PAKDD2020 阿里巴巴智能运维算法大赛

队伍：fengyang95
答辩人：李元鹏
2020/4/16

目录

content

问题分析

需要自定义任务

二分类、回归、排序问题都可以尝试，需要取舍

样本不平衡

所有数据共5600万余条记录，
其中正样本3.8万余条
负样本5600万余条
正负样本比例约为1:1500

容易过拟合

特征噪声多
二分类问题的分类界限不明显

方案介绍

数据处理

pyspark处理原始数据

根据logs数据和fault_tag数据打标签

特征构造

原始特征

窗口统计特征

组合特征

模型

对负样本降采样

self-paced ensemble

LightGBM

方案介绍

特征构造

原始特征

smart_7raw
smart_7_normalized
smart_9raw
smart_9_normalized
smart_4raw
smart_12raw
smart_192raw
smart_193raw 等

窗口统计特征

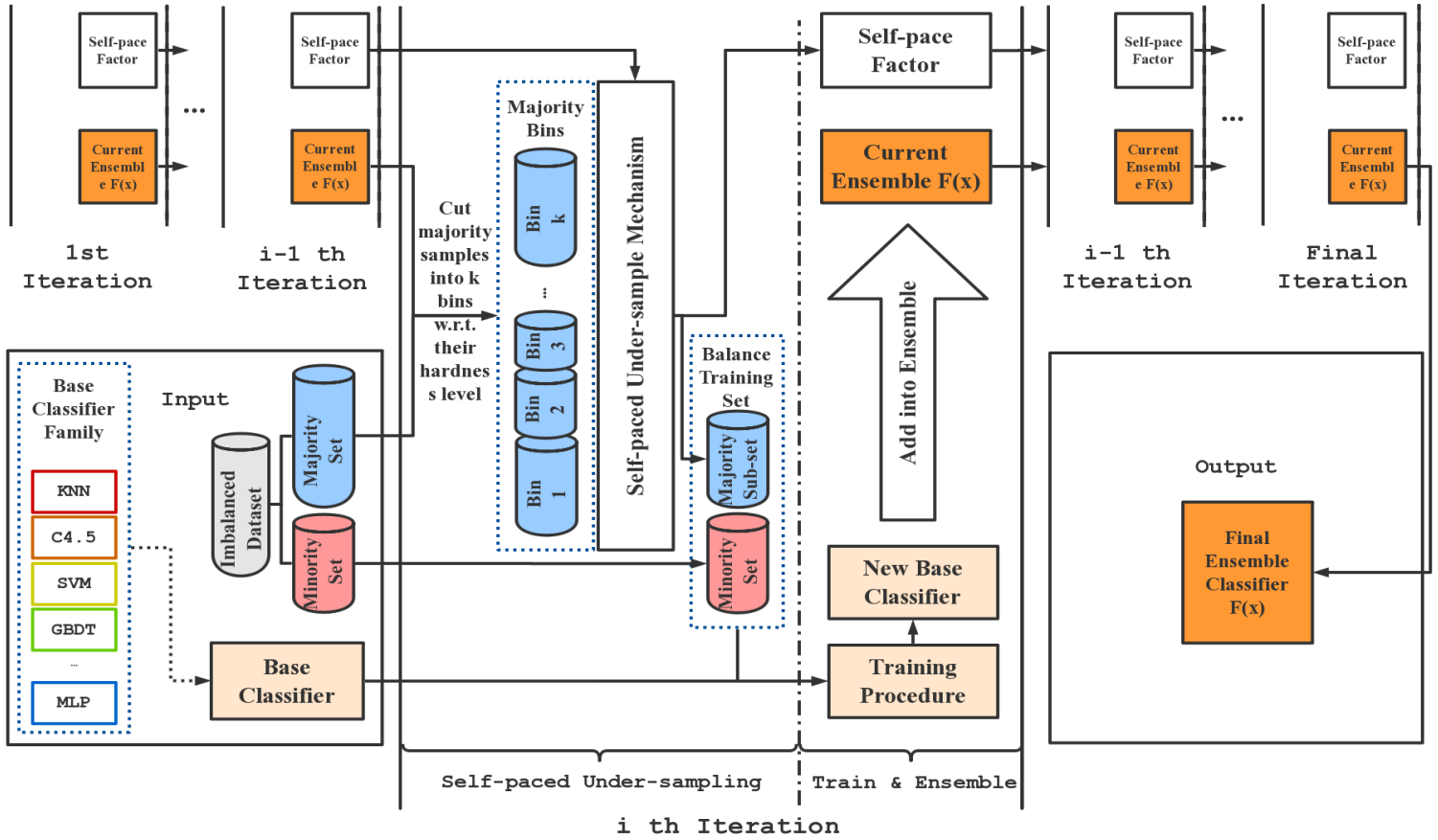
smart_1_normalized
smart_5raw
smart_190raw
smart_191raw
smart_193raw等的7天内变化范围值以及方差

组合特征

smart_4raw
smart_5raw
smart_187raw
smart_191raw
smart_198raw的两两组合特征

方案介绍

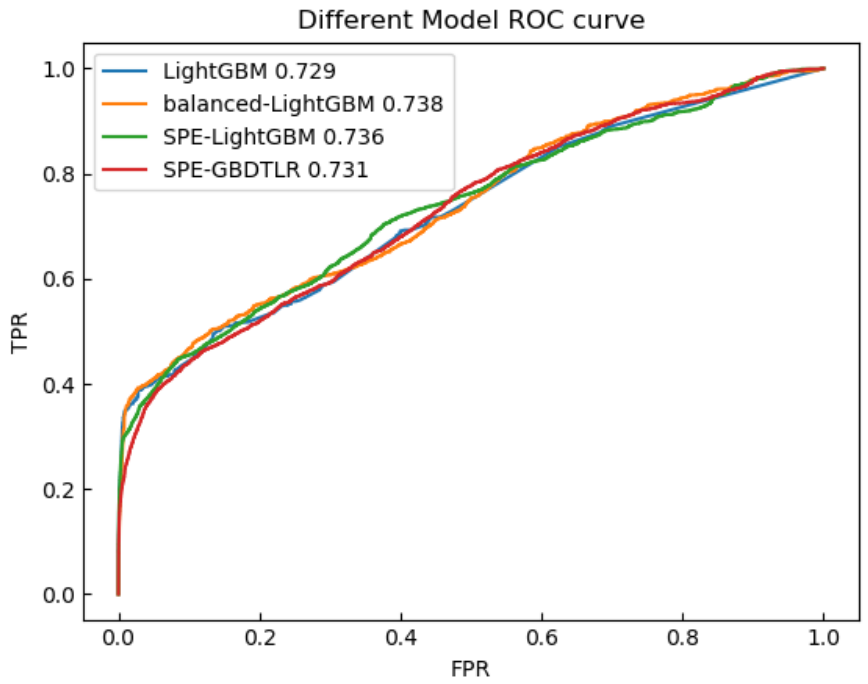
算法选择: self-paced ensemble

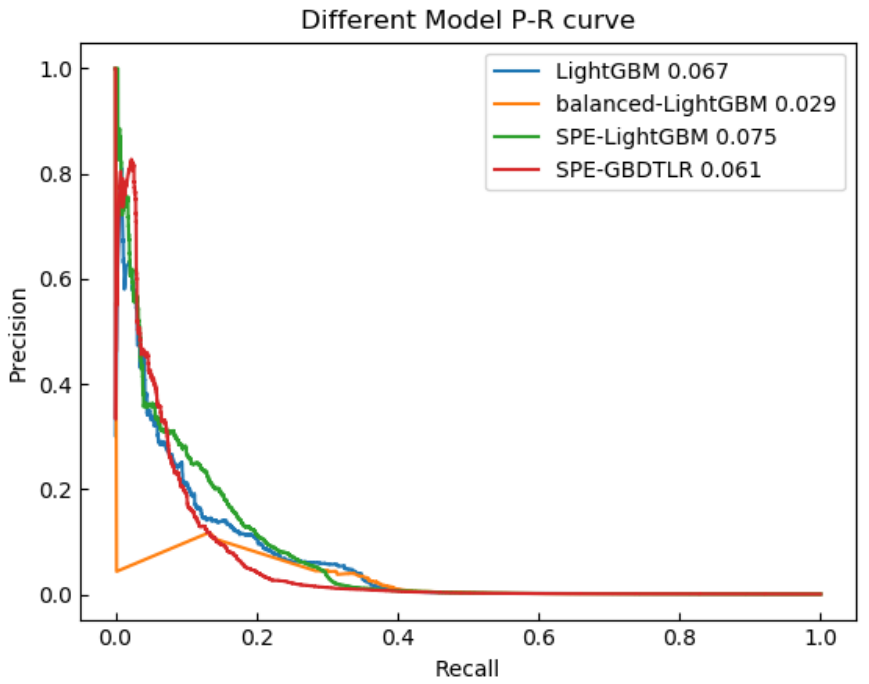


方案介绍

算法选择对比

- LightGBM及balanced LightGBM
- self-paced ensemble+(LightGBM+LR)
- **self-paced ensemble + LightGBM**



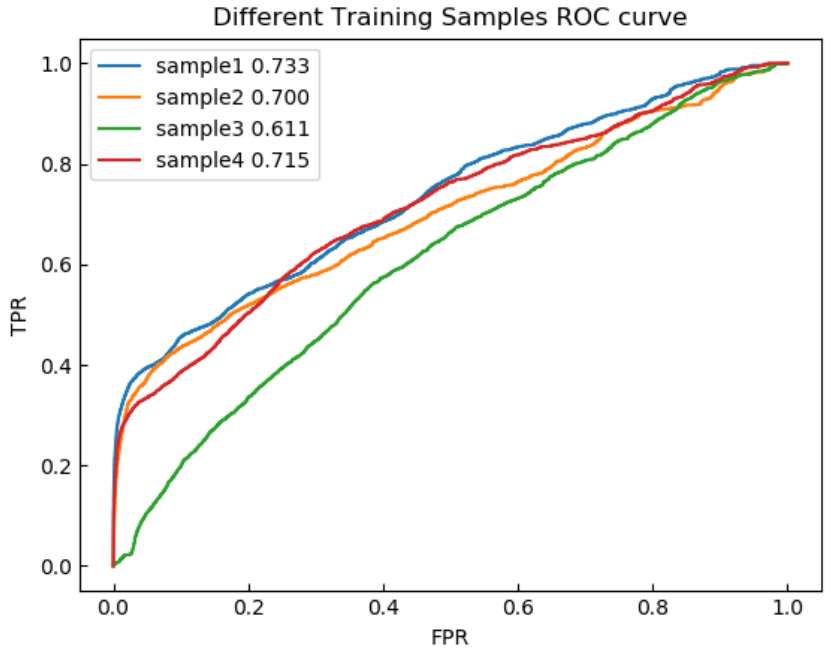


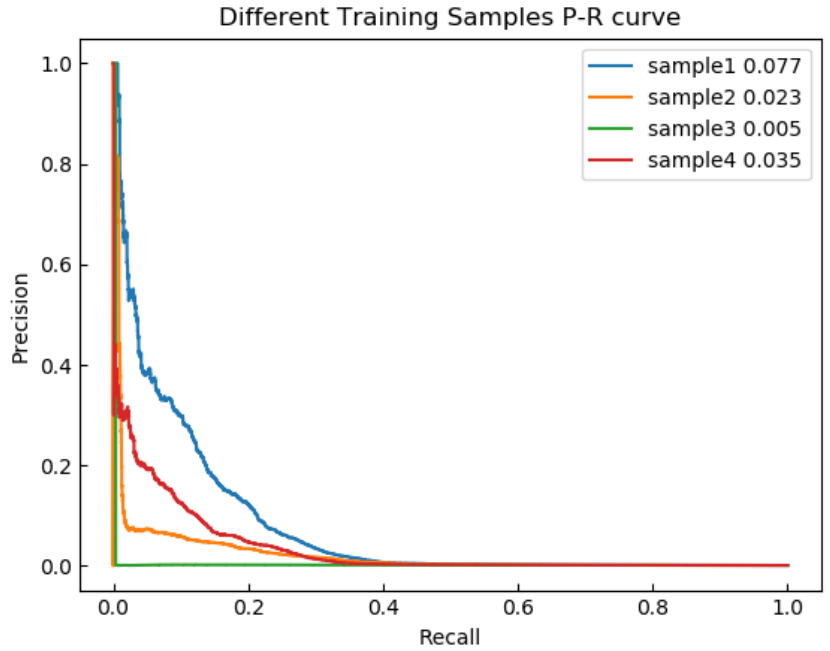
方案介绍

训练样本选取

- 1. **负样本按月抽样（每月抽2条记录）+全部正样本**
- 2. 两个月的负样本+全部正样本
- 3. 有损坏记录的盘对应的所有样本
- 4. 两个月的所有数据

区别只在于负样本





特点&不足

特点

pyspark

处理速度快

特征简单

原始特征+统计窗口

特征+组合特征

self-paced ensemble

方法新 训练速度快

不足

EDA不充分

没有挖掘出特别有

意义的额外特征

特征选择做得不好

特征选择方法不当、

存在噪声、过拟合问

题依旧没有解决

后处理欠缺

没有找到特别有

用的后处理方法

。

展望&总结

尝试挖掘更多特征

特征选择

转换思路

试一下异常检测，利用关键指标的时间序列

模型的参数优化

线下指标的选择

多交流

感谢观看！欢迎指正！

答辩人：李元鹏