用户人品预测大赛--wifi队--竞赛分享

•发布于 2016-03-24 14:59 •作者 DataCastle运营 (/user/5451) •37 次浏览 •来自 微额借款用户人品预测大赛 (/?tab=148)

参赛队队名：wifi

竞赛报告书

一、参赛作品概述
算法概述：

本问题是要对微额借款人群的信用度进行评价。从预测目标为按时还钱与不按时还钱角度来看这是一个分类问题；但是如果从还款概率大小的角度来看，这又是一个排序的问题。本算法结合 xgboost,gbdt,FM,NN 等算法的优势，将分类问题与回归问题相结合。最后利用线性加权融合法对每个用户在不同模型下的预测位次进行加权。通过实验发现两两进行分层的融合会比简单的直接线性加权融合具有更好的效果。灵活运用融合技术可以发挥各个算法的长处，然后提高预测结果的质量。同时算法还有对数据的预处理。针对不同模型的特性进行了不同的处理。比如 gbdt 对于高维稀疏数据效果较差，我们对其进行了特征选择，选择信息增益率最高的 600 维特征。对于 FM 算法我们队离散类型数据进行了 one hot encoding。

关键技术：

xgboost，gbdt，NN 分类模型的训练与调参。

FM 模型的回归预测。

多模型的融合。

二、参赛作品技术路线
算法基本思路

1. 训练基本分类器（ xgboost 和 gbdt ）

通过参数调优使得 xgboost，gbdt 获得最佳线上评分。其中 xgboost 线上得分为 0.7178，gbdt 线上得分为 0.6949。

2. 利用过拟合的 xgboost 为 train 数据排序

利用在训练集上 AUC 接近 1 的 xgboost 模型，得出 train 数据中所有人的得分值。对其排序得出所有 train 数据所有人的位次，并将该位次规范化到 0，1 区间。

3. 利用 xgboost 对 train 数据的排序，训练二阶基本分类器（ FM ， NN （神经网络 ））由于只用 FM 和 NN 对训练数据进行训练，其得分值聚集在 0,1 附近，没有较好的区分度，对于利用 AUC 作为评分值难以获得良好的效果。于是我们利用第二步中得出的规范化的人员位次作为 target。将分类问题转化为回归问题。由此得到 FM 的线上测评结果为 0.702NN 的线上测评结果为 0.7。

4. 将四个基本分类器两两融合

首先将 xgboost 与 gbdt 得到的结果文件进行（0，1）区间规范化并将两个结果以 0.8：0.2 的比例进行均值融合，其结果为 temp1。Temp1 的线上评测得分为 0.72。然后将 temp1 与 FM 得到的结果文件进行（0，1）区间规范化，并将两个结果以 0.9：0.2 的比例均值融合，得到结果 temp2。temp2 的线上得分为 0.7212。最后将 temp2 与 NN 得到的结果文件进行（0，1）区间规范化，并将两个结果以 0.9：0.1 的比例进行均值融合，得到最终结果。最终结果线上评分为 0.7214

算法原理：

1.问题本质是一个排序问题。但是数据集给出的只有类别。于是我想到用 xgboost 训练一个过拟合的分类器。当 AUC 值接近 1 的时候，对于训练集近乎似一个完美分类器。因此，其概率值位次基本上可以作为对训练集的一个排名。

2.多模型算法可以比单个模型获得较大的效果提升。传统的线性加权模型由于两两加权之后其分布可能会偏斜，与第三个融合的时候因为两者分布不同，导致结果提升不明显。于是我们在每次加权之后加入了结果归一化，保证了要融合的两个模型分布相当。

三、作品总结

算法优势：

（1）可以很方便的与新的模型进行融合吗，具有较强的可扩展性。

（2）将分类与回归相结合，利用模型的差异性获得更好的融合效果。

可能的改进方向：

（1）并行化策略：单机性能较弱，训练时间太长。

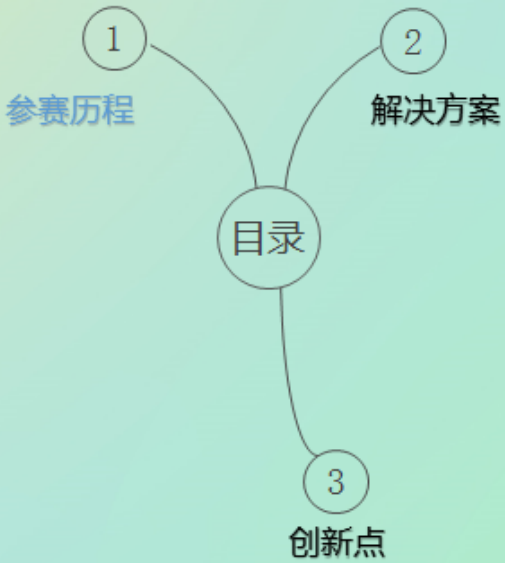
（2）参数的选择：更合理的参数选择可能会带来更好的预测结果。

（3）更好的排序策略：利用 xgboost 模型时，对于分类与排序运用了同样的参数，这样导致算法内部耦合教大，对样本的排序有一定的欠缺，可以使用更大的树的深度和树的数目获得更加过拟合的 xgboost 模型用来排序。也可以引入排序学习的方式，获得更好的排序模型。

用户人品预测大赛

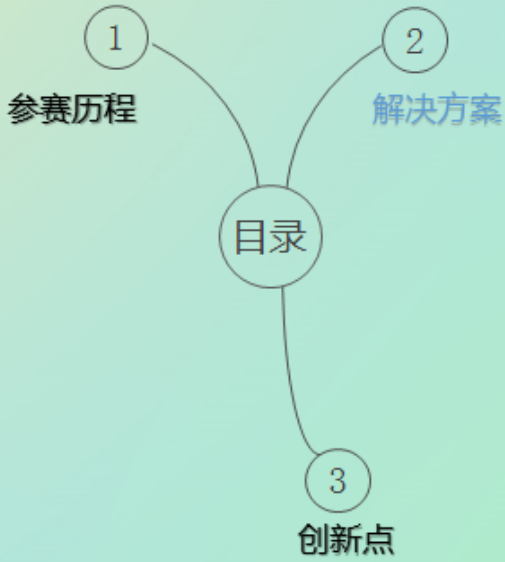
队名：wifi

2016-1-9



1 参赛历程

- ◆单个分类模型参数调节（xgboost,gbdt,NN）
- ◆利用FM算法对待测目标的人品进行排序
- ◆多个模型进行融合
- ◆模型融合权重调优



2 解决方案——比赛任务

- 数据：
  微额借款信用数据（训练集15000，测试集5000，特征维数1138）
- 任务：
  构建分类器：对用户人品状况进行预测
- 评价标准：
  AUC值（Area Under Curve）
- 工具：python+sklearn+xgboost

2 解决方案——数据特点

- 正负样本比例：9:1
- 数据维数高：1138
- 数据有类别型数据和数值型数据
- 样本缺失值多：横向看：缺失维度大于100维的样例占全部数据的1/3
  纵向看：有189个维度缺失样例数目超过1/5
- 特征含义不明

2 解决方案——问题建模

分类

回归

借款不还的概率

好人

坏人

训练集：特征 ——→人品好的概率排名

预测：训练集中的用户在训练集中的排名

2 解决方案——数据预处理

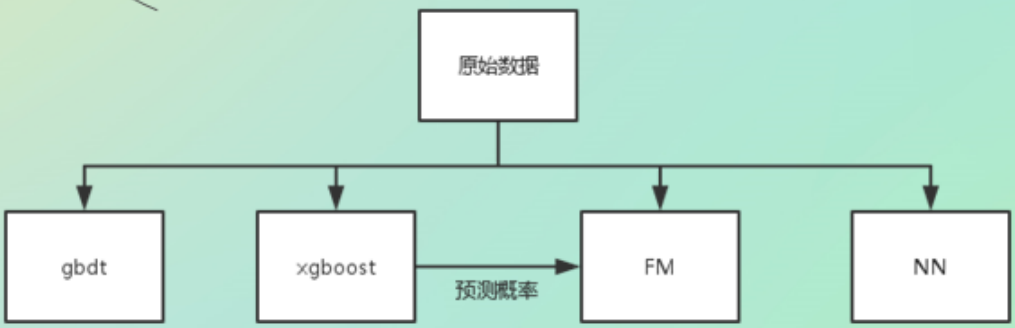
特征选择：
（1）过滤掉缺失值过多的特征
（2）过滤掉与分类目标相关性较弱的特征。
（3）单个特征训练，过滤掉预测结果很差的特征

降维

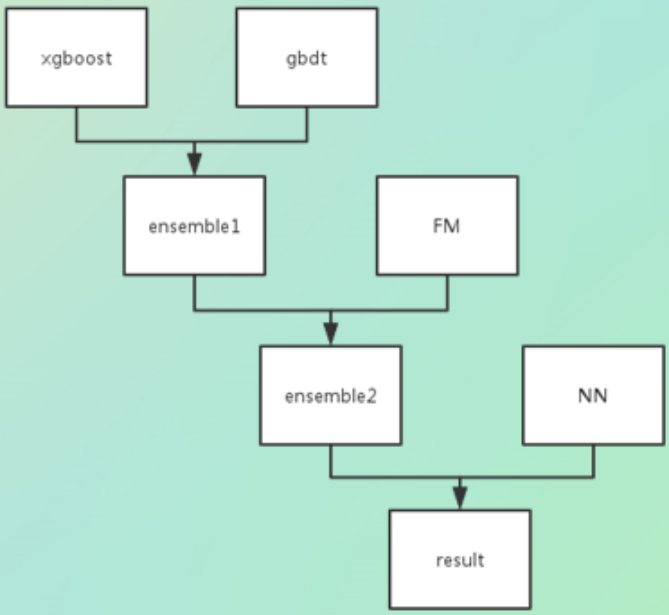
缺失值处理：
（1）对于数值类型缺失值，利用同类样本的平均值填充
（2）对于离散类型缺失值，利用同类样本的众数填充

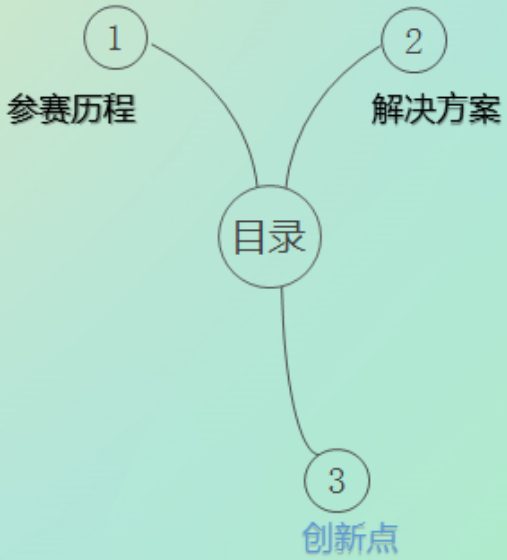
对于类别型数据：
转换为One hot encoding使其可以适应不同的算法；
提高算法的适应性

2 解决方案——模型训练


原始数据
gbdt
xgboost
预测概率
FM
NN

2 解决方案——模型融合



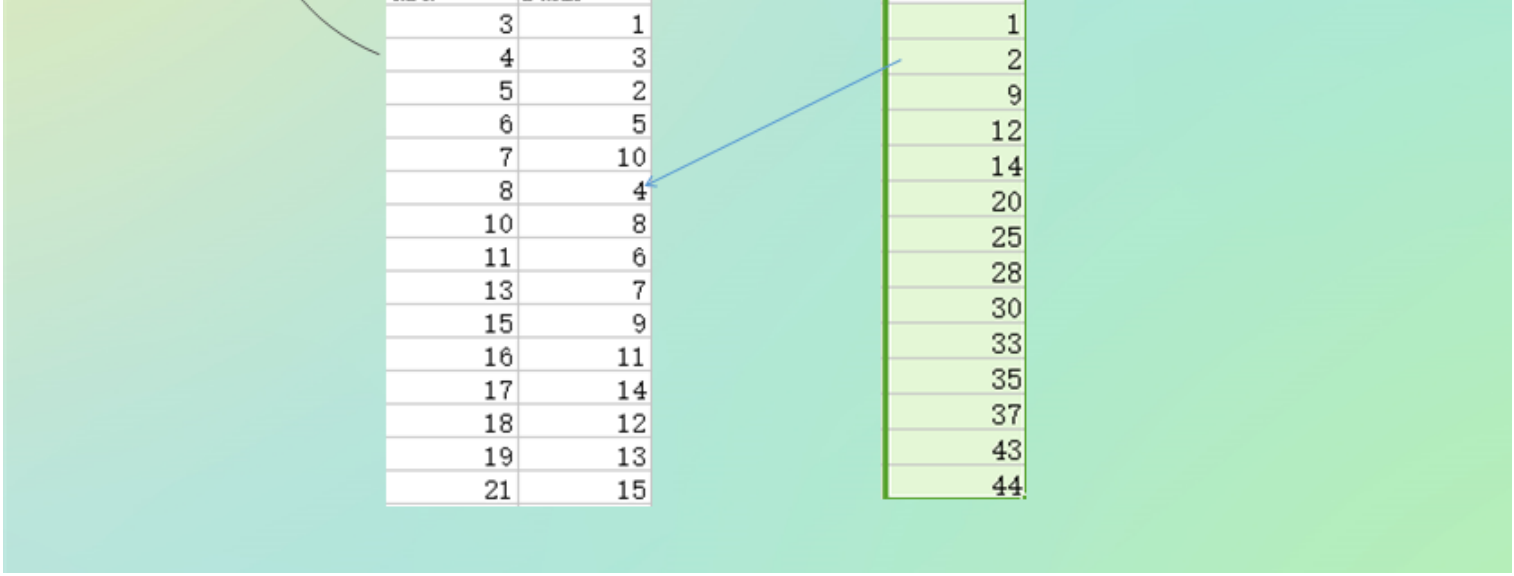


3 创新点-分类模型与回归模型的结合

train

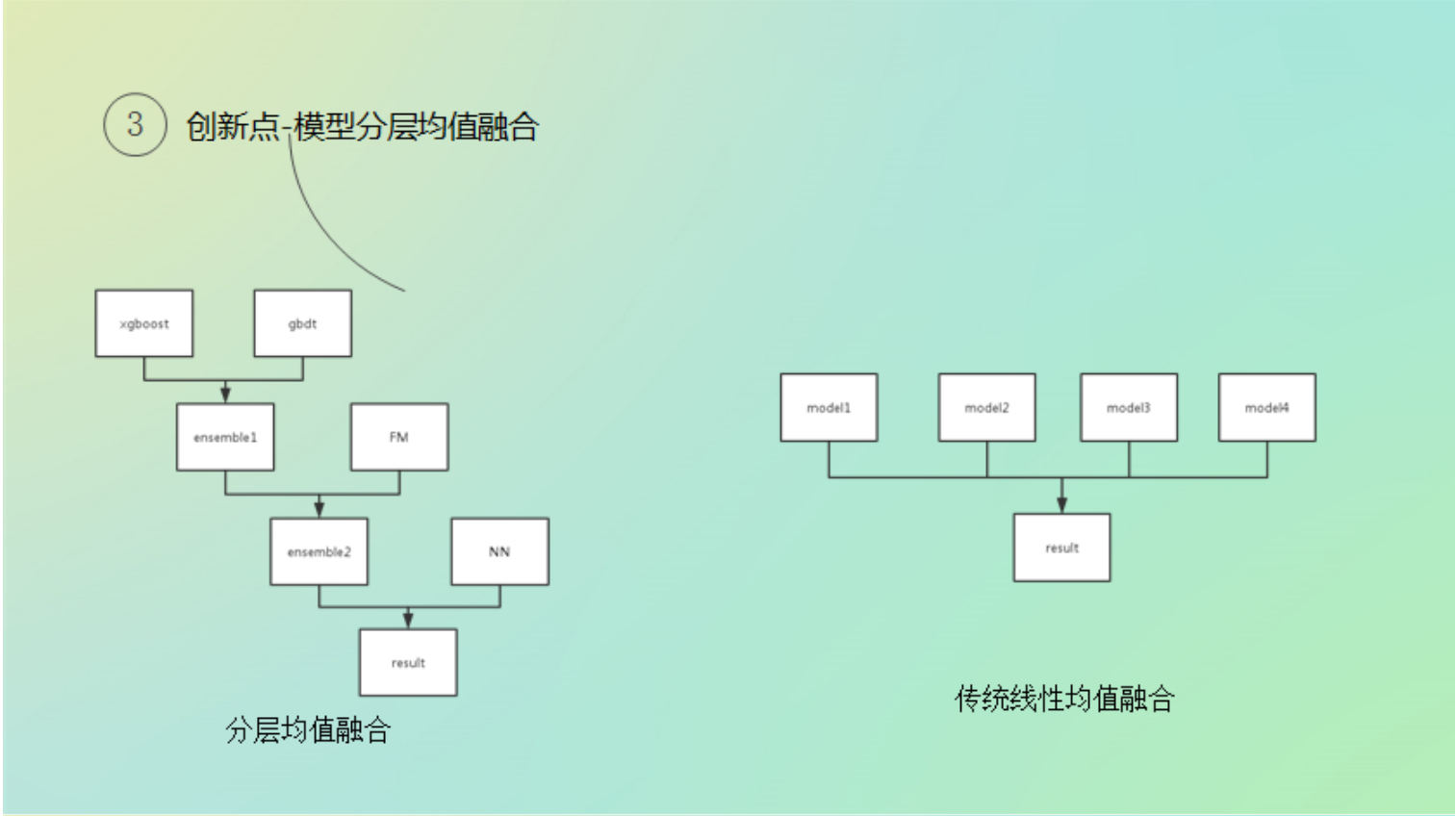
uid	rank

uid



3	1
4	3
5	2
6	5
7	10
8	4
10	8
11	6
13	7
15	9
16	11
17	14
18	12
19	13
21	15

1
2
9
12
14
20
25
28
30
33
35
37
43
44



0 回复

添加回复 **注:回复会奖励1点DC币，但被管理员删除回复，将扣除作者2DC币;可以使用@符号回复其他人**

回复

作者

(/user/5451) DataCastle运营 (/user/5451)

DC币: 428

无人回复话题

用户人品预测大赛--宝宝心里苦宝宝要说队--竞赛分享 (/topic/0f9a866d6d1e4f84bea2176db8237031.html)

用户人品预测大赛--wifi队--竞赛分享 (/topic/10078c08aecb44c18e1620686f0aa462.html)

用户人品预测大赛--火星人队--竞赛分享 (/topic/c5b1ce84f9ed42e7a933bbfcd2d6269a.html)

用户人品预测大赛--getmax队--竞赛分享 (/topic/cac927b5eff94193894f7dc588e1745a.html)

用户人品预测大赛--挖掘业务队--竞赛分享 (/topic/17416447cdab4bd5ad6a4bc00053f91e.html)

作者其他话题

用户人品预测大赛获奖团队分享 (/topic/58870500b2f84ddb9cbd4f6a45f180df.html)

用户人品预测大赛--宝宝心里苦宝宝要说队--竞赛分享 (/topic/0f9a866d6d1e4f84bea2176db8237031.html)

用户人品预测大赛--火星人队--竞赛分享 (/topic/c5b1ce84f9ed42e7a933bbfcd2d6269a.html)

用户人品预测大赛--getmax队--竞赛分享 (/topic/cac927b5eff94193894f7dc588e1745a.html)

用户人品预测大赛--挖掘业务队--竞赛分享 (/topic/17416447cdab4bd5ad6a4bc00053f91e.html)