2018中国高校计算机大赛
——大数据挑战赛

From: Ictry

目录

- 团队介绍
- 特征工程
- 树模型
- 神经网络模型
- 短期模型
- 模型融合

团队介绍

刘畅　　陈大浩

博士　　博士

哈尔滨工业大学

计算机科学与技术

模式识别与智能系统研究中心

指导教师: 刘鹏

赛题分析

“快手”新注册用户脱敏和采样后的数据，预测未来一段时间活跃的用户

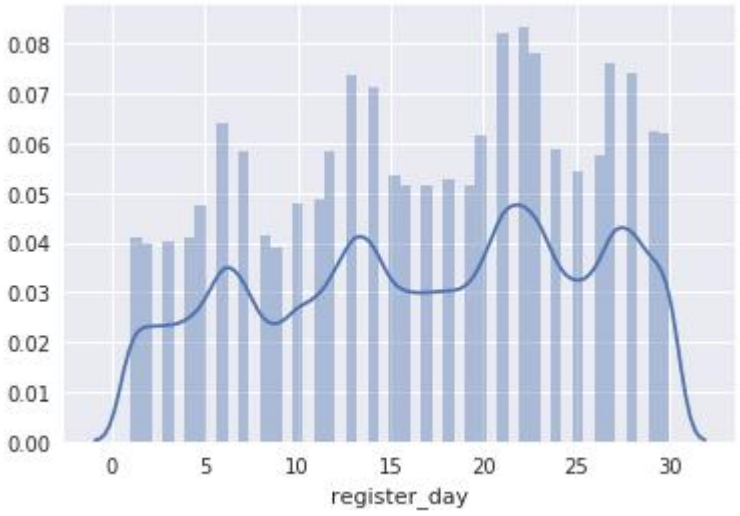
注册

启动

拍摄

行为

去除异常



31—37？

特点：注册时间越长越稳定

难点：节假日

滑窗法

	标签	用户	特征
测试	31—37	1—30	
训练	24—30	1—23	相对区间一致
	17—23	1—16	

对注册越久的用户预测越准

特征工程

普通特征：

- 三个表day的min, max, mean, std, nunique；
- max-min, last_gap, max_continous_days,diff;
- 均值特征：不同区间， 除以区间长度、注册时间、登陆次数、行为次数;
- 衰减：最近连续n天总的统计
- page, action_type, (page, action_type)分组， 对行为数量，交互视频、作者数量统计和计算比率、均值

特征工程

注册相关:

- 注册必登陆，绝大部分有行为
- 注册特征：
  - 注册类型，设备类型，注册时间
  - 设备类型：全局统计，分箱
  - 注册当天是否拍过视频，是否有行为，是否有点赞、关注等，是否访问过除发现外其他页面，个人主页
  - 因为滑窗导致数据中用户有重叠，要防止过拟合
- 剔除注册特征：
  - 注册本身是一种异常，去除每个用户注册当天数据再统计，比如登陆率

特征工程

用户粘性(业务特征):

- 自己对自己视频。。。别人对自己视频。。。
- 喜爱的视频
- 喜爱的作者
- 最大行为、观看、关注、点赞、转发
- 有多天交互的数量，最大天数，
- 有多次交互的数量，最大数量
- 每天观看视频中来自看过的作者的比例
- 取关、取赞

特征工程

高阶特征：

- 数据：
  - 按天统计数据
  - 交互用户数最多的video，author
    - Top(50,100,500)
    - 对用户统计，TFIDF
- 处理：
  - 降维
  - 岭回归
  - 01特征相乘（行为同时发生）

主要树模型

- lightgbm 线下890~891，A榜单模型0.9118+
- 数据量很大防止过拟合
  - num_leaves = 13,
  - max_depth = 4,
  - max_bin = 90,
  - min_data_in_leaf = 300
- 调参准则: 线下最大

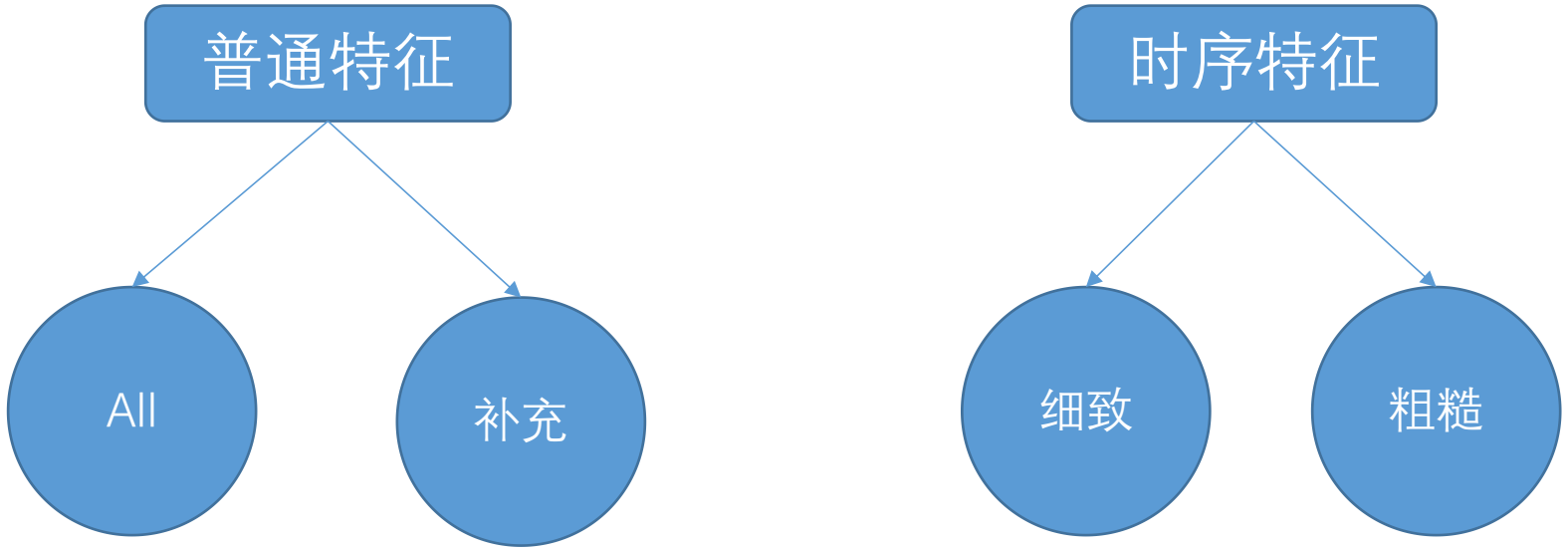
特征选择

- 线下验证增长
- 直觉
- 特征相关度
- lgb重要性
- 同组特征共同选择
- 多学习率

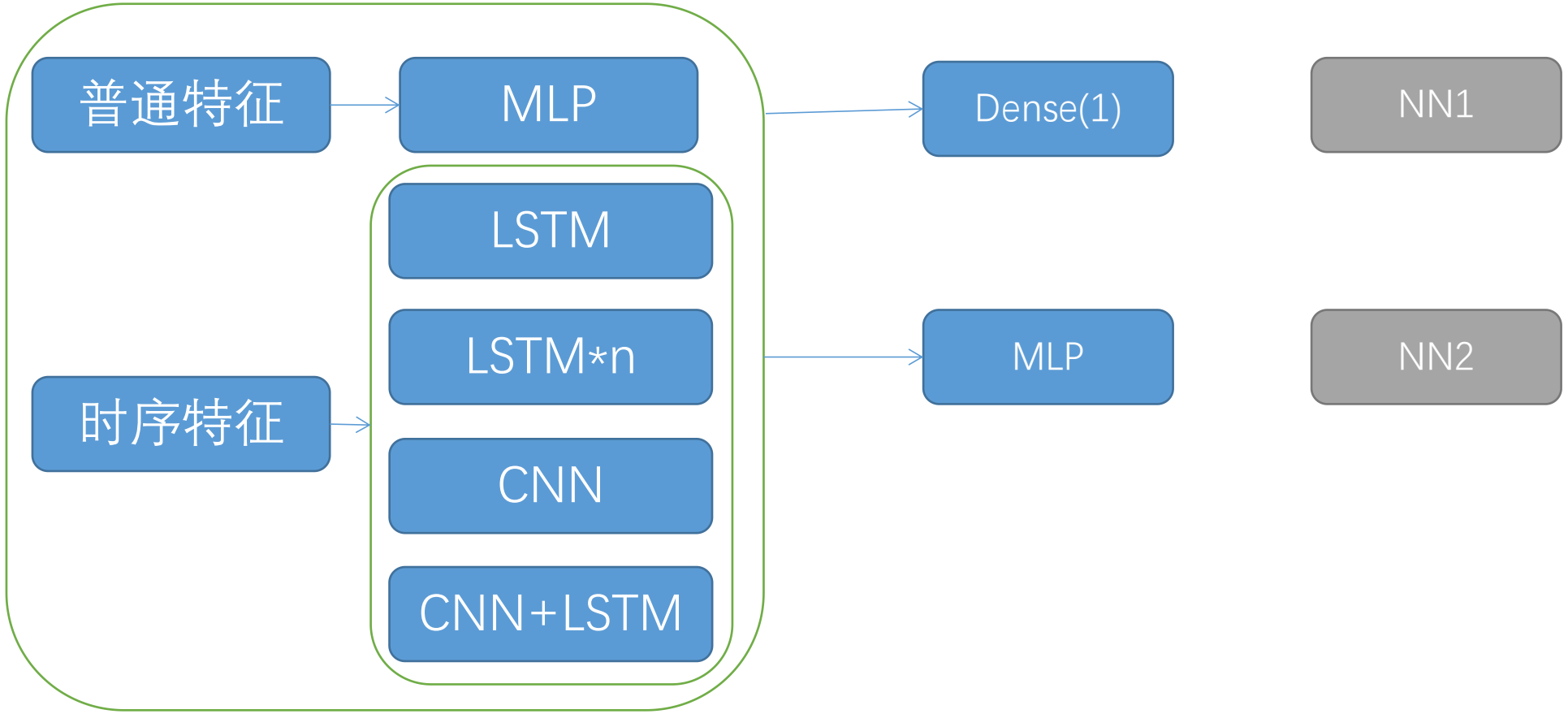
其他树模型

- xgboost　线下略低于lightgbm
- catboost　线下略低于xgboost，差异性比xgboost大
- 融合
  - 线下有万分位提升
  - 线上有十万分位提升

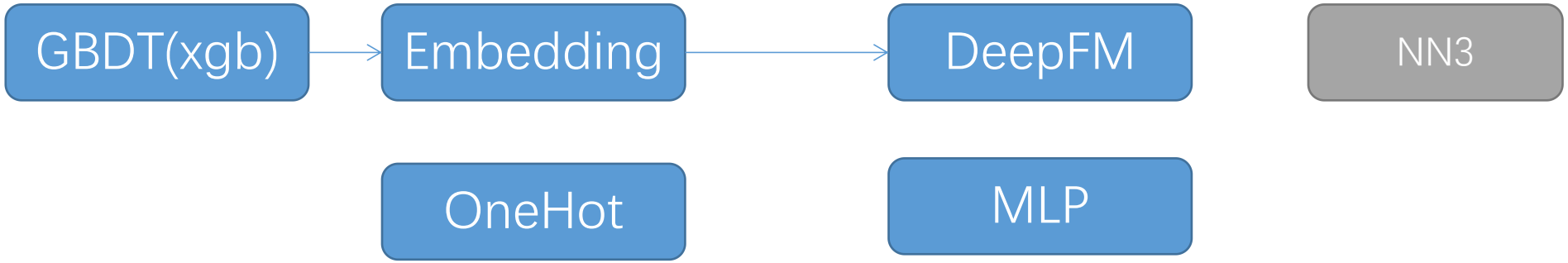
神经网络特征



神经网络



神经网络



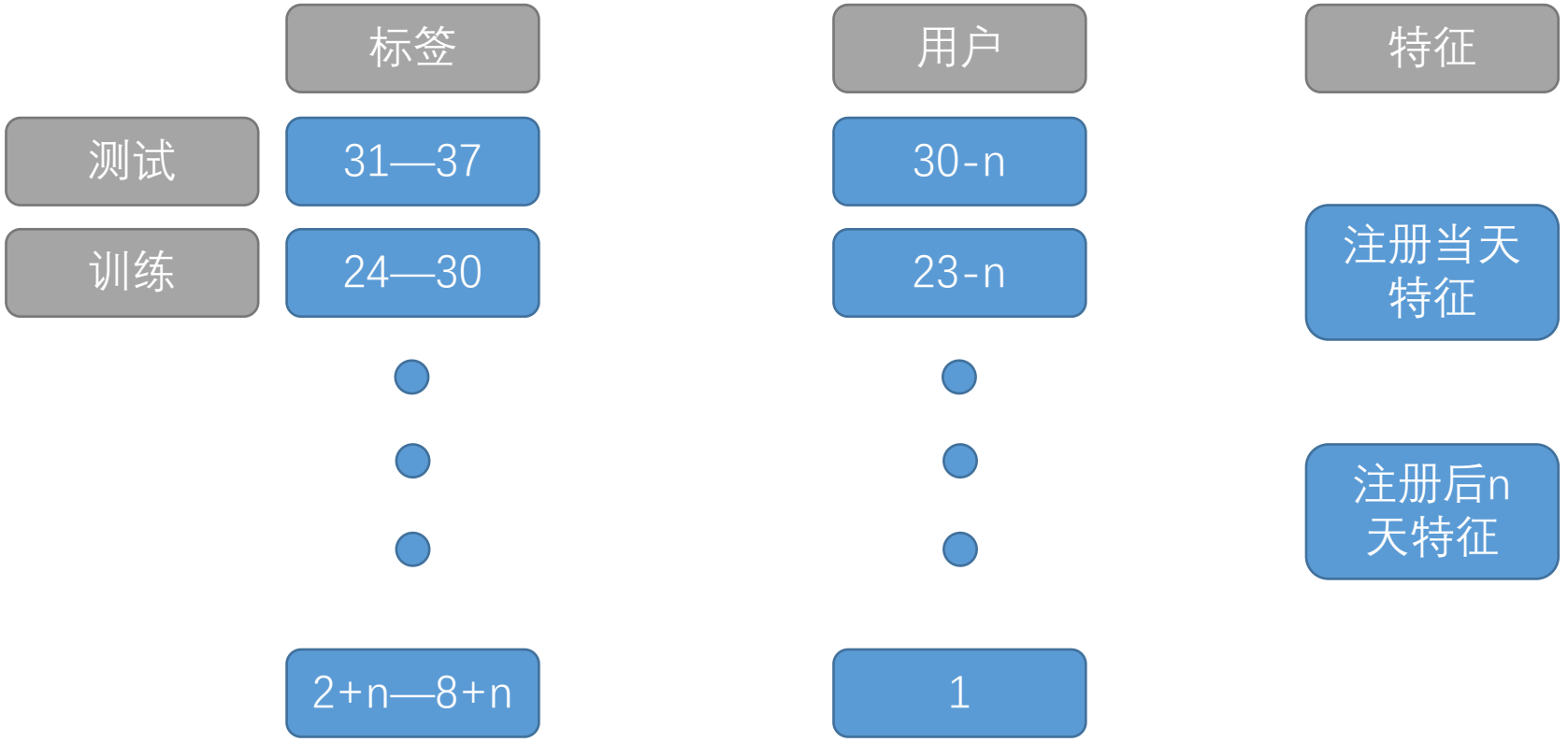
神经网络融合累计线下提升约一个千分点，
线上提升二点几个万分点

滑窗法不足

对近期注册的用户预测不准

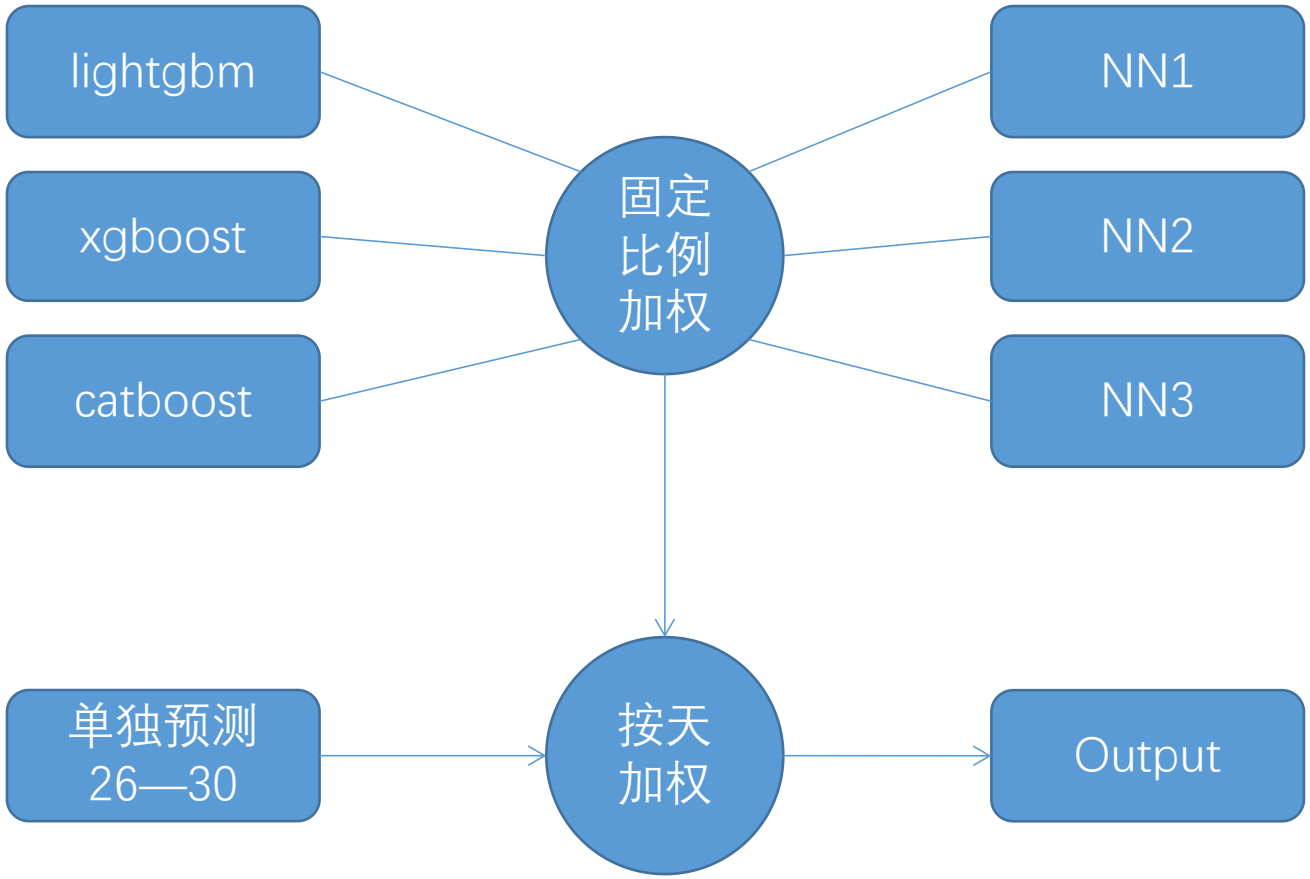
- 注册人数逐天增加，近期新用户很多
- 滑窗法对新用户预测不准，
  - 线下非常严重（18~23注册用户）
  - 线上还行（26~30注册用户）
- 解决方法
  - 预测用户注册n天后的七天是否会登陆

短期模型——单天滑窗预测



线下增益：两个千分点
线上增益：一个万分点

模型融合



最终线上成绩：
A: 0.91216145
B: 0.91313958

总结

- 注册特征
- 特征选择
- 合理的模型
- 模型融合
- 特别尝试


清華大學
Tsinghua University
快手
谢谢聆听