

阿里移动推荐算法大赛答辩

2015 天池大数据竞赛

TIANCHI 天池

新浪微博预测大赛分享

科学院南路6号

团队介绍

郭天佑



机器学习

主题模型

数学建模

周楠



自然语言处理

机器学习

侯建鹏



Spark

ACM

机器学习

中国科学院计算技术研究所

目录



问题分析



数据处理



特征工程



模型融合



总结思考



问题分析

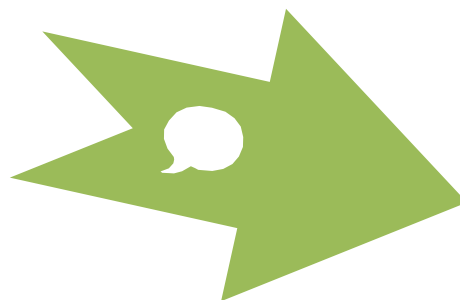
问题分析

博文的互动数一共划分为5档，0-5为1档，6-10为2档，11-50为3档，51-100为4档，100+为5档。

根据题目要求，互动数在同一档内的微博权重相同，因此我们认为没有必要精确地预测微博互动数本身，而是要确定其所在的档位。于是我们将互动数的预测转化为一个经典的多分类问题。

回归问题

直接对微博
的互动数预测



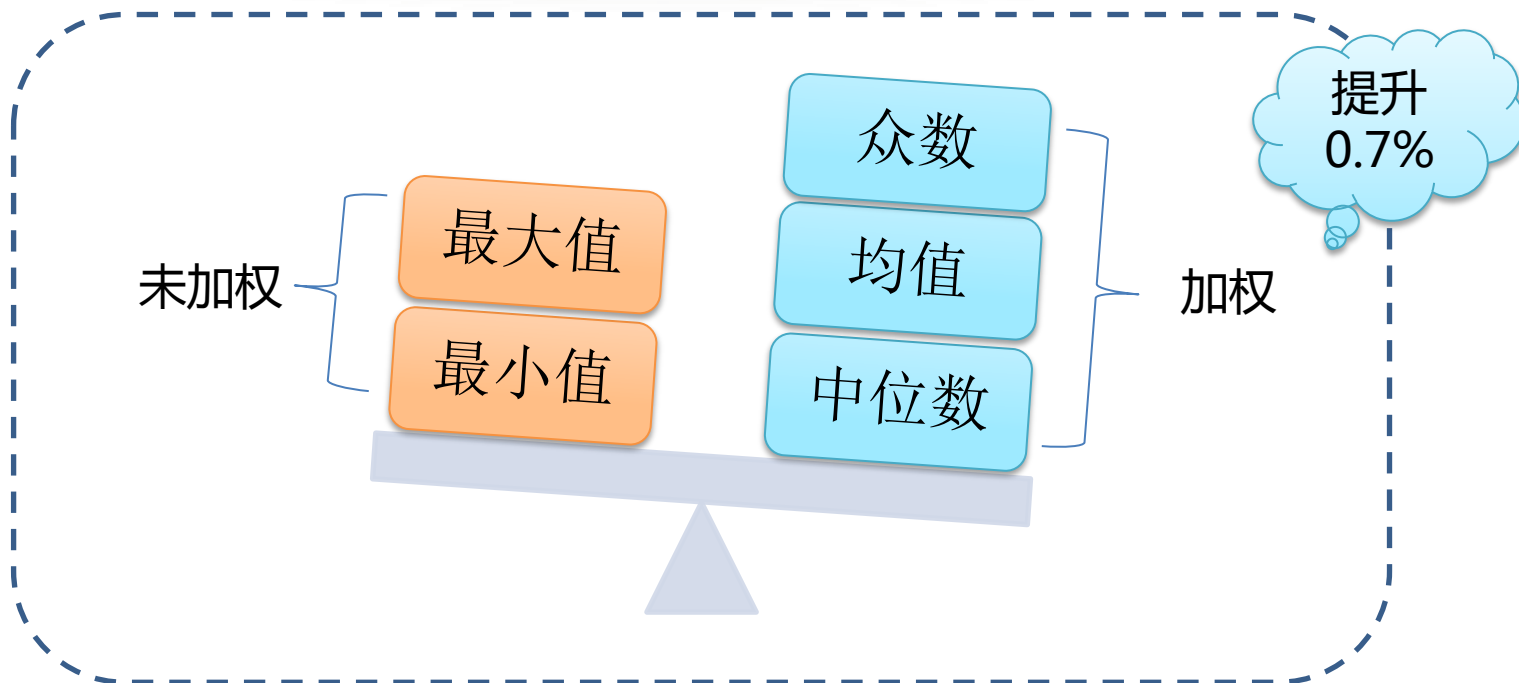
分类问题

对微博互动数所在档
位预测



特征工程

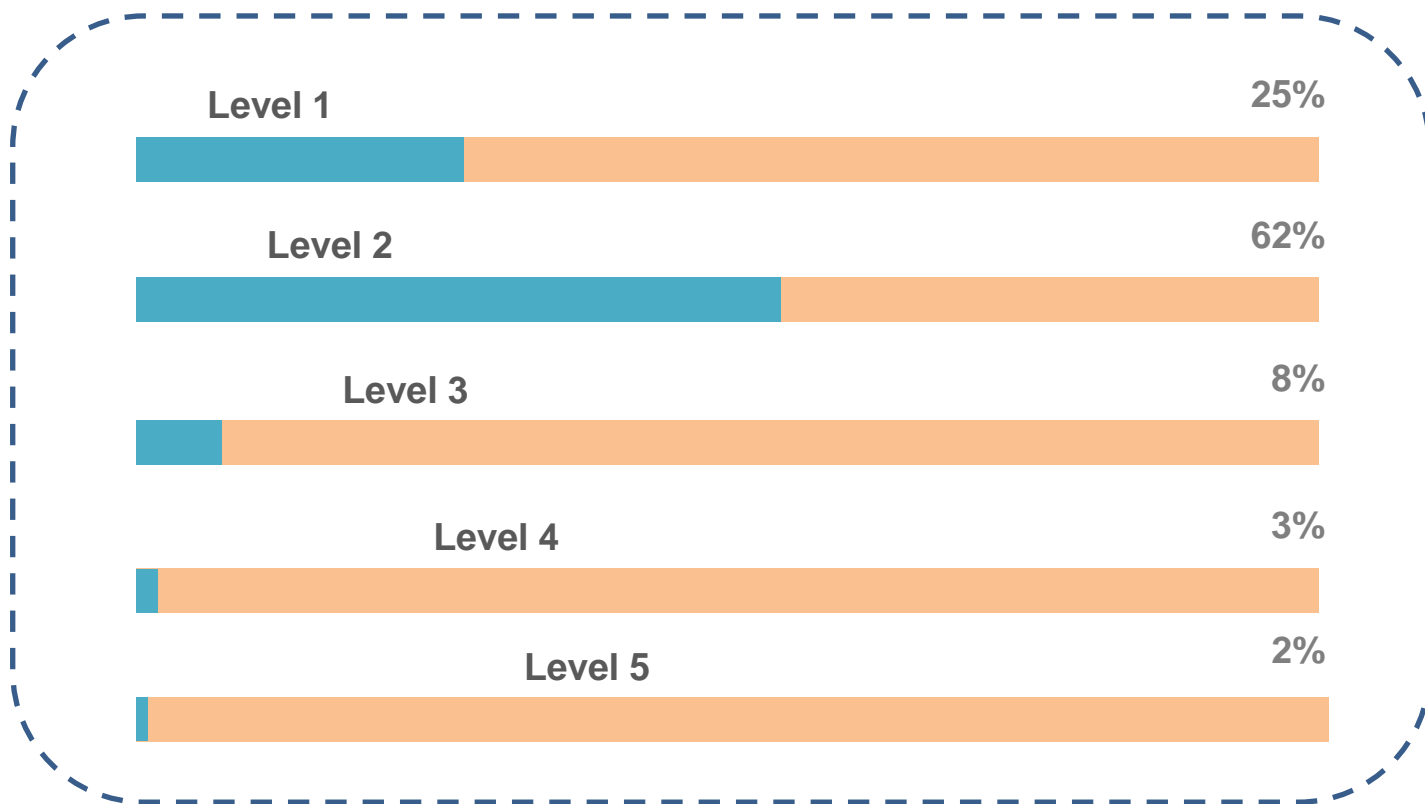
用户微博统计特征



最初引入中位数、众数是因为中位数和众数对噪声相对不敏感，这个系列的特征值往往能够起到互为补充的作用，并且从统计的角度来看能够可靠的反映真实数据的分布情况。

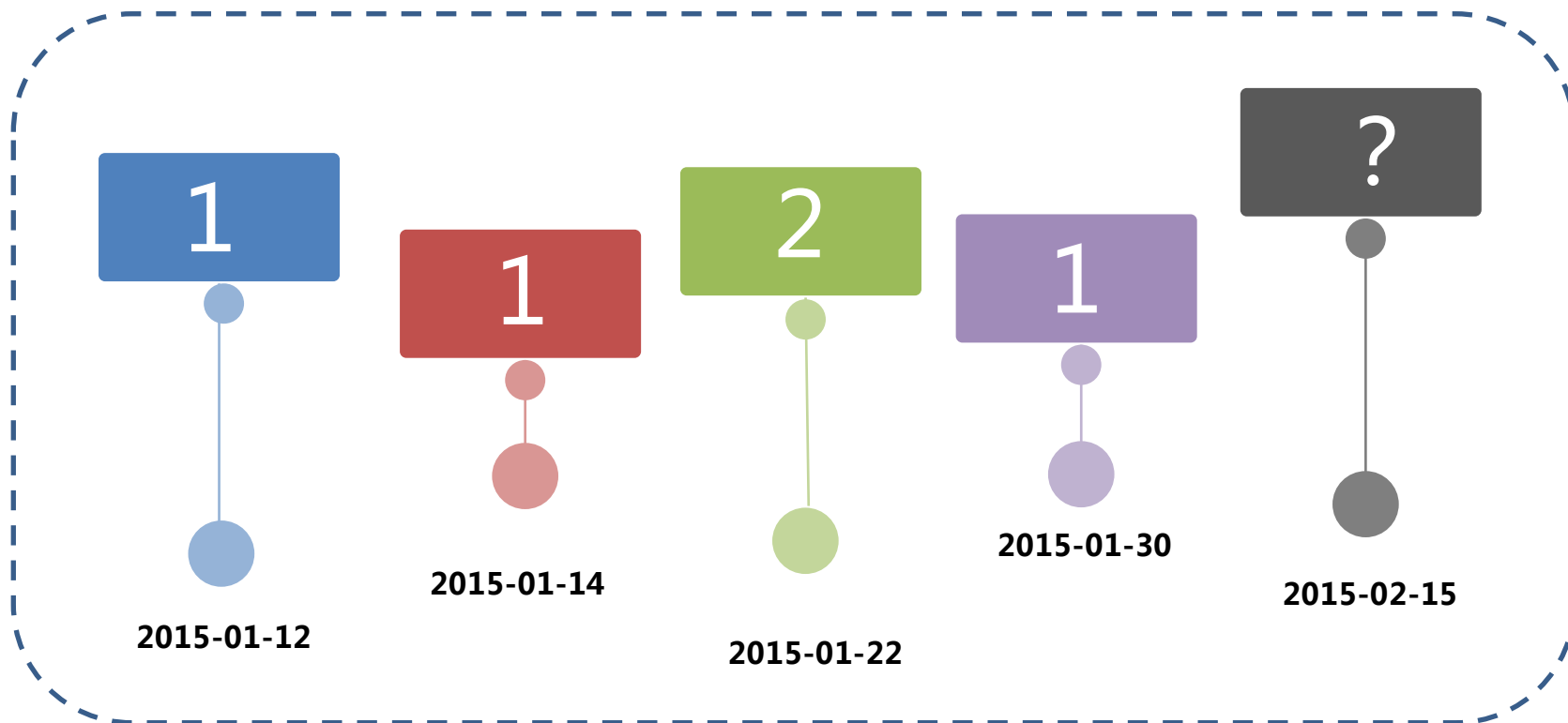
用户微博统计特征

根据不同的**时间窗口**，统计用户所发微博互动数在5个档位上的分布归一化得到，用户历史所发微博在不同档位的概率



用户最近微博档位特征

假设：用户最近所发微博互动数所在的档位有一定的**相关性**，可以根据用户最近微博的档位推测新微博的档位



用户社交关系特征



粉丝特征

主要关注用户的粉丝数，有效粉丝数以及关注他人微博数量



链接分析

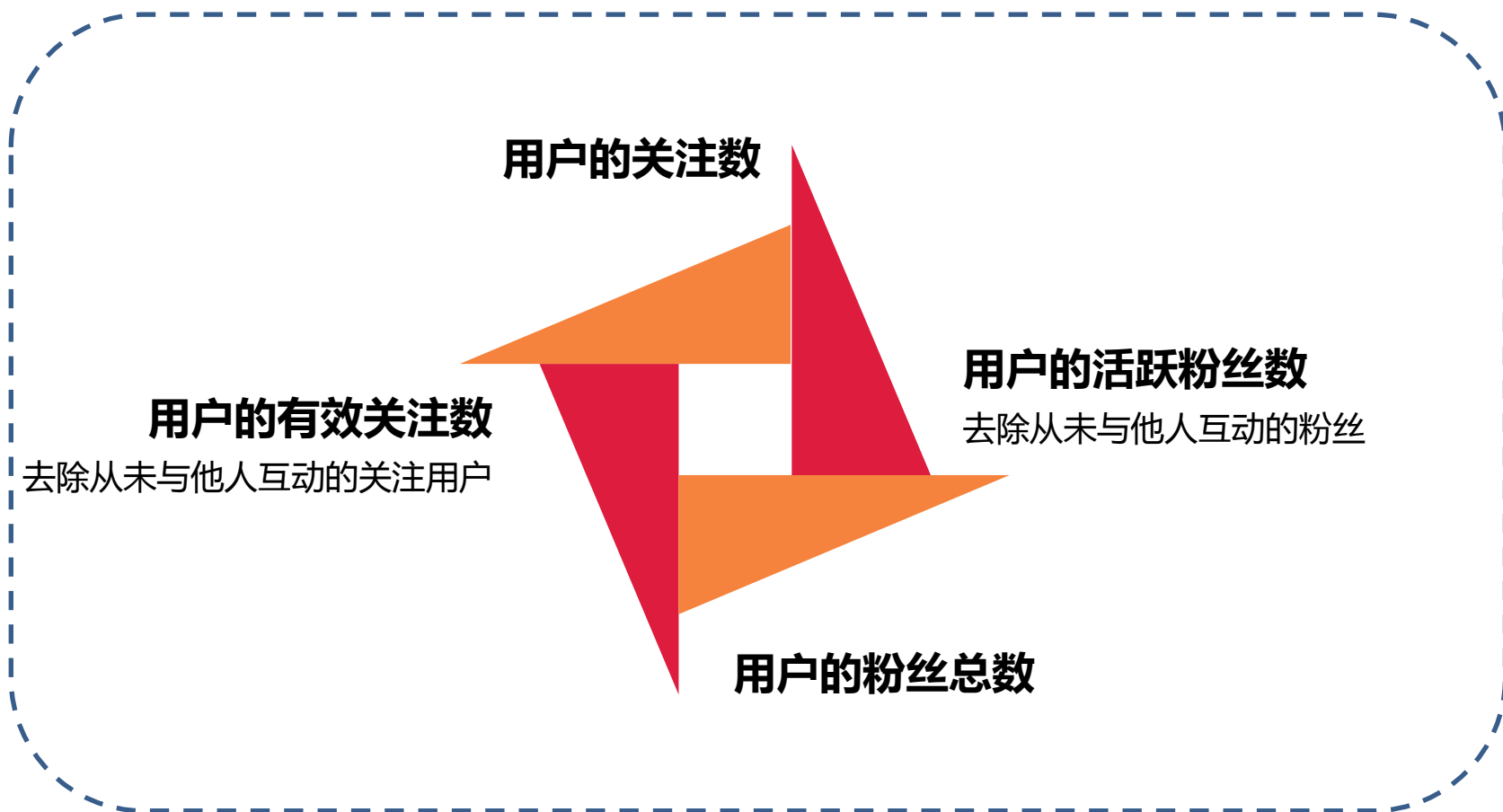
利用成熟的链接分析算法，对用户社交关系网络中的地位进行分析



用户互动情况分析

考察用户与他人互动的频度，根据不同时间窗口分别计算

粉丝特征



链接特征 PageRank

根据网页之间相互的超链接计算的技术，而作为网页排名的要素之一，Google用它来体现网页的相关性和重要性，在搜索引擎优化操作中是经常被用来评估网页优化的成效因素之一。



将用户在社交网络中的重要性，类比为网页在web中的重要性。将用户视为节点，关注关系视为有向边，利用PageRank算法，计算用户在网络中的权重

链接特征 Hits

按照HITS算法，用户输入关键词后，算法对返回的匹配页面计算两种值，一种是枢纽值（Hub Scores），另一种是权威值（Authority Scores），这两种值是互相依存、互相影响的。所谓枢纽值，指的是页面上所有导出链接指向页面的权威值之和。权威值是指所有导入链接所在的页面中枢纽之和。



用户在社交网络中的特征可以借助HITS算法中的hub和authority两个特征值进行建模，前者反映了当前用户关注了哪些“高质量”用户，而后者则体现了用户本身被哪些“高质量”用户关注，进而可以表现用户的重要程度。

用户互动特征

用户互动总数（转评赞其他用户）

用户互动总数（时间窗口为最近2个月）

用户互动总数（时间窗口为最近1个月）

微博内容特征



基本特征



相同前缀微博特征



word2vec文本聚类特征



微博在各档位的概率分布




表情特征

微博基本特征

知

知乎 V

「如何阅读艺术书籍？以《加德纳艺术通史》为例，我们可以先翻翻整本书里面的插图，从里面选出自己比较喜欢的画，或者艺术家，比如莫奈。那不妨从这一段看起。这种阅读方法，就像是往池塘里丢下一颗石子，围绕着石子的一圈圈波纹一样...」详细： 网页链接 @翁昕很欢乐 发表于知乎专栏。



链接

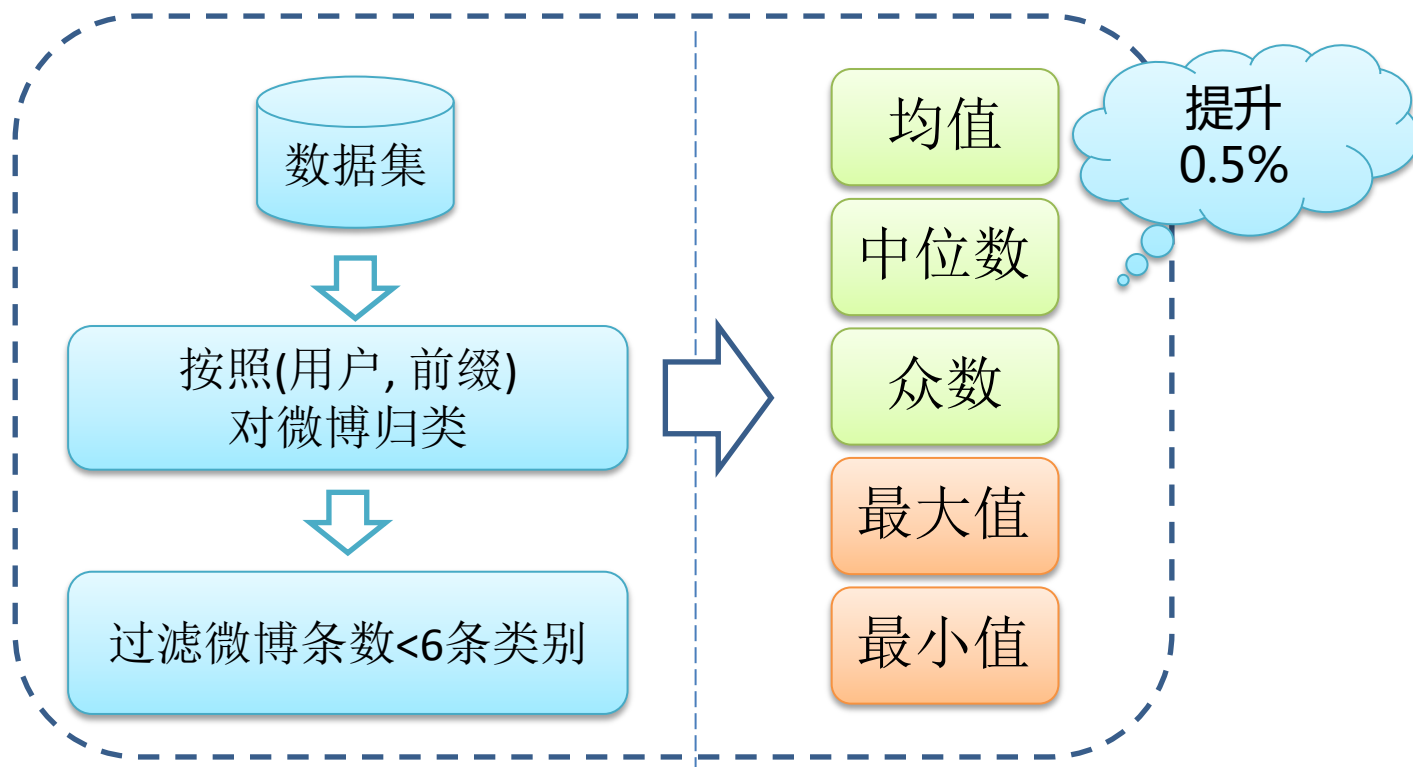
@其他用户

今天 11:09 来自 微博 weibo.com

关键词：分享、抽奖、中奖、转发、红包

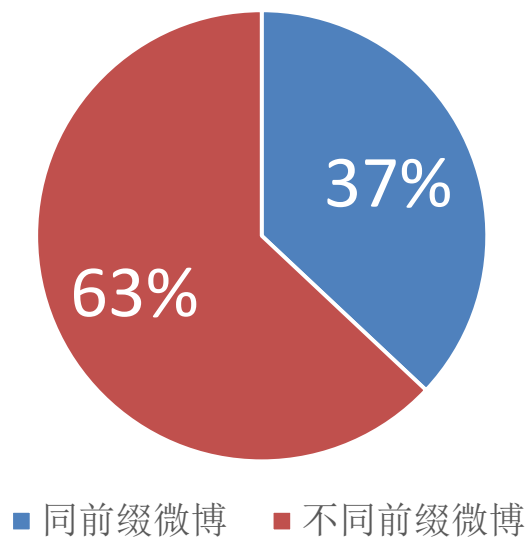
用户相同前缀微博统计特征

把具有相同前缀（前四个字相同）的同一用户微博看成同一类微博，统计该类微博互动数的**均值**、**中位数**、**众数**、**最大值**、**最小值**。

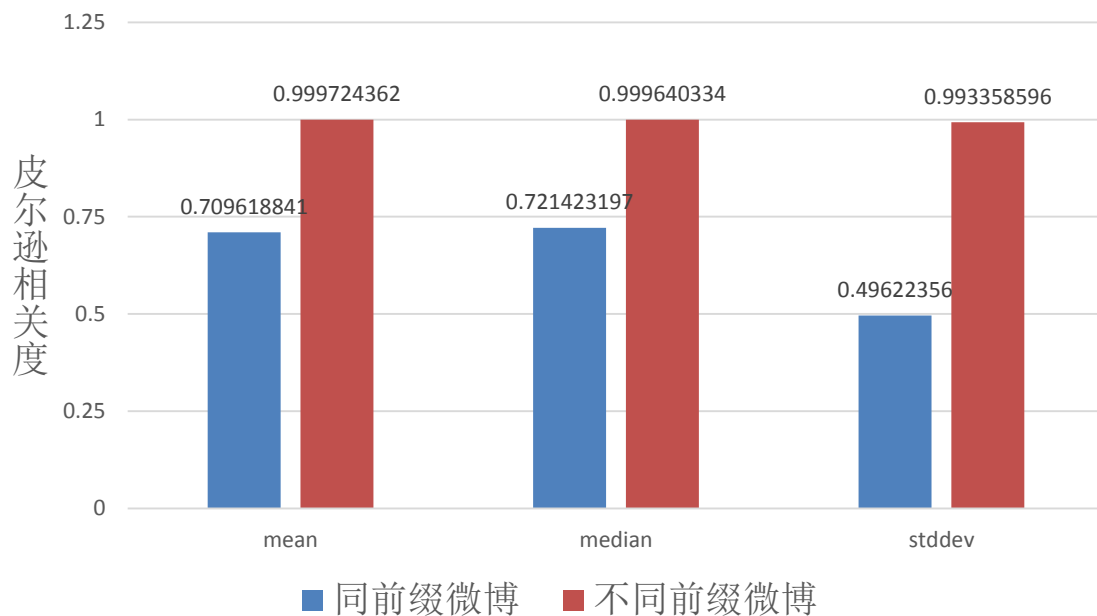


数据分析

同前缀微博占比

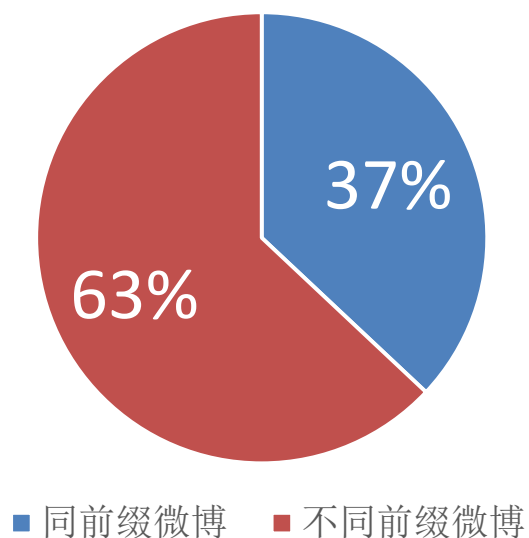


各统计量与所有微博的相关度

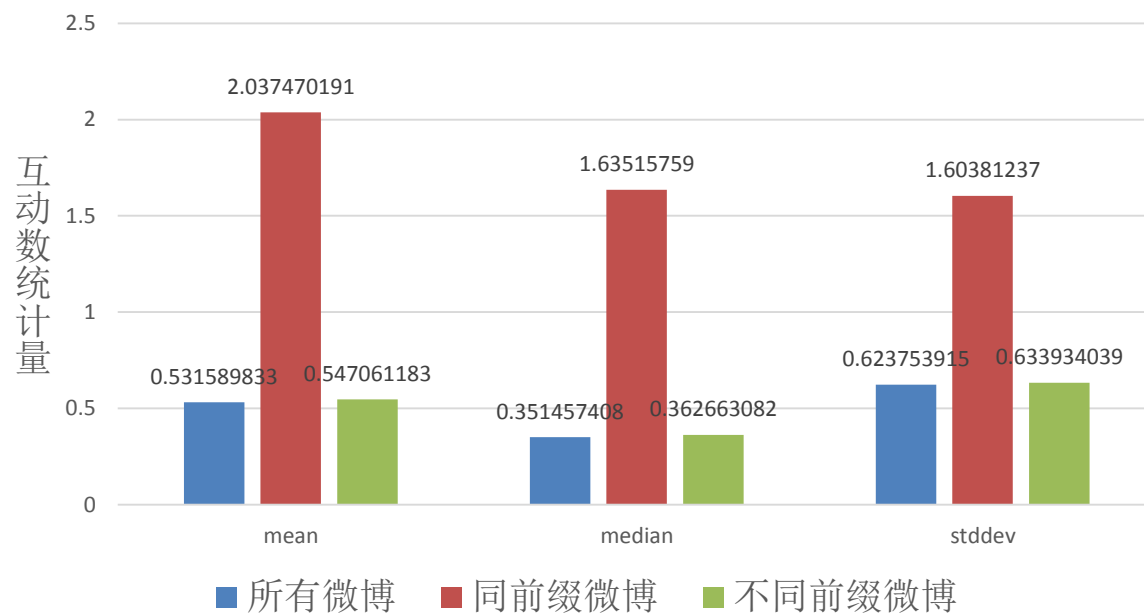


数据分析

同前缀微博占比

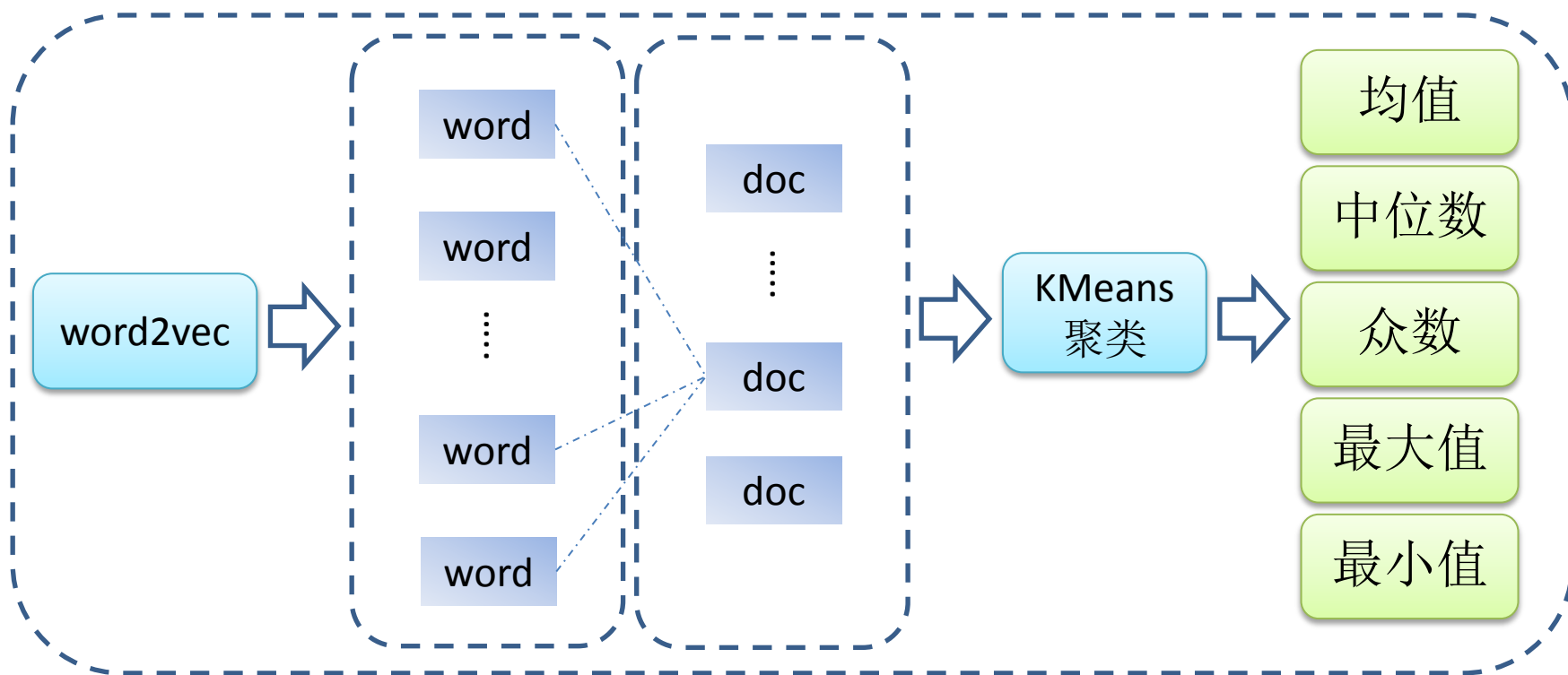


互动数统计量对比



word2vec聚类特征

物以类聚，人以群分。内容相似的微博收到的行为也是高度相似的。这是区别于用户的另外一个不可忽视的巨大挖掘空间。



word2vec聚类特征

正源地产 “在建住宅智能三星级实施计划楼盘” “北京地产资信20强”、“全国房地产综合品质创新奖”等荣誉。

<http://t.cn/z89msu>

从大连、北京，西安、南京到湖南宁乡……“尚峰尚水”作为正源地产城市山水豪宅典范。<http://t.cn/zRiDIB>

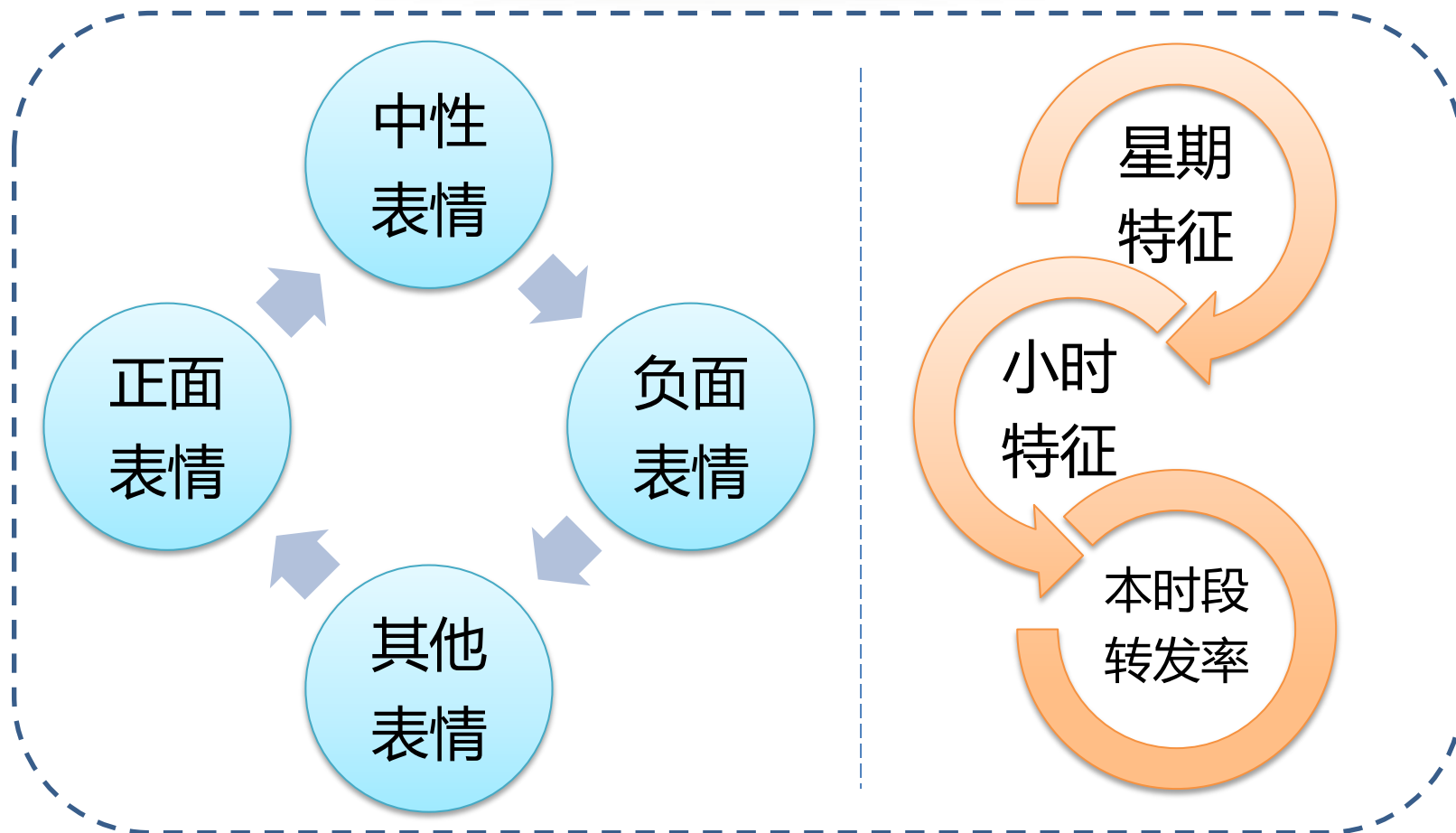
正源地产 “大连消费者购房首选楼盘” “商品房质量保修服务先进企业” “工程质量管理先进开发企业” “大连成品住宅示范项目” <http://t.cn/zRie2s>

欧洲央行执委普雷特：欧洲央行准备好调整资产负债表的规模与干预措施。

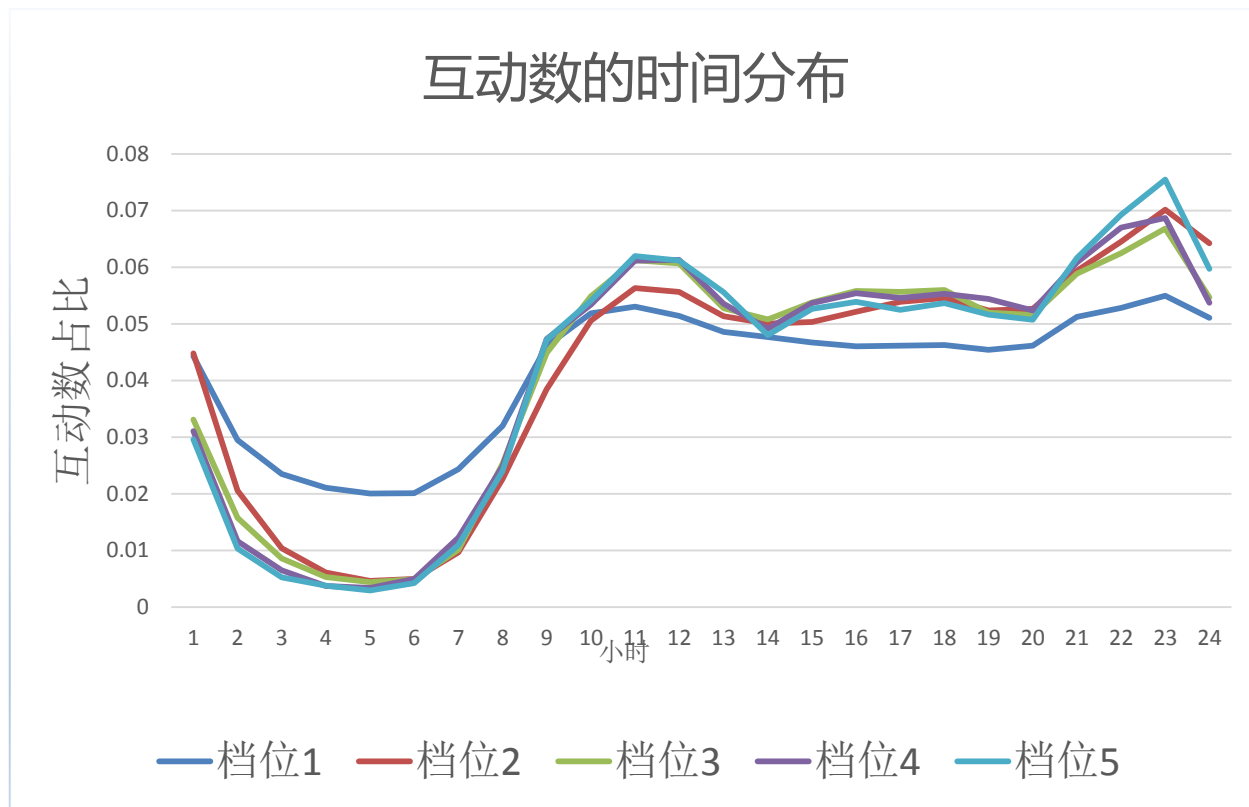
解套固然是投资者必须掌握的基本功，但投资者更应该把精力放在套牢之前，雪豹团队帮你提高分析技巧和买卖水平，尽可能减少被套牢的次数...

在高风险的外汇市场上投资投机买卖外汇，其实最重要的就是一开始的买入。当投资者回顾自己过去的外汇投资生涯中，都会发现所有的错误都是先犯
<http://t.cn/RzbNjmQ>

表情&时间 特征



表情&时间 特征



微博在各档位的概率

ONE

词袋模型假设

词项满足多项式分布。

TWO

**统计每个词
在各个档位的概率**

预处理，去除停用词

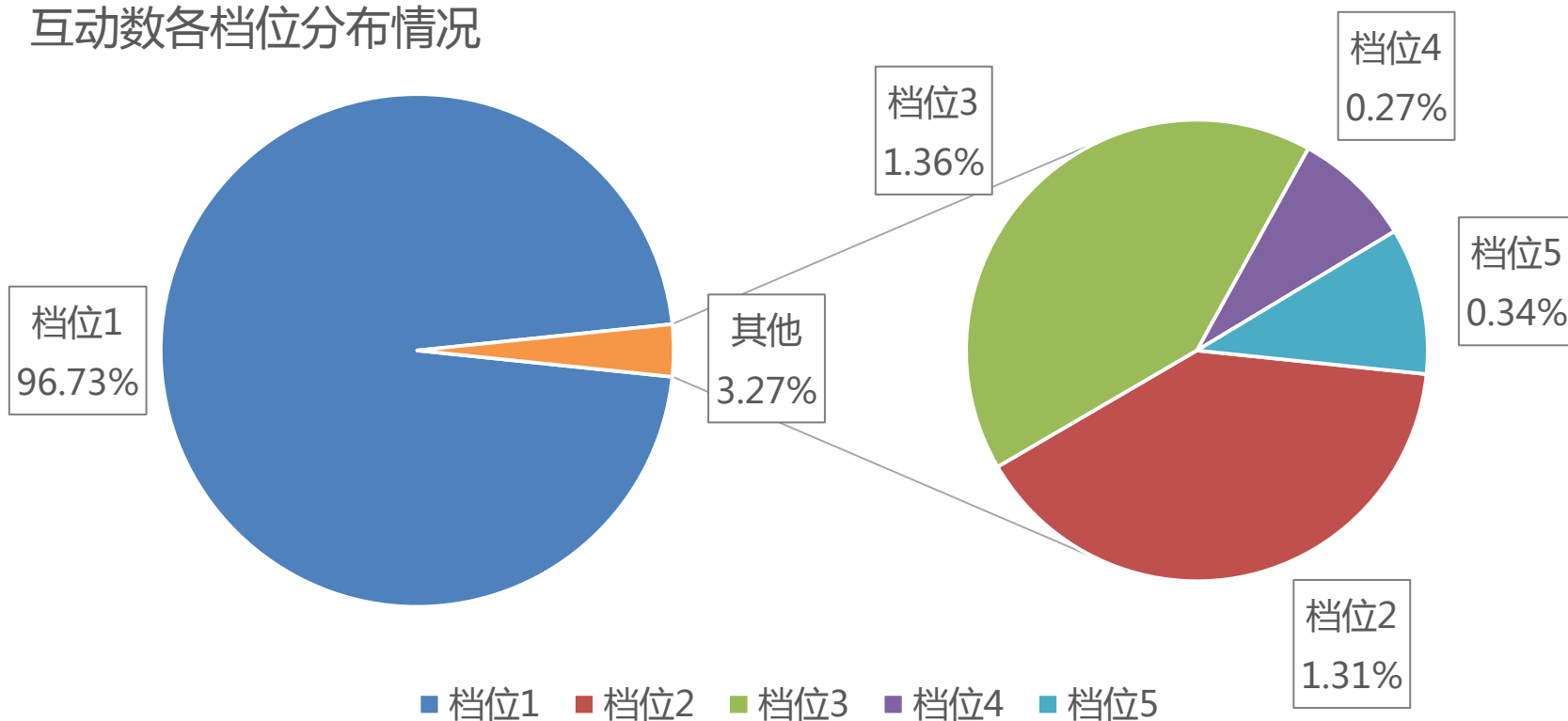
THREE

**计算博文
在各个档位
的概率分布**

贝叶斯原理，设定先验分布

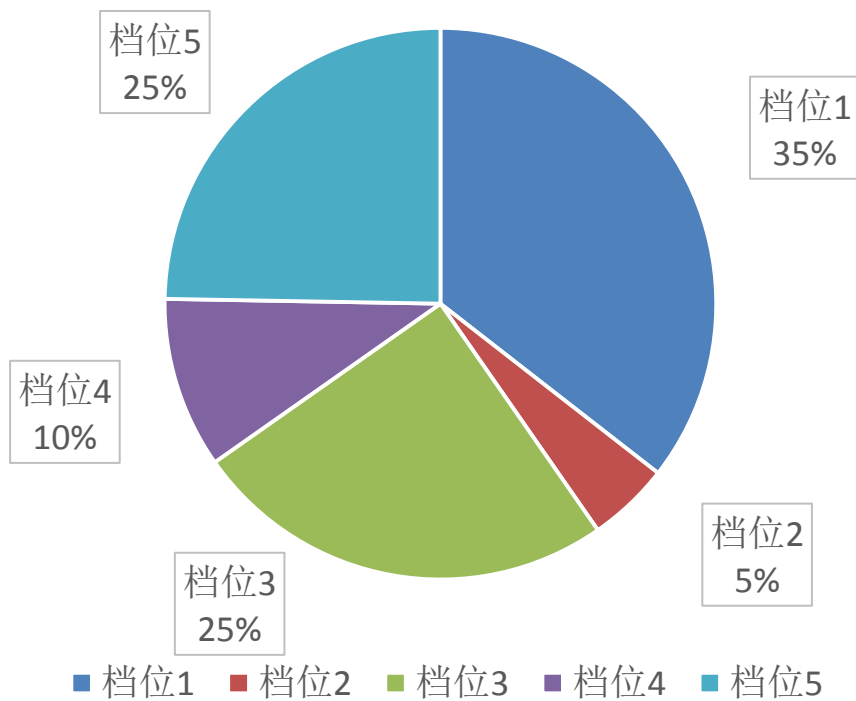
数据分析

互动数各档位分布情况

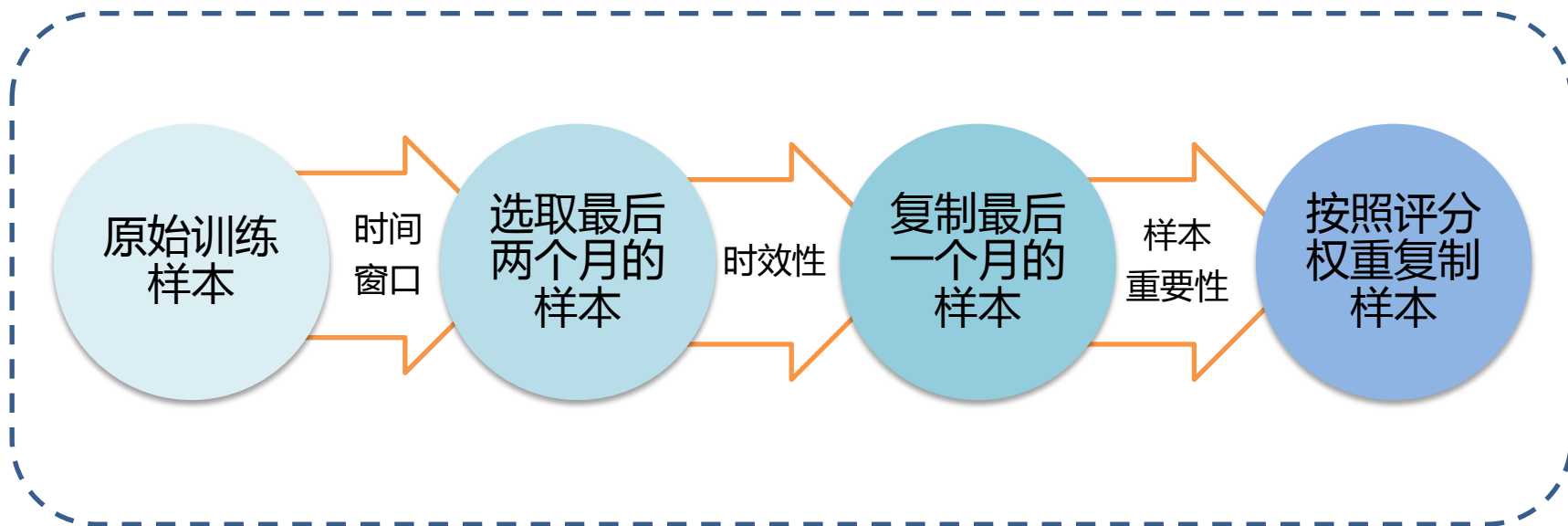


数据分析

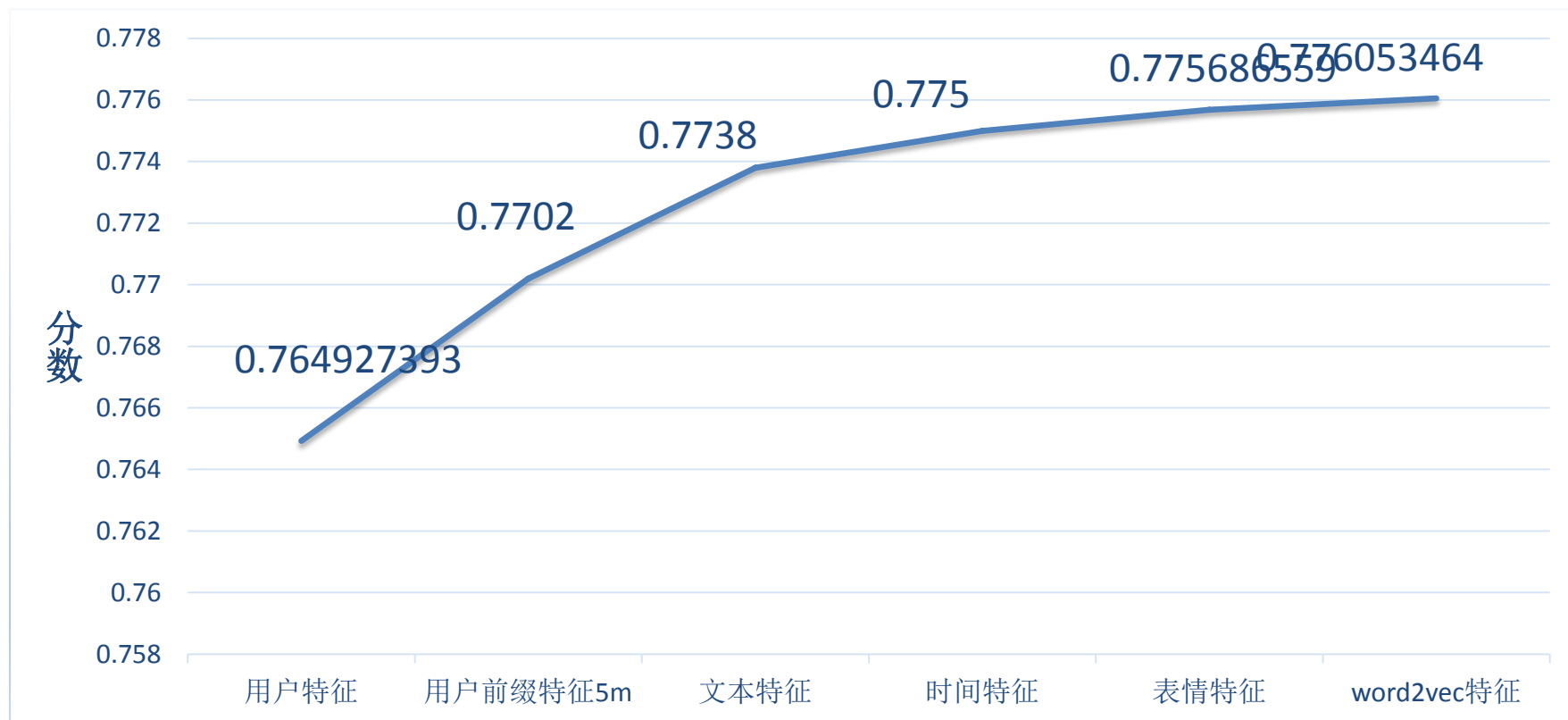
图表标题



训练样本的构造



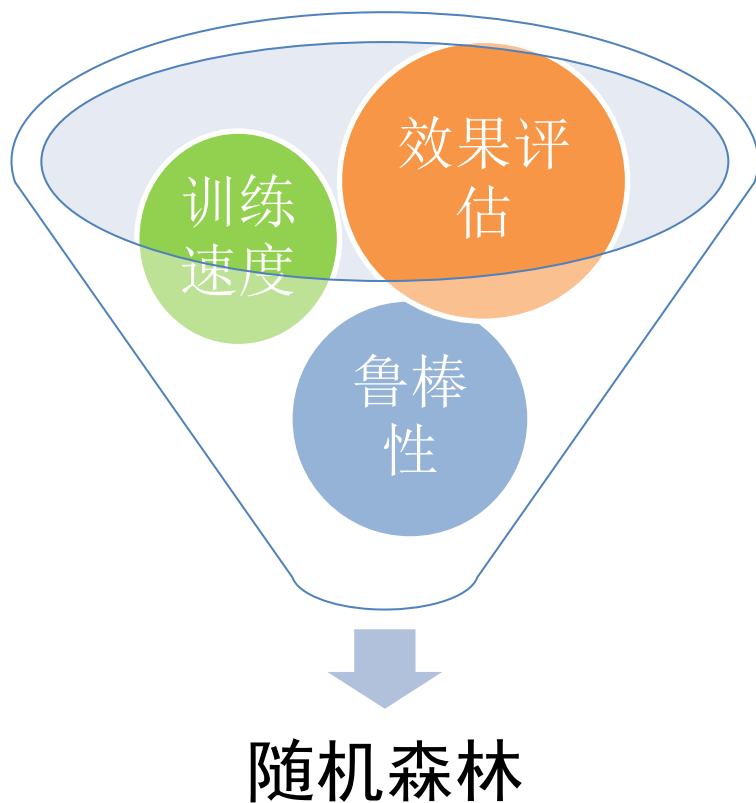
线下成绩





模型融合

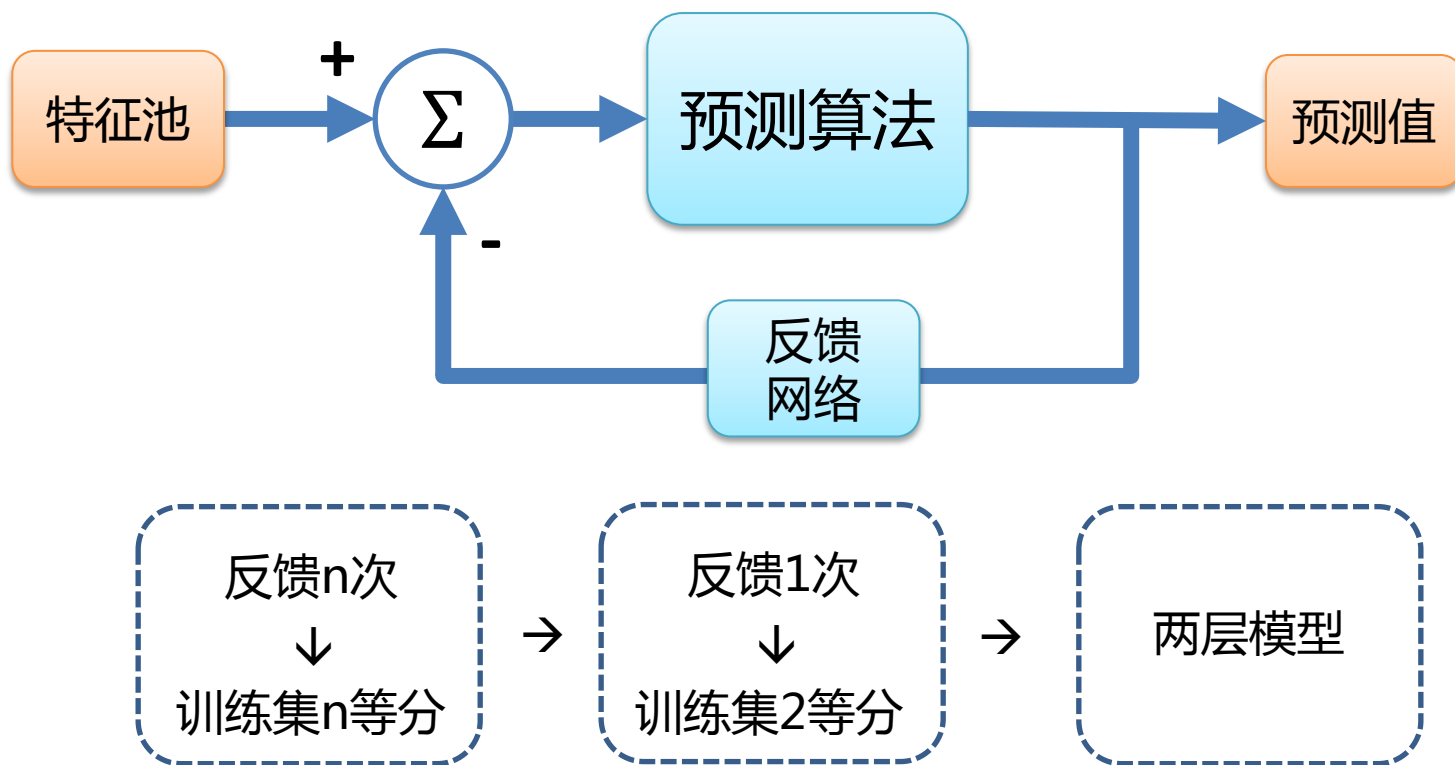
单模型选择



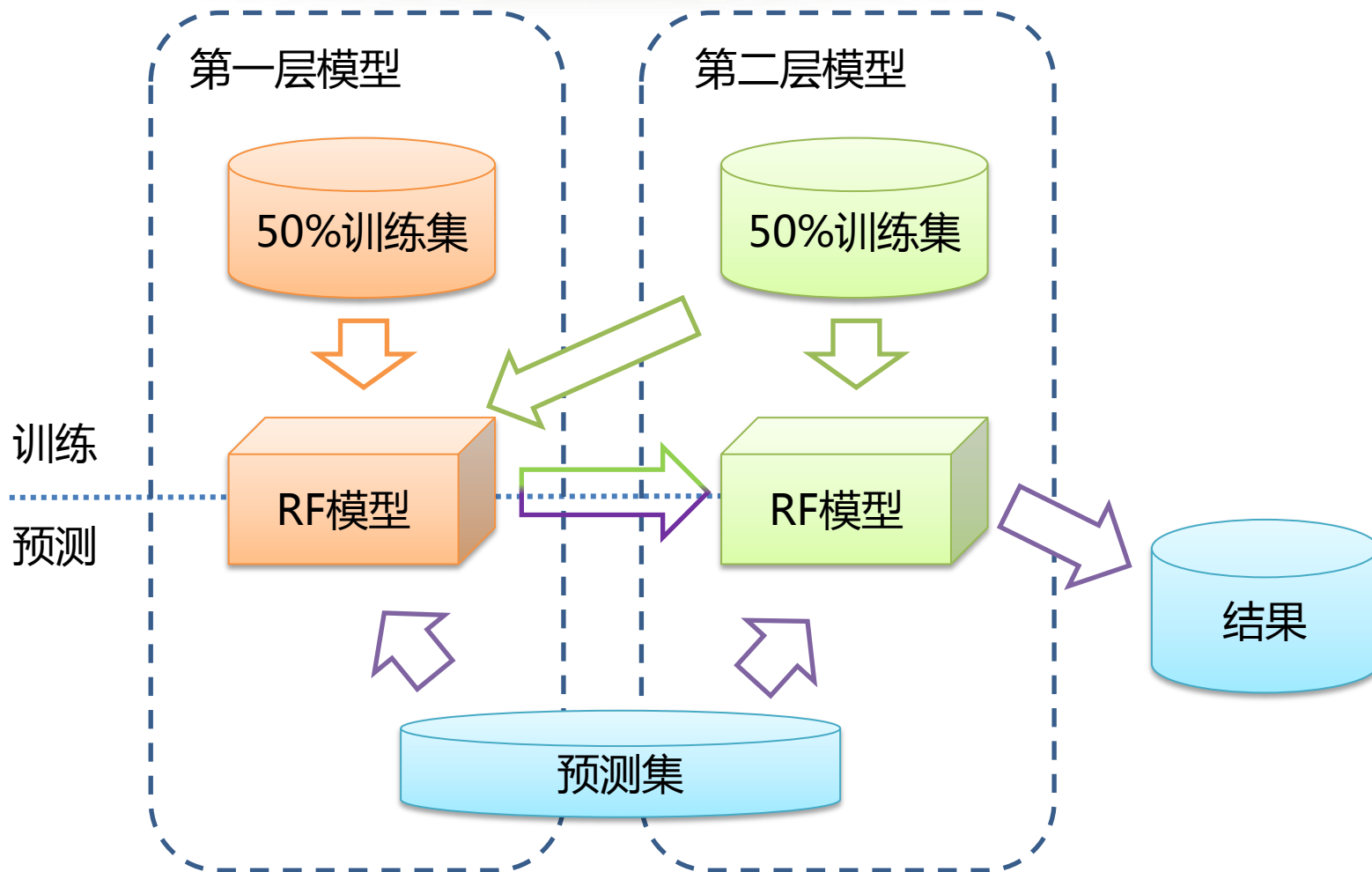
- 效果 : $RF > GBRT > LR$
- 速度 : $RF > LR > GBRT$
- 鲁棒 : $GBRT > RF > LR$

多模型组合

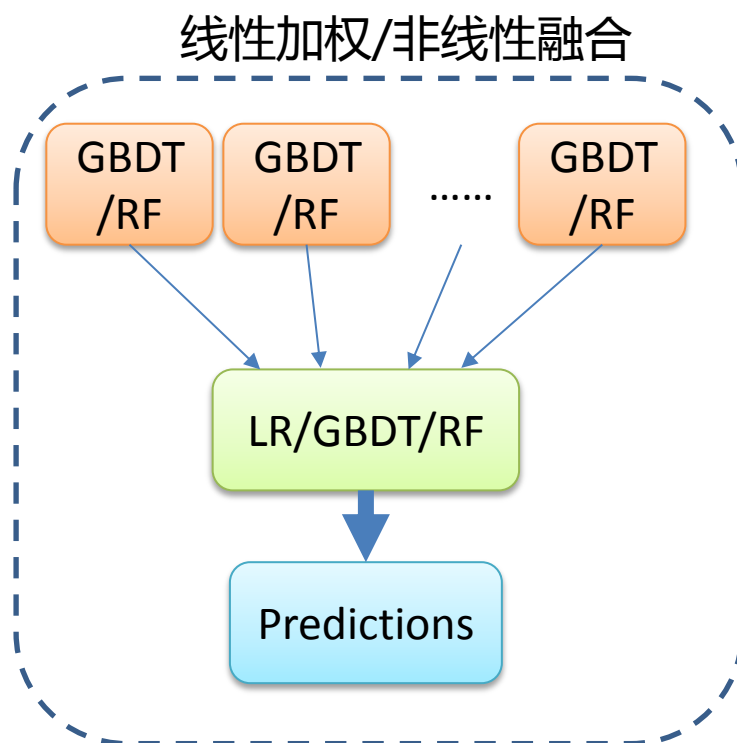
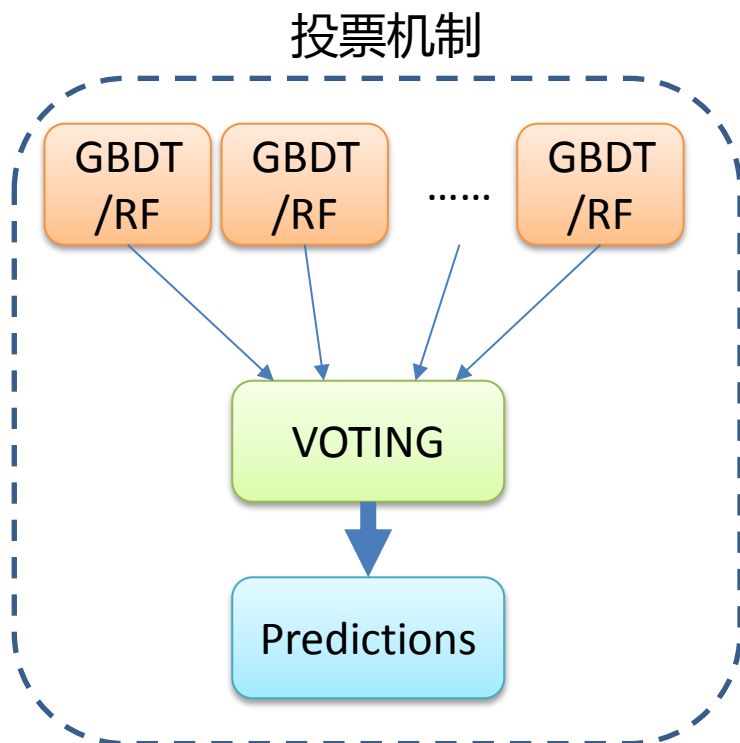
源自 闭环反馈控制算法



多模型组合



其他融合方法



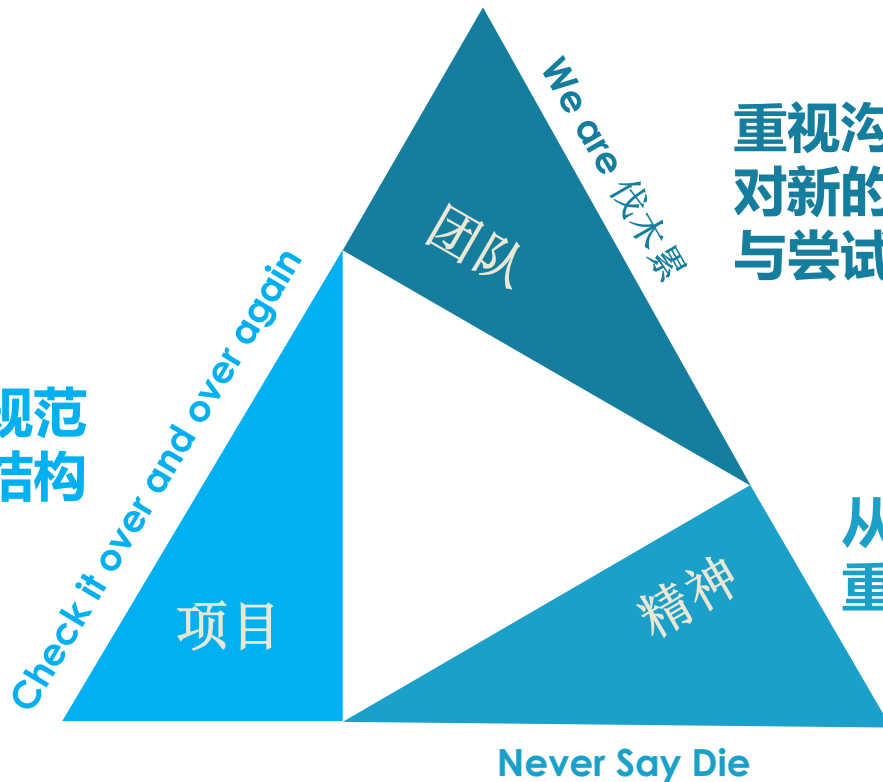
效果相当，BUT **模型复杂+速度慢** ← 线下训练 / 线下预测
线上训练 / 线上预测



总结思考

收获

多人同时开发：
严格遵守的命名规范
条理清晰的目录结构



重视沟通交流，
对新的想法及时记录
与尝试

从100名之外到Top5，
重视收获，永不言弃

致谢



感谢**阿里巴巴**和**新浪微博**提供的数据及平台



感谢**天池团队**的精心组织和默默付出



感谢一起坚持一起奋斗一起成长的小伙伴