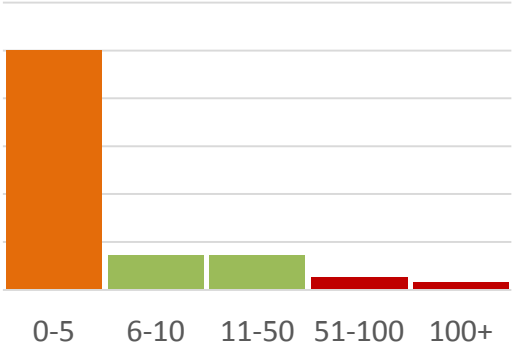
任务描述

- 目标：预测用户发表博文一天后的互动数位于第几档（总共5档）

赛题建模: Multi-Class Cost-Sensitive Learning

- 多分类问题

档位数量大致分布



数据倾斜

模型评价：

$$\text{precision}=\frac{\sum_{i=1}^{5} weight_i \times count\_r_i}{\sum_{i=1}^{5} weight_i \times count_i}$$

Cost-sensitive

问题一：如何处理**Cost-sensitive**的多分类问题？

赛题建模：样本选择

- 用户的互动数随时间的波动

我是歌手播出前
600+

我是歌手播出后
10000+

问题二：如何处理用户**不同时期**的互动数的变化？

赛题建模：解题思路

- 样本选择

方案一：历史所有用户的微博作为训练样本

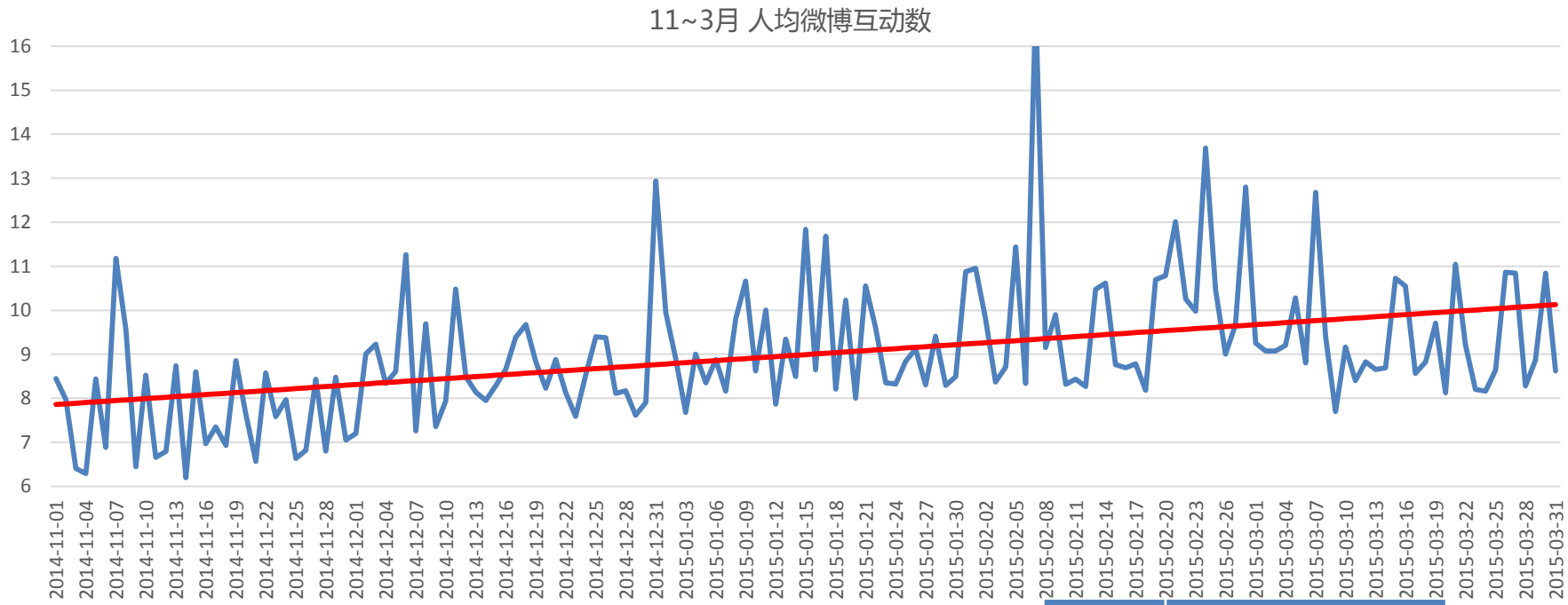
方案二：特定时间区间内用户的微博作为训练样本

- Cost-sensitive 多分类问题

方案一：对样本按权重进行采样

方案二：对样本按权重重复生成样本

赛题建模：数据分析



- 用户的人均微博数随着时间的增长有一定增长趋势
- 过早的数据与当前的数据存在较大的差异

月份	人均互动数
11	7.6584
12	8.73
1	9.177
2	10.1251
3	9.33

赛题建模：解决方案

- 样本选择方案

  我们选择最近一个月的**<用户，微博>**作为我们的训练样本

- Cost-sensitive 多分类问题

  按赛题给定得类别权重进行重复样本生成

特征工程

用户特征

- 历史统计数值特征
- 用户的影响力特征

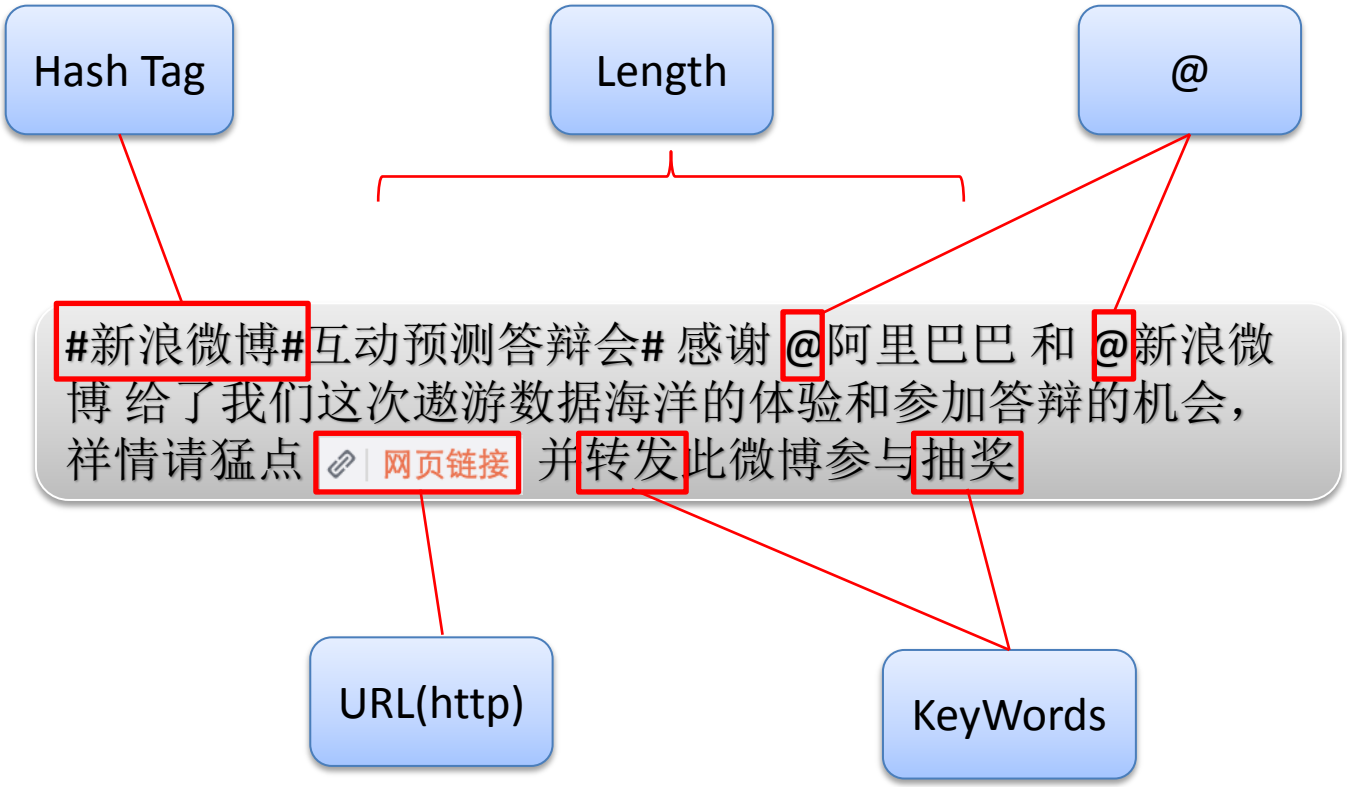
文本特征

- 微博博文特征

特征工程-用户特征

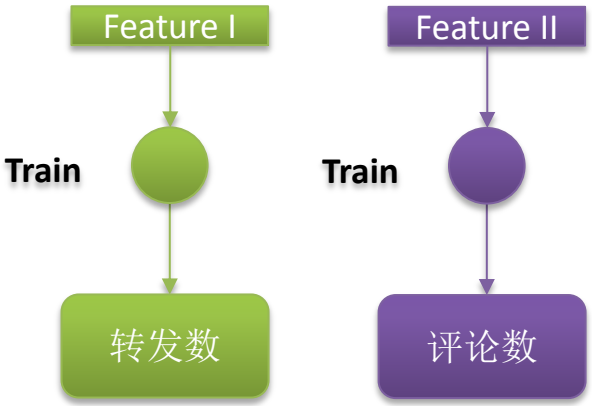
整体的统计特征	• 最大值、最小值、均值、方差、中位数
分类别的统计特征	• 细化整体的统计特征，对每个类别分别统计
用户的影响力	• PageRank • 粉丝数、关注数 • 是否大v：fans>1000 and fans/follow > 1

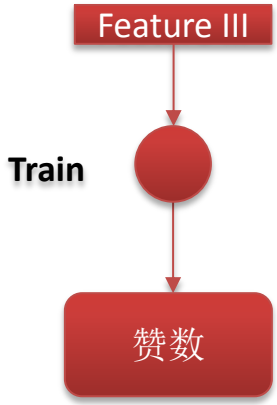
特征工程-文本特征


Hash Tag
Length
@
#新浪微博#互动预测答辩会# 感谢 @阿里巴巴 和 @新浪微博 给了我们这次遨游数据海洋的体验和参加答辩的机会，详情请猛点 网页链接 并转发此微博参与抽奖
URL(http)
KeyWords

模型框架:(一赛季) Multioutput Regression

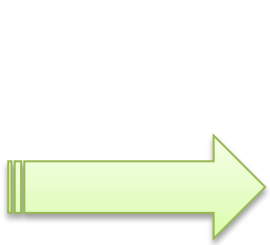
对每个目标训练一个模型：单输出节点优化

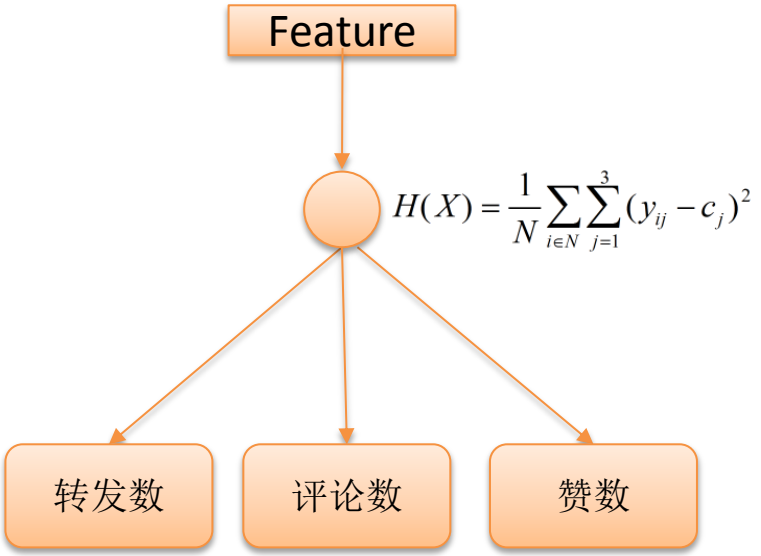




$$H(X)=\frac{1}{N}\sum_{i\in N}(y_i-c)^2$$

对多个目标训练一个模型：多个输出节点优化

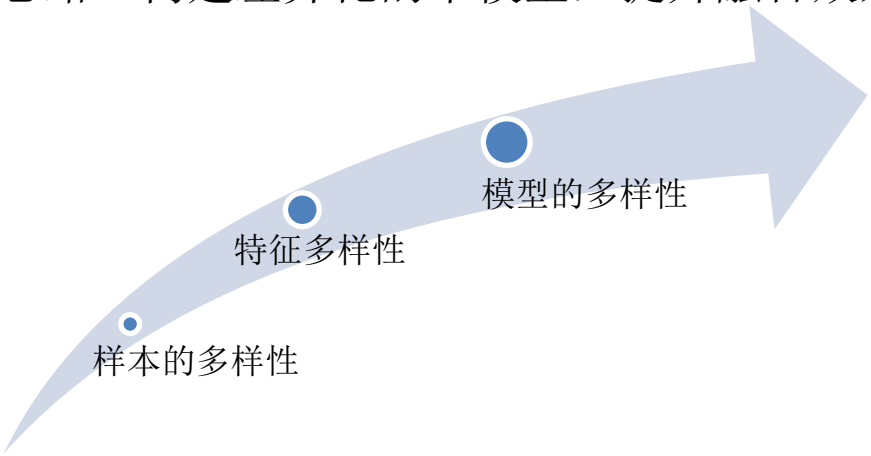




9 9 给女朋友赢旅游经费 中国科学院 31.17%

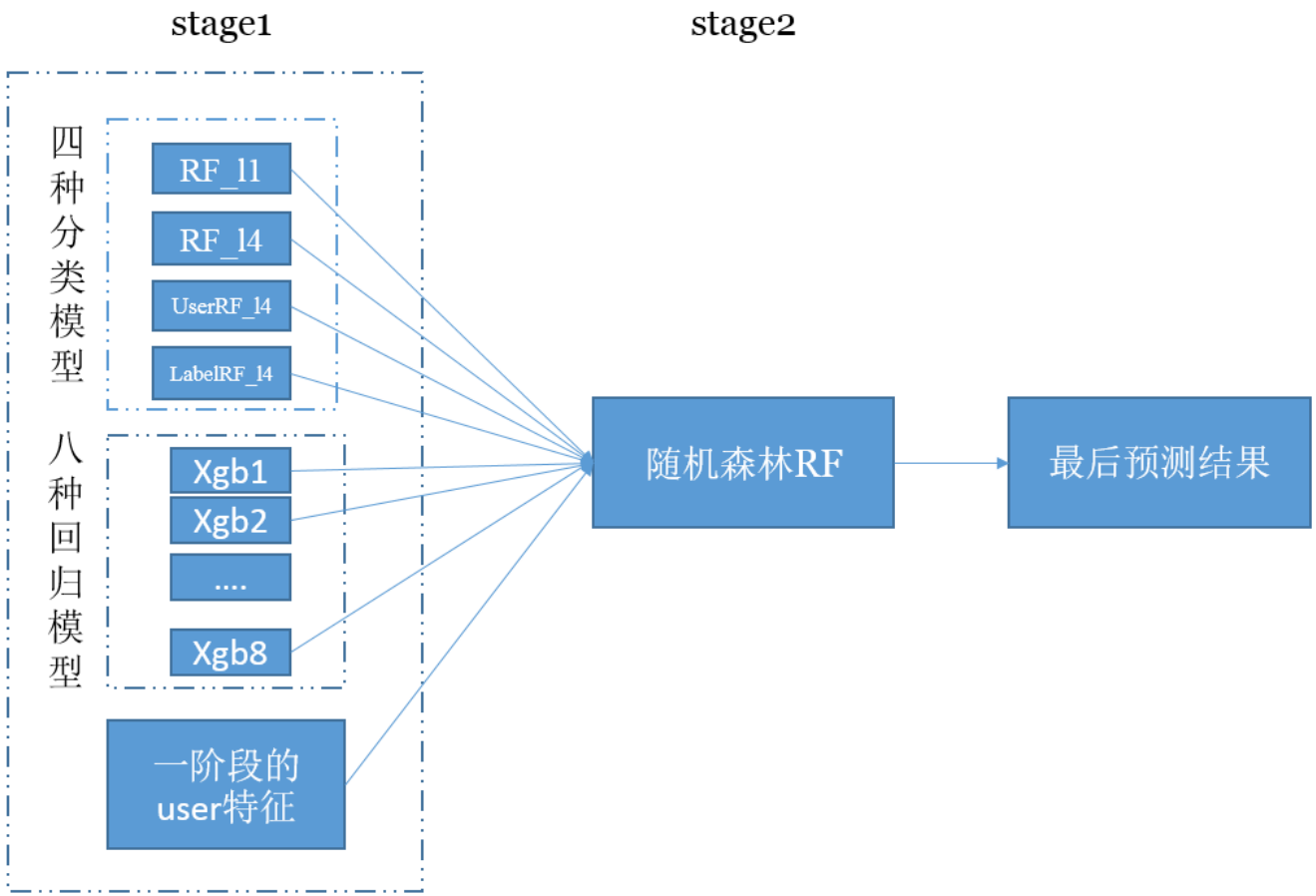
模型框架:(二赛季）二阶段模型融合框架

模型融合思路：构建差异化的单模型，提升融合效果



- 样本多样性：不同月份的训练样本
- 特征的多样性：不同特征 → 训练不同模型
- 模型多样性：不同模型具有差异性 → 分类模型 + 回归模型

二赛季：二阶段模型融合框架



分类单模型

- 我们主要训练了4种差异化的分类模型

RF_L1	RF_L4	UserRF_L4	LabelRF_L4
基于最近一个月的统计特征+文本特征	基于最近四个月的统计特征+文本特征	只使用最近4个月用户的特征（无文本特征）	基于最近4个月分区间的统计特征+文本特征

eg. 差异化结果

label_real	catg_pred	catg_real	catg_precision
1	20230630	20983810	0.9641066136225976
2	12050	295800	0.040736984448952
3	251915	320654	0.7856287462498519
4	13032	67363	0.193459317429449
5	74976	86515	0.866624284806103

RF_L1各个类别的分类结果

label_real	catg_pred	catg_real	catg_precision
1	20257240	20983810	0.9653747341402729
2	1276	295800	0.004313725490196078
3	258308	320654	0.8055661242335974
4	12036	67363	0.17867375265353383
5	75239	86515	0.8696642200774433

RF_L4各个类别的分类结果

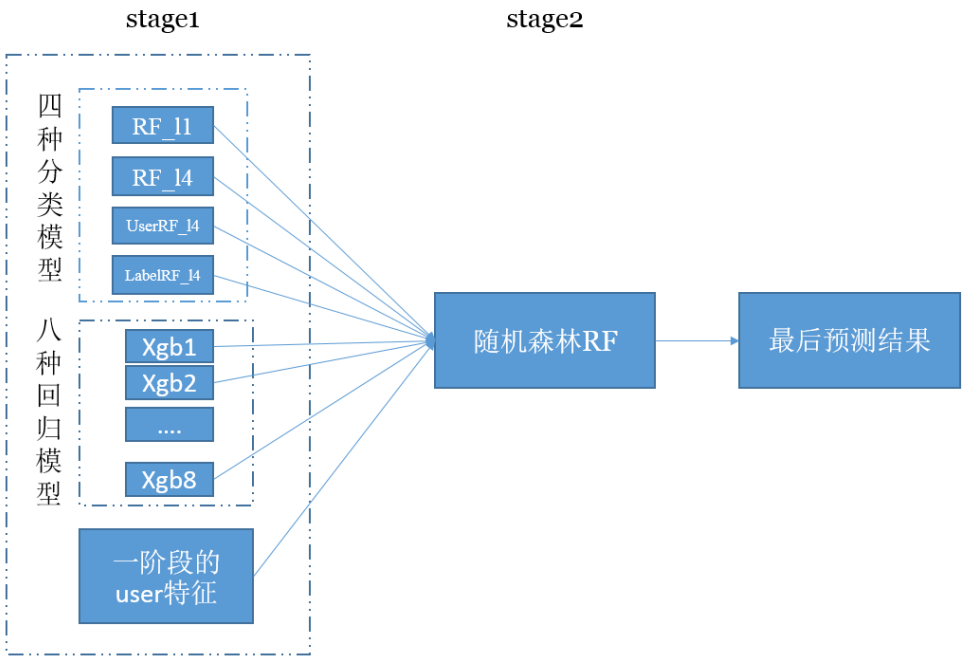
回归单模型

构造多样性

回归单模型细节

XGBOOST 1...8

二赛季：二阶段模型融合框架



- 二阶段的输入结合了4种分类模型+8种回归模型+user特征，采用RF进行多分类

总结

- 特征工程决定上界！！！
- 重视数据分布、样本选择
- 理论结合实际、批判的态度看paper
- 耐心和毅力——坚持到最后

致谢

- 感谢阿里巴巴和新浪微博提供找到自己乐趣和探索自己热爱东西的平台
- 感谢所有热情、认真、友善的工作人员
- 感谢一起比赛竞争的小伙伴使得我们不断超越自己

谢谢！