

新浪微博天池大赛答辩

2015天池大数据竞赛

TIANCHI天池

队伍：电光火石



吴朝恬

西南交通大学

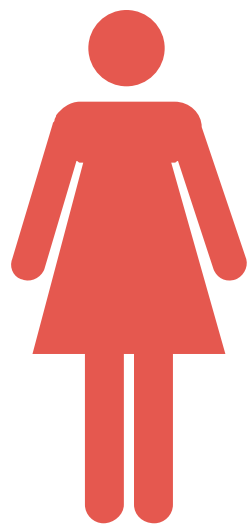
整体架构，特征提取



蔡少阳

中国科学与技术大学

模型调优，特征组合



张涛

国防科学与技术大学

数据分析，代码管理



目录

问题理解

特征提取

模型选择

比赛总结

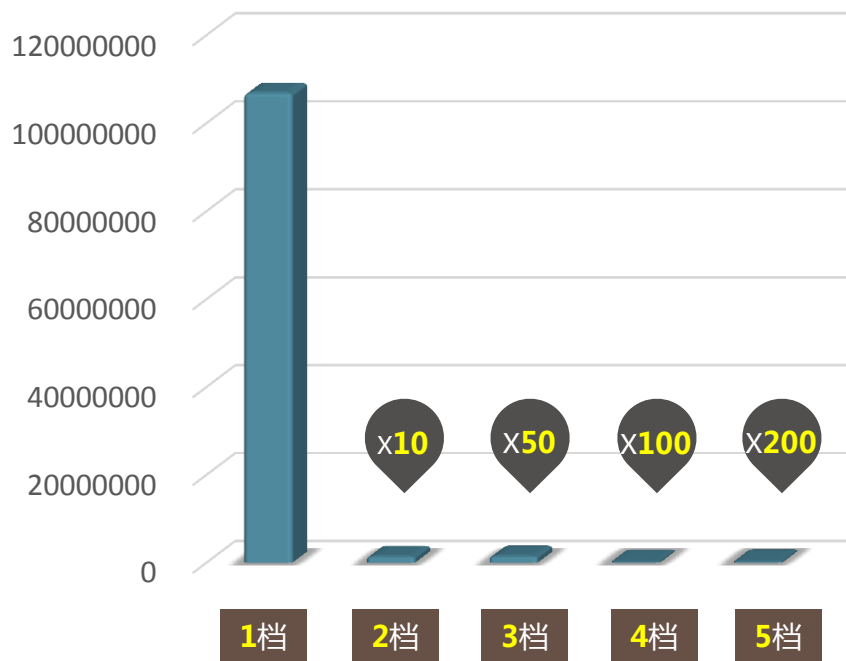
问题理解



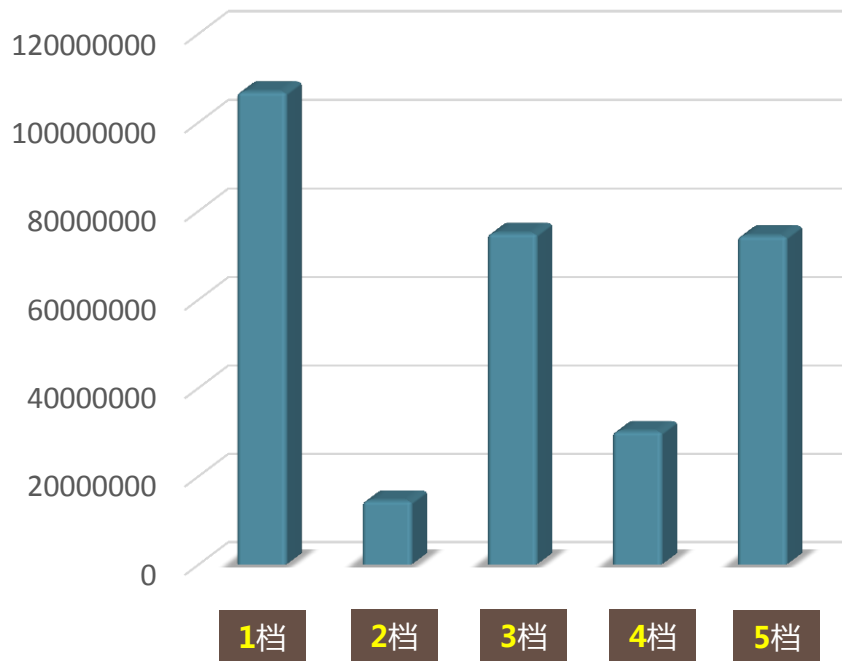
问题理解

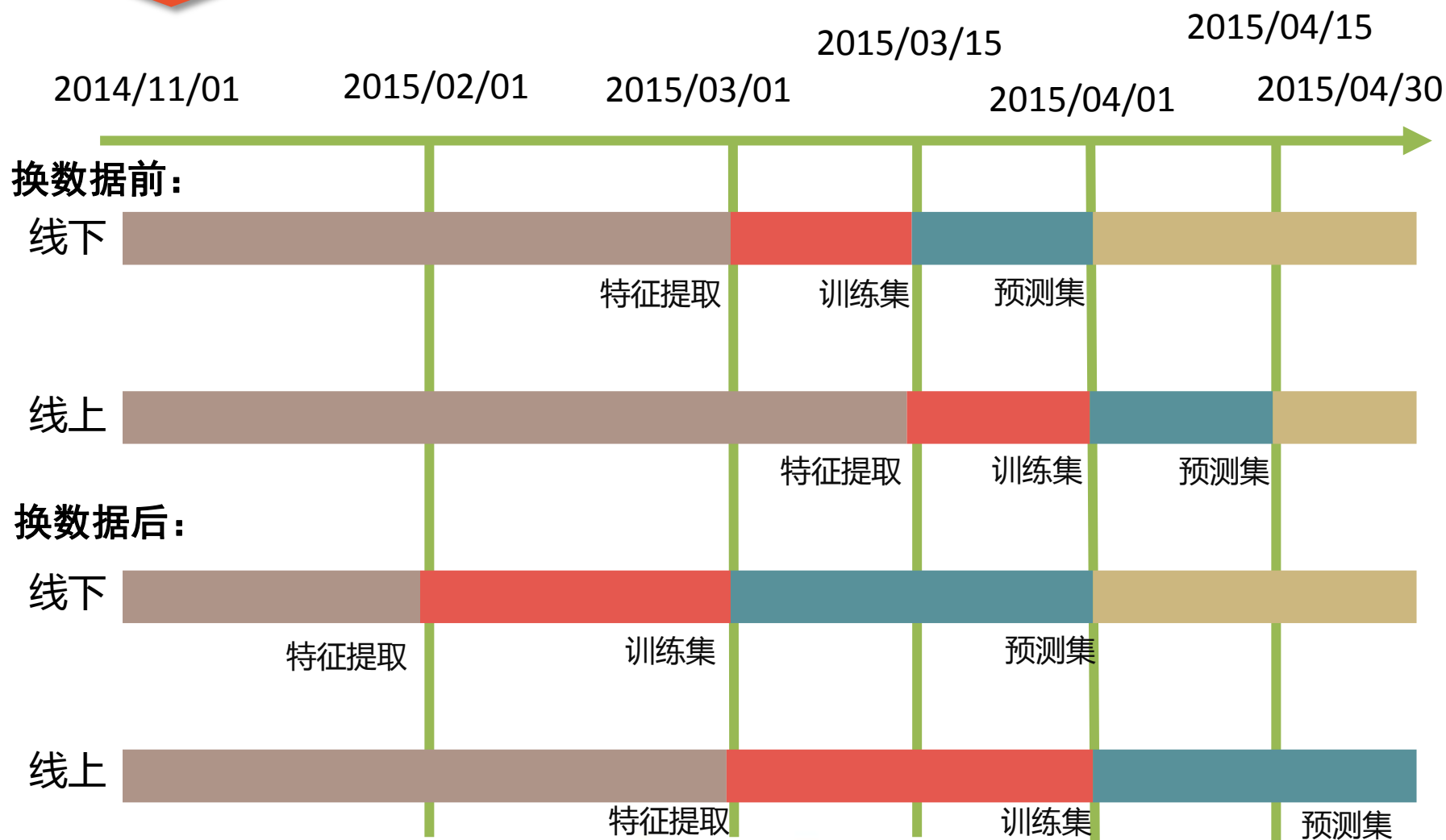


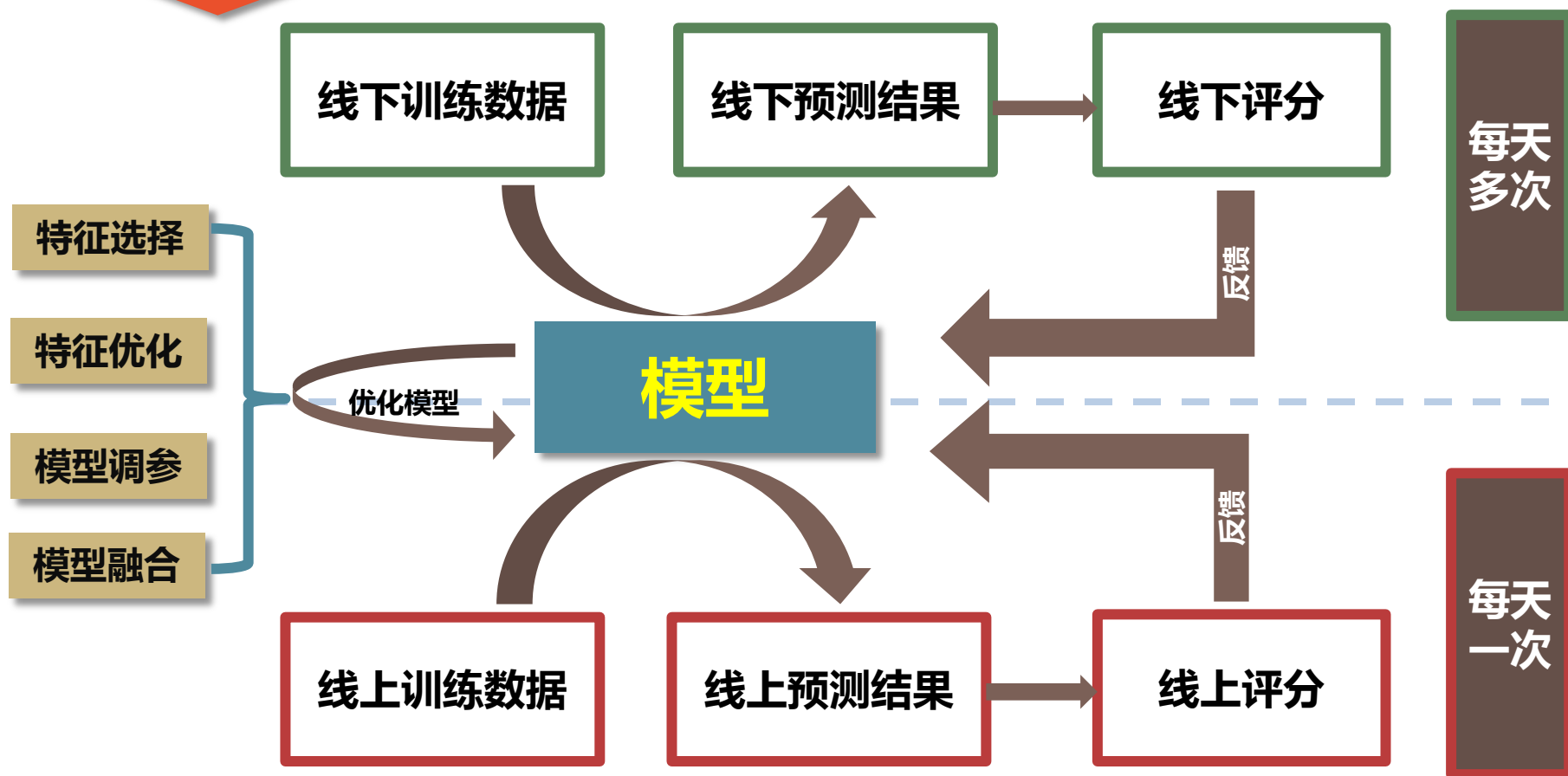
重复采样前



重复采样后







特征提取

用户特征

历史互动数

用户第一档概率
用户第二档概率
用户第三档概率
用户第四档概率
用户第五档概率

粉丝

粉丝数量
粉丝活跃度

互粉数
关注数

.....

微博特征

发微博时间

星期
小时

微博长度

微博文本特征

广告
热门话题

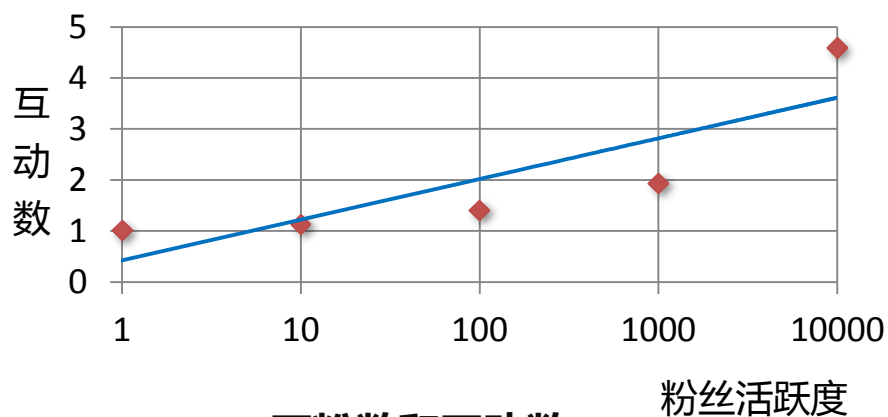
...

...

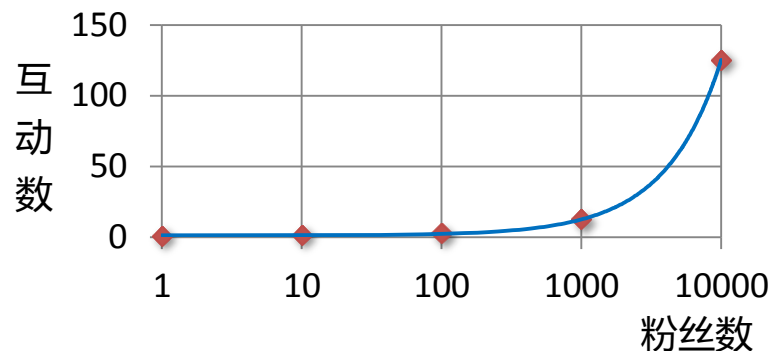
特征提取

2.1 用户特征

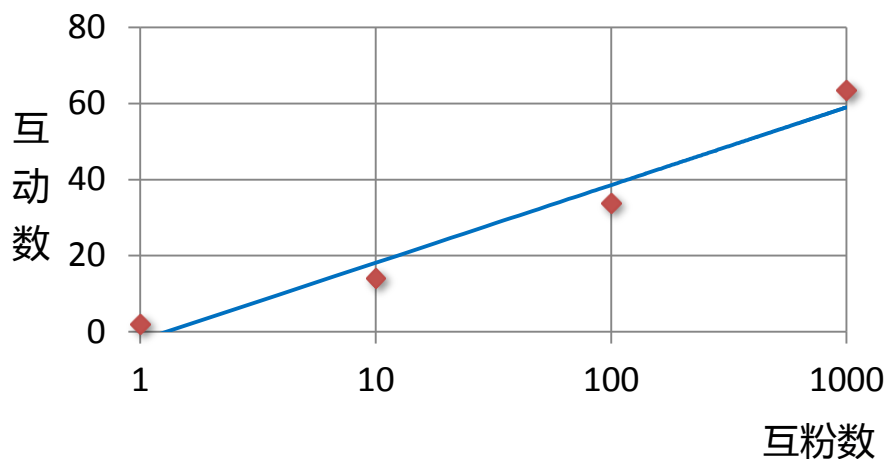
粉丝活跃度和互动数



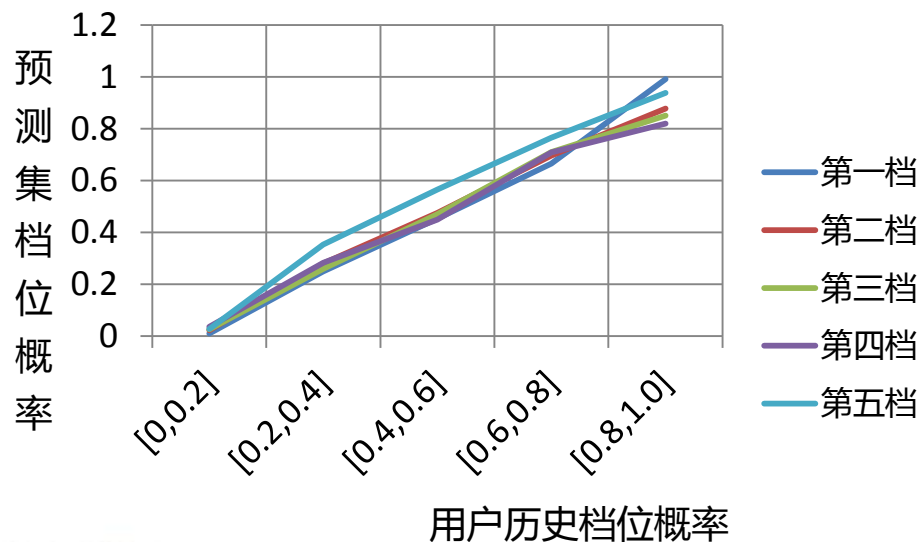
粉丝数和互动数



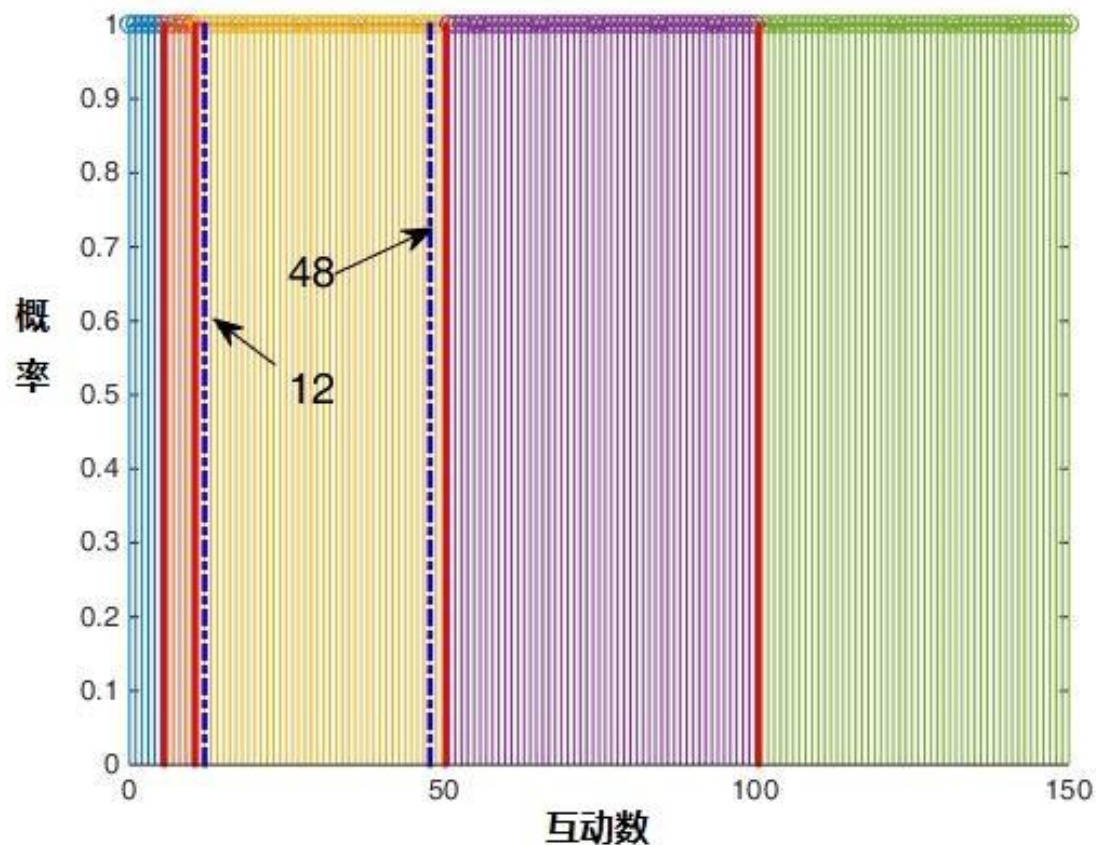
互粉数和互动数



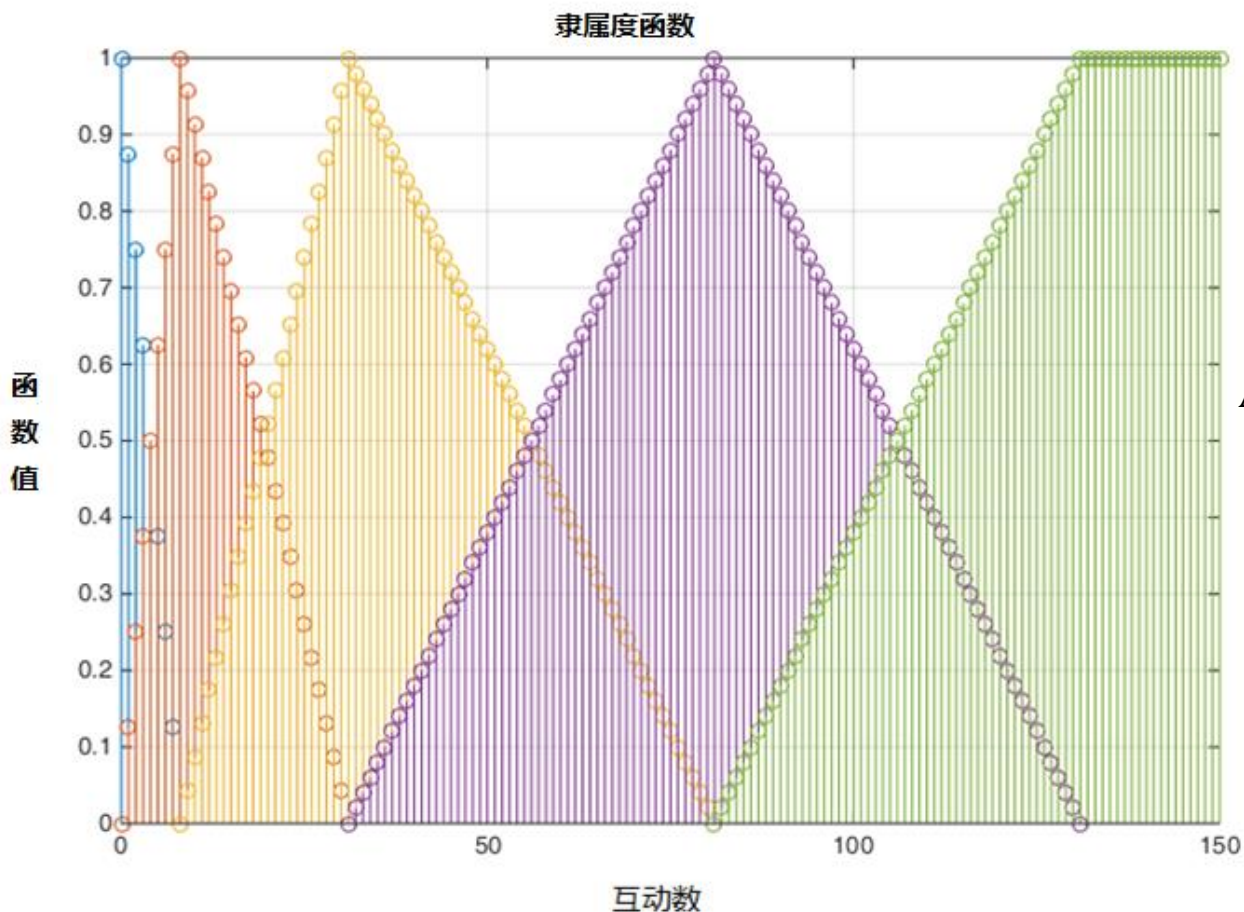
档位概率分布



模糊集档位特征概率模型



模糊集档位特征概率模型

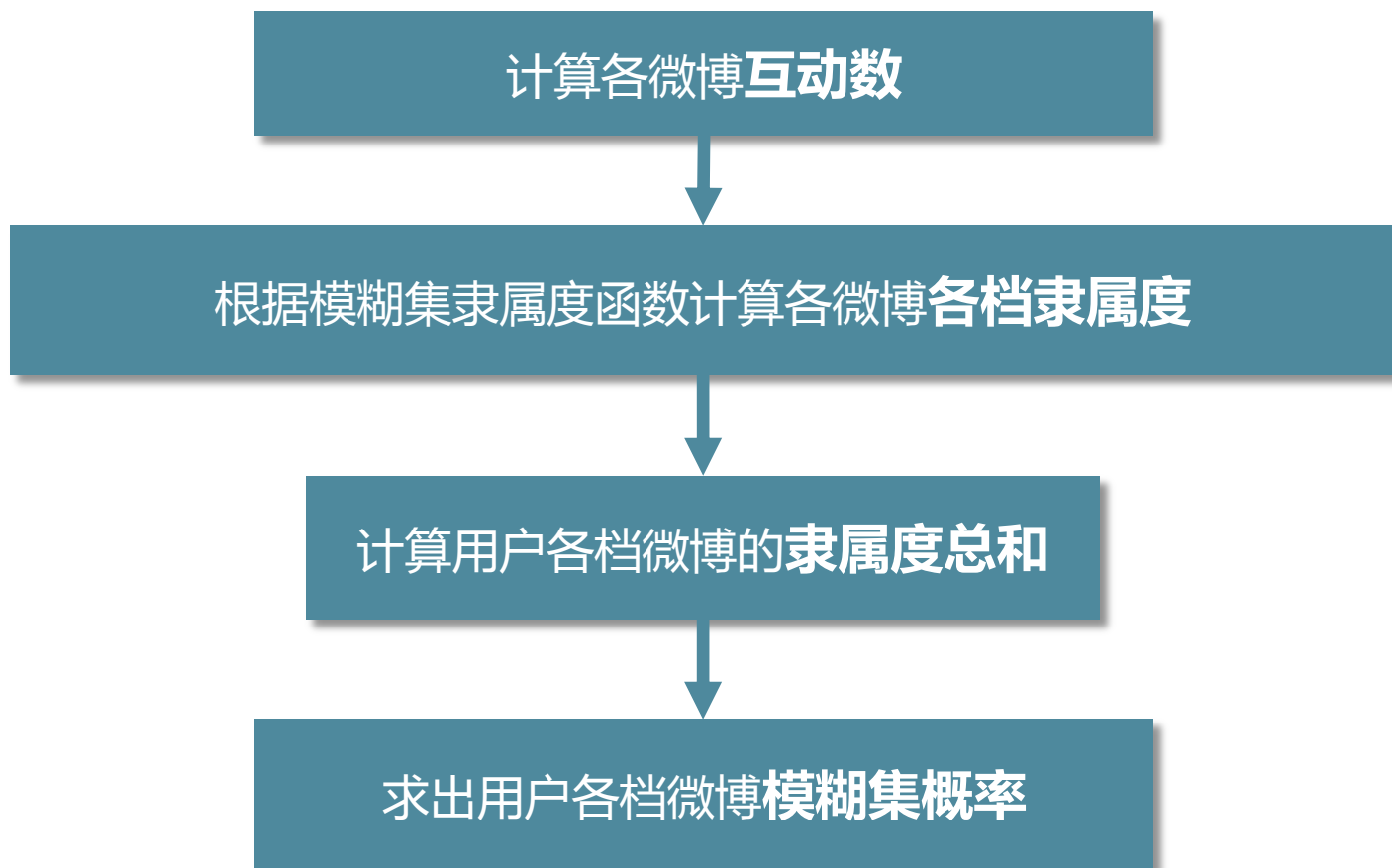


$$A(n,m)=\begin{cases} \frac{8-n}{8} & m=1, 0 \leq n \leq 8 \\ 1-\frac{8-n}{8} & m=2, 0 \leq n \leq 8 \\ \frac{31-n}{23} & m=2, 8 \leq n < 31 \\ 1-\frac{31-n}{23} & m=3, 8 \leq n < 31 \\ \frac{81-n}{50} & m=3, 31 \leq n < 81 \\ 1-\frac{81-n}{50} & m=4, 31 \leq n < 81 \\ \frac{131-n}{50} & m=4, 81 \leq n < 131 \\ 1-\frac{131-n}{50} & m=5, 81 \leq n < 131 \\ 1 & m=5, 131 \leq n \\ 0 & \text{others} \end{cases}$$

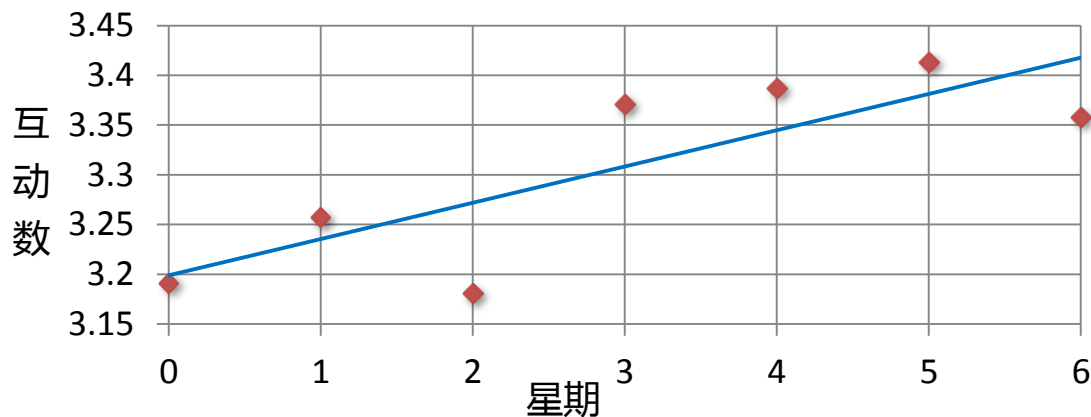
模糊集档位特征概率模型



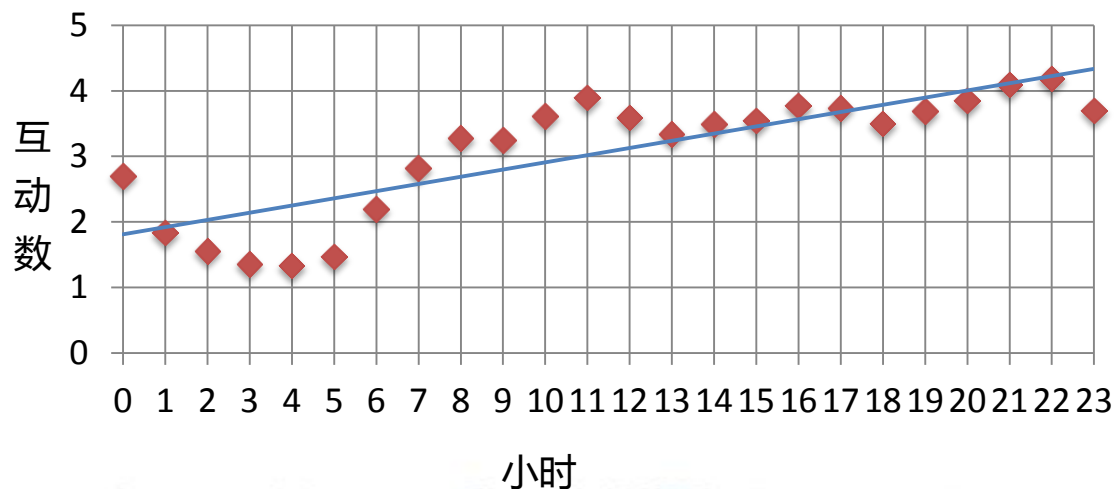
模糊集档位特征概率模型计算流程



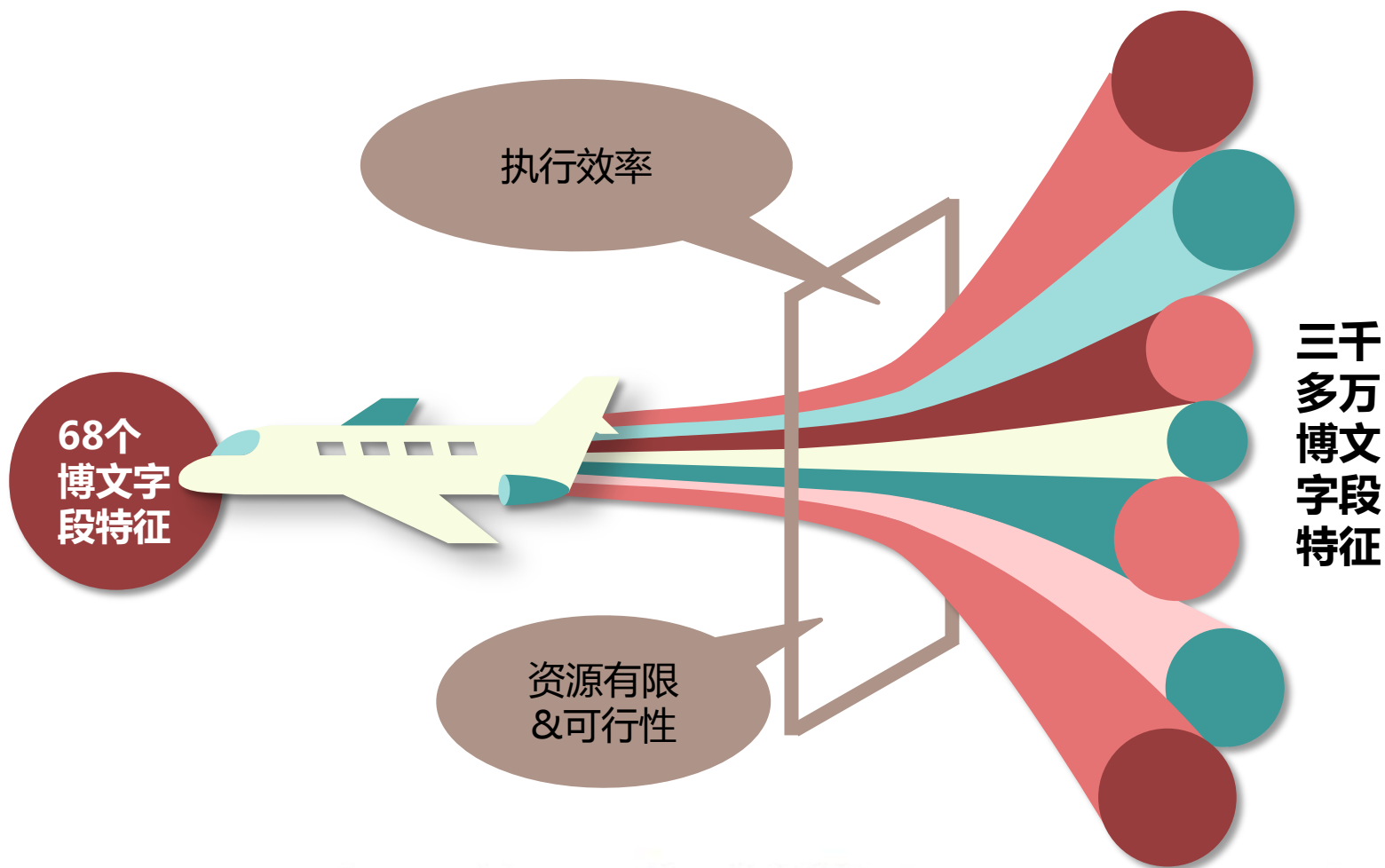
星期和互动数



小时与互动数

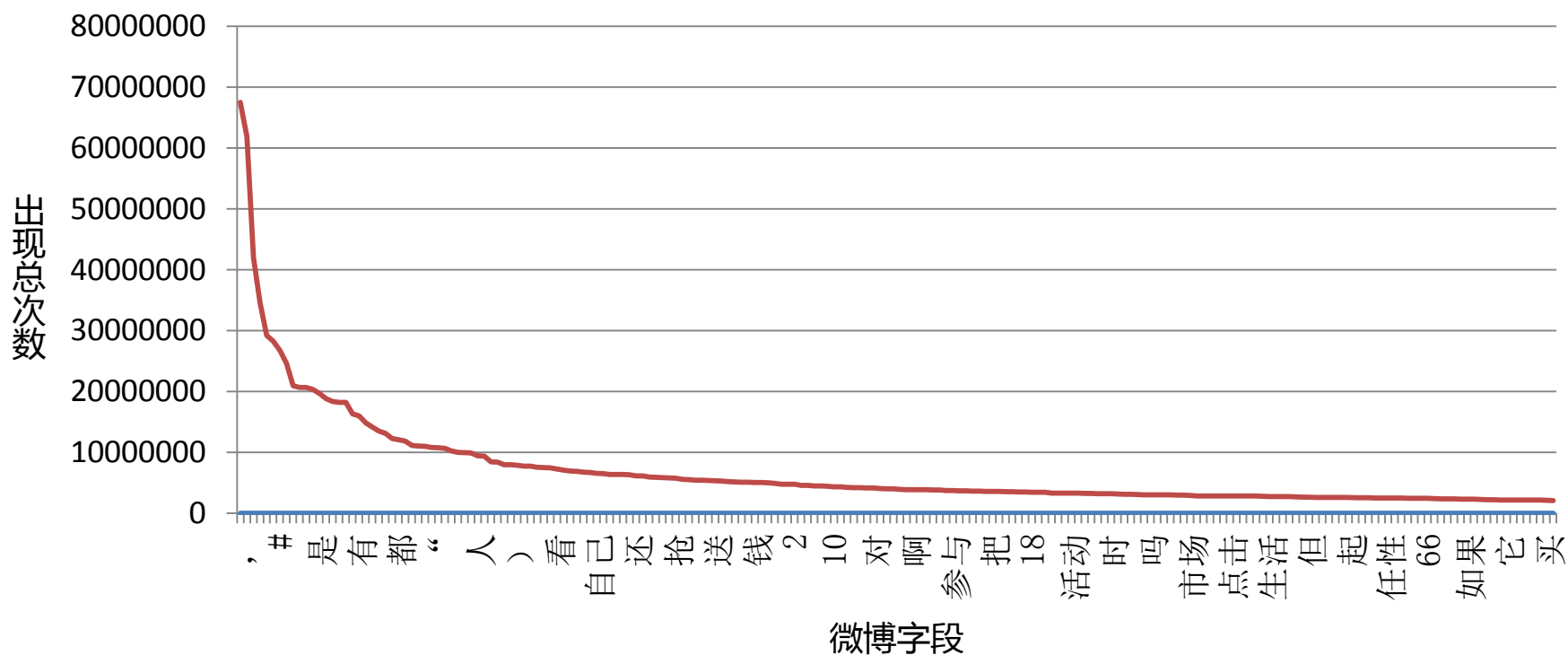


高效字段特征选择模型

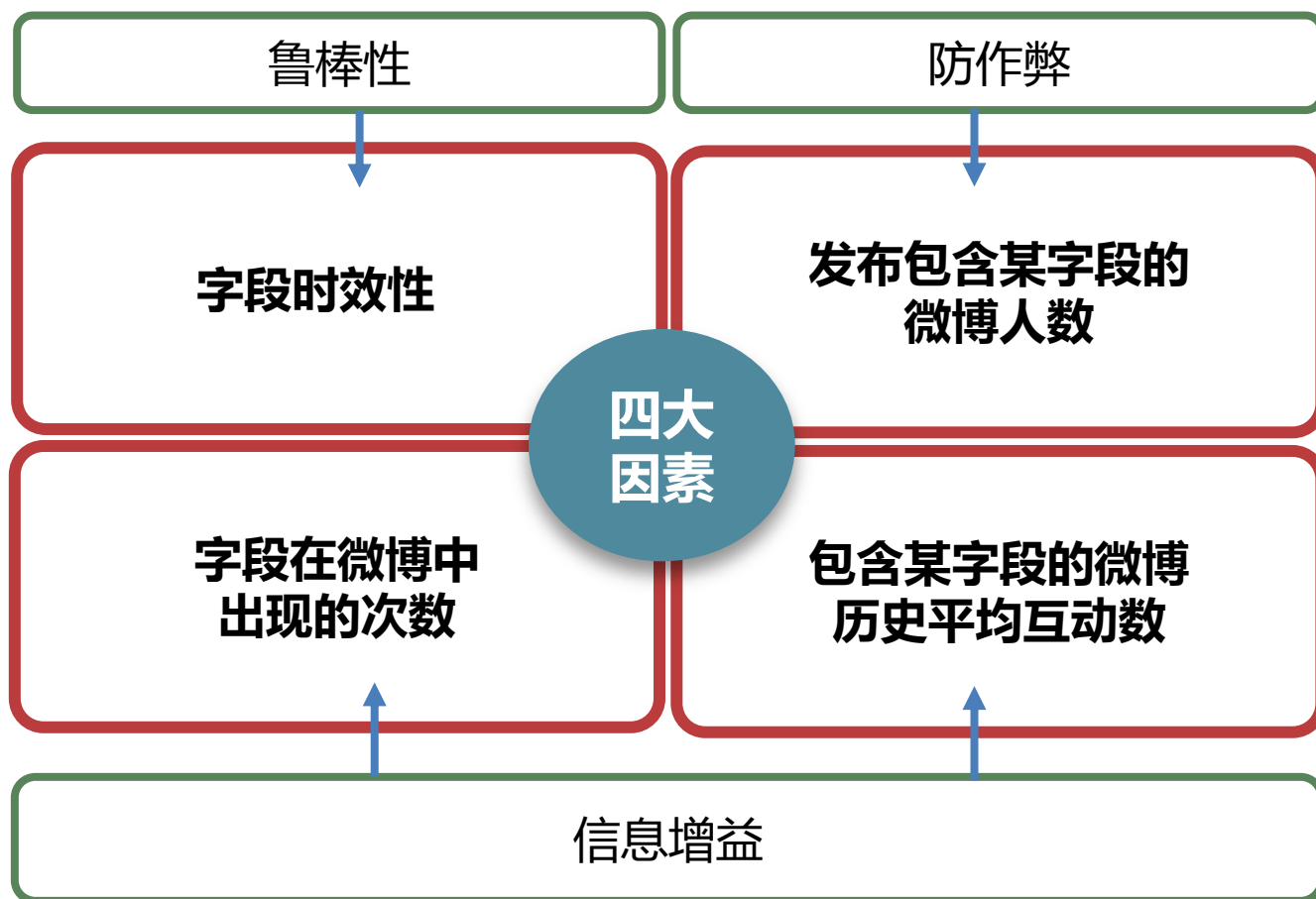


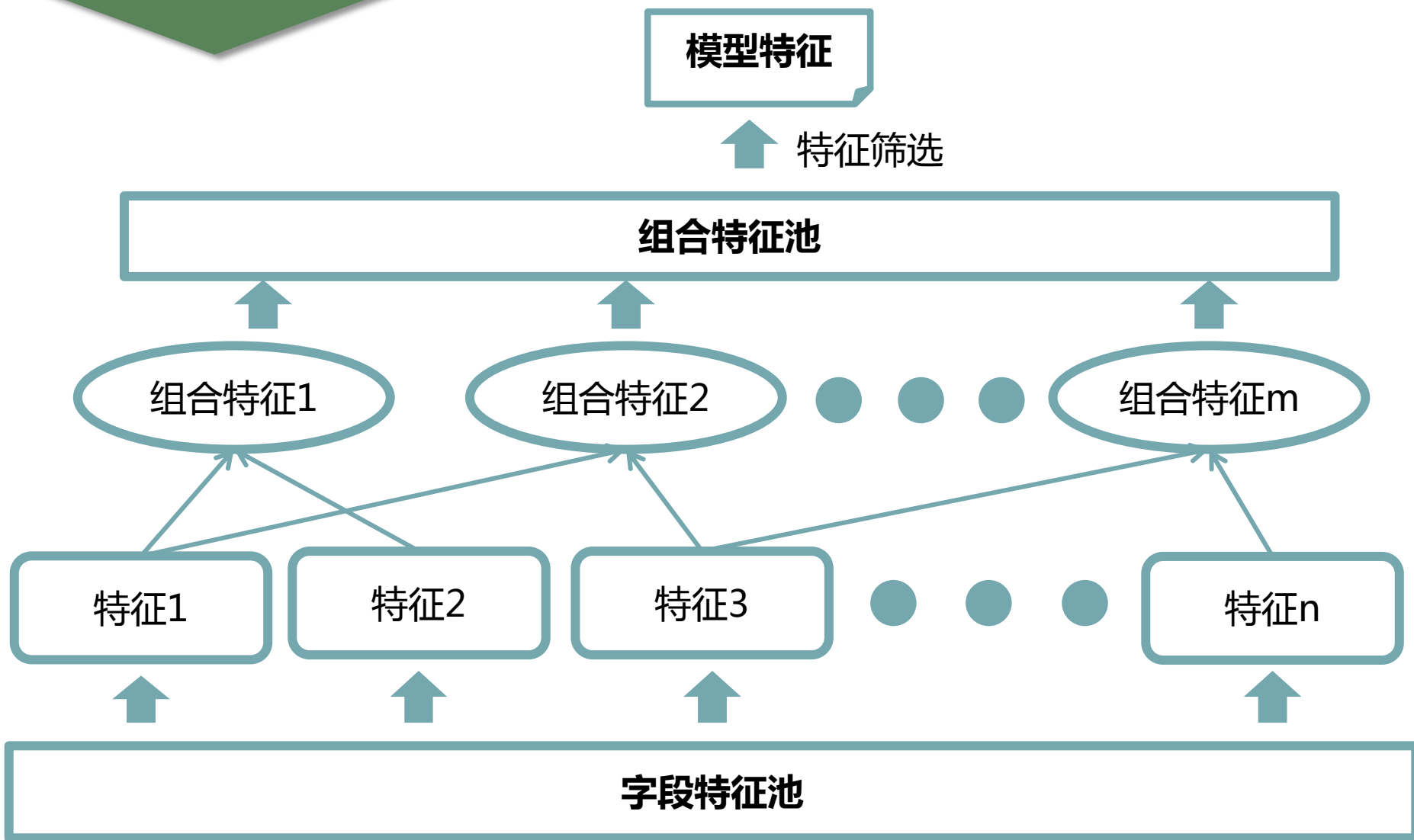
高效字段特征选择模型

微博字段出现频率排名



高效字段特征选择模型





特征提取

■ 2.3 特征改进

用户微博
互动数

用户各档
微博数

用户各档
微博概率

用户各档微
博模糊集概率

用户粉丝
数量

用户互粉
数量

用户互粉
粉丝活跃度

随机字段特
征

高效筛选字段
特征

组合字段特征

特征在改进，成绩在提高

多分类模型

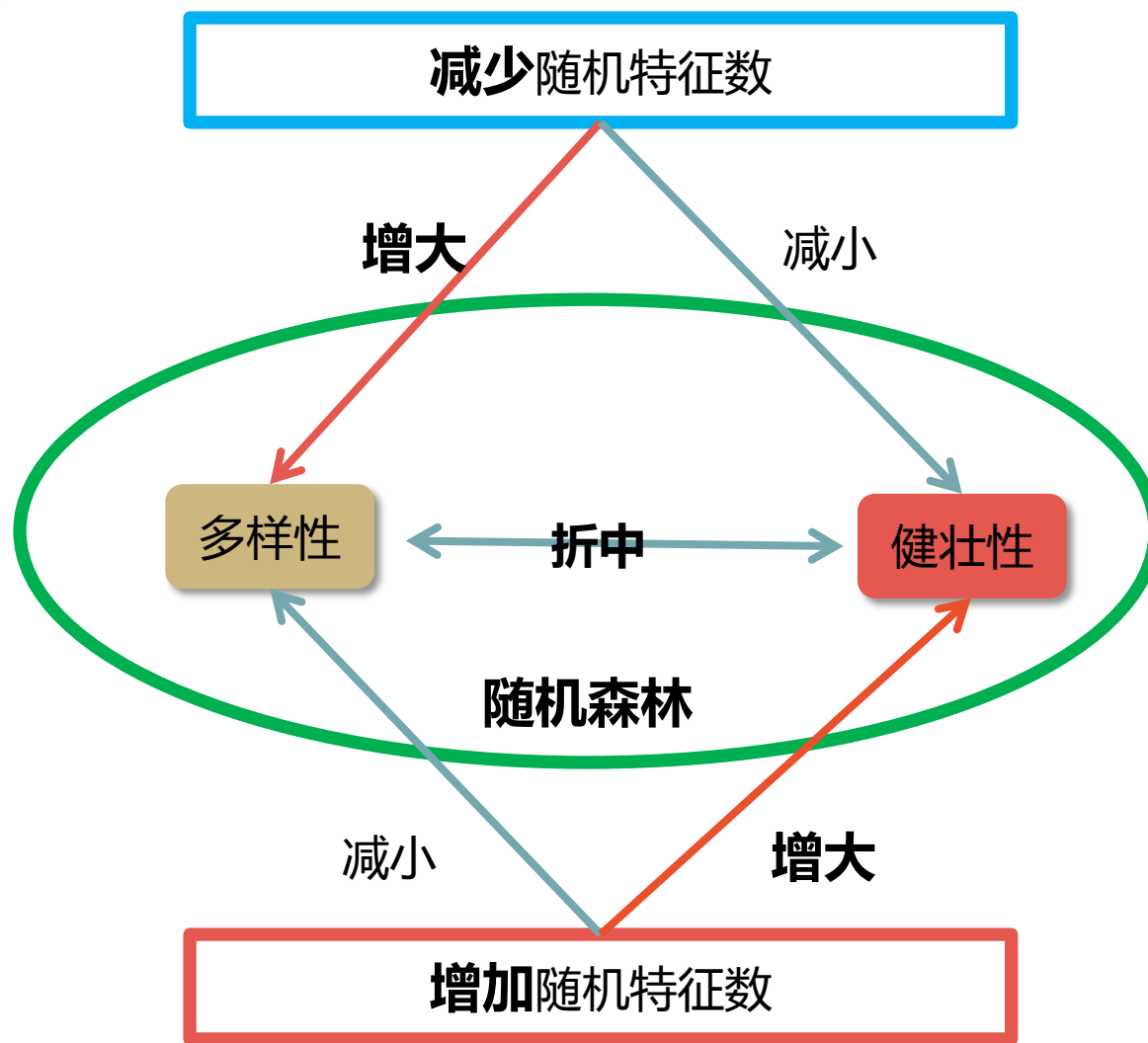
逻辑回归 (LR)	预测较快，易过拟合
随机森林 (RF) 	训练较快，抗噪能力强 预测较慢，泛化能力强
朴素贝叶斯 (NB)	训练、预测都较快，精度差

二分类模型

梯度下降回归树 (GBDT)	训练较慢，预测较快，精度高
支持向量机 (SVM)	平台没有引入核函数，精度差

五分类问题化二分类形式描述：

{ 第一档, { 第二档, 第三档, 第四档, 第五档 } }
{ 第二档, { 第一档, 第三档, 第四档, 第五档 } }
{ 第三档, { 第一档, 第二档, 第四档, 第五档 } }
{ 第四档, { 第一档, 第二档, 第三档, 第五档 } }
{ 第五档, { 第一档, 第二档, 第三档, 第四档 } }



■ 3.3 模型融合

■ 多分类问题化为多个二分类问题

01
多分类化二分类

02
模型投票

■ **多**个分类器分别预测，投票择优预测结果。

模型融合

03
精化分类

■ **分**类问题先粗分，再细分，多次逐步分类预测

04
分时预测

05
多参预测

■ **不**同参数的模型分别预测再融合

■ **不**同时间窗口的训练集分别预测再融合



92维特征



单模型

一次预测完成




模型训练

590.162s



预测时间

1764.894s

- 
- 字段特征选择模型使得文本**字段特征**选取会随着训练集变化而变化
 - 用户特征的取值具有**时效性**，会随着训练集变化而变化
 - 换数据前后都长期**保持第一名**，说明鲁棒性强



重复采样

把模型训练
目标与赛题
评分目标**统**
一起来



模糊集档位

模糊集档位
特征概率模
型，有效反
应**用户互动**
情况



字段选择特征

字段特征选
择模型，优
选**预测价值**
明显的关键
词特征

比赛总结

■ 4.4 历史成绩 & 经验分享



经验分享

- 精挑特征
- PAI命令取代画布
- 线下与线上**差距**的动态把握
- 线上**调参**勿过早

1

连续保持排行榜
第一名**20天**

获得**四次**周星星

2

3

换数据前后均长期
保持**第一名**

Thanks

Q&A