第三届融360天机智能金融算法挑战赛

第二题 特征挖掘

SUPERVISEDLEARNING

目录

- 概述
- 建模思路
- 特征工程
- 训练和融合
- 参赛感想

概述

- 本题要求根据题目中提供的用户数据（包括关联关系、危险行为、标签类型、app情况，均已脱敏），通过数据挖掘技术，组合出有显著效果的特征，并利用这些特征构建模型预测用户的逾期情况。
- 评价指标为AUC，初赛成绩为0.7149，排名第一

特征挖掘

排名	预测评分	参赛队伍	所属单位	提交时间
1	0.7149	SupervisedLearning	仰望星空大学	2018-11-10
2	0.7087	zzzz	zz	2018-11-10
3	0.7081	bee	null	2018-11-10
4	0.7071	EST	P2C	2018-11-10
5	0.7065	ylx	互联网公司	2018-11-10

建模思路

- 用户本身
  - 关联关系
  - App
  - 标签类型
  - 危险行为
- 周围联系人
  - 一度
  - 二度

数据集	用户数量
Train	18959
Valid	4000
Test	6000
all	28959

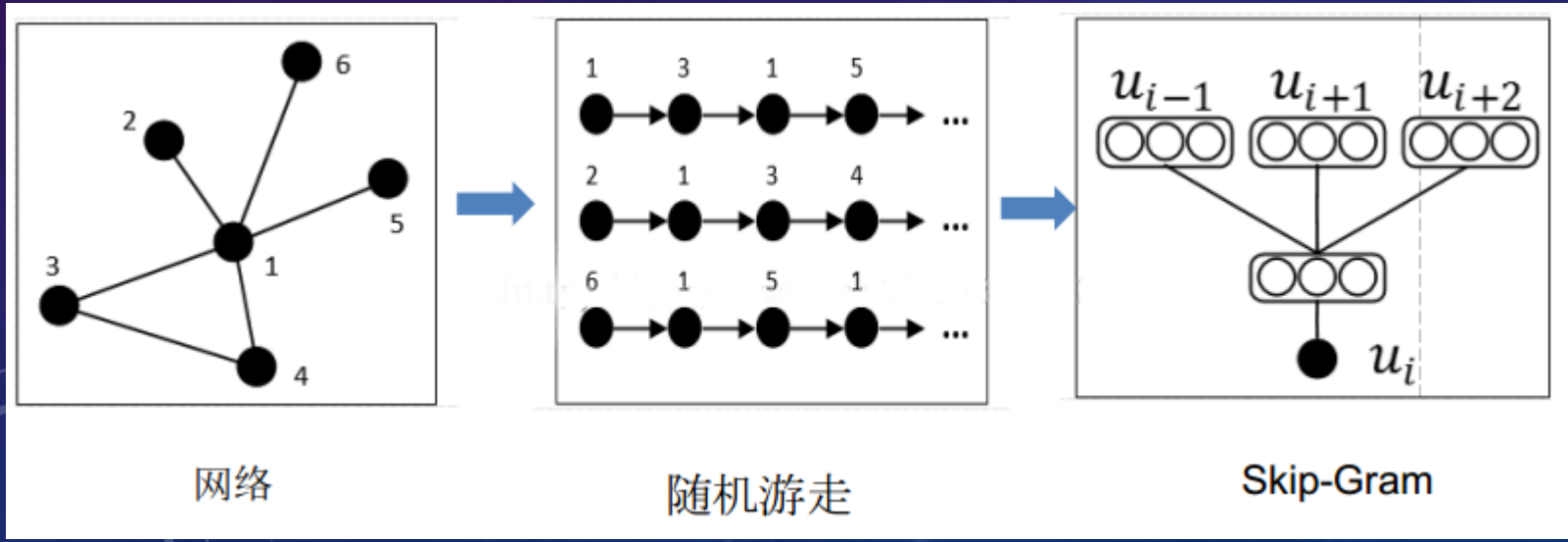
数据集	用户数量	Id ∩ all
APP	2759440	6966
关联关系	33728365	28442
标签类型	2266640	1544
危险行为	7437689	18735

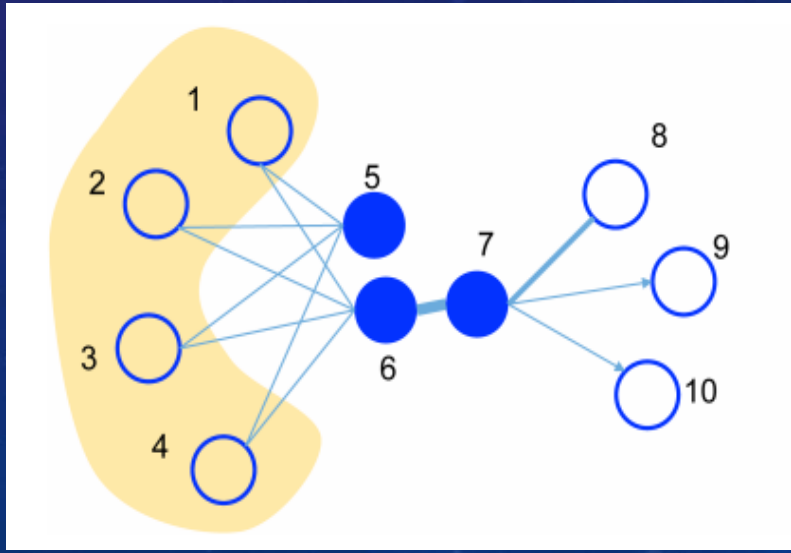
特征工程-关联关系

- 网络表示学习(network embedding)
  - 网络节点→低维向量
  - 随机游走 + word2vec
  - 一阶邻接性 + 二阶邻接性
  - Window = 5, dim=128, 192, 256

https://github.com/thunlp/OpenNE

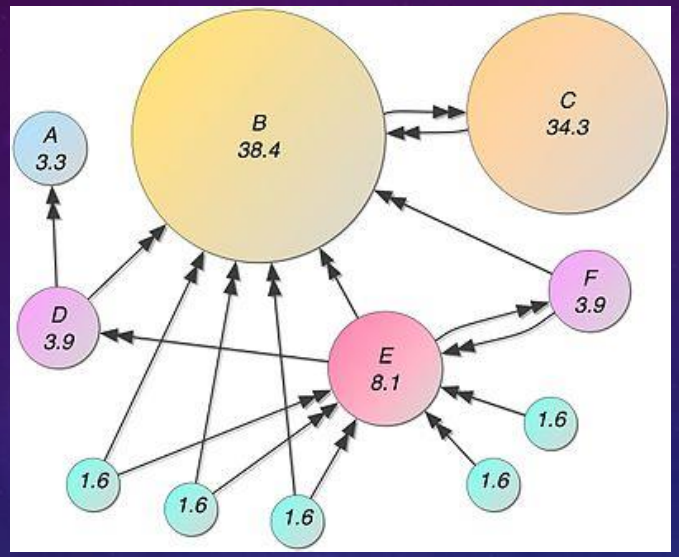
Algorithm	Time	Micro-F1	Macro-F1
DeepWalk	52s	0.669	0.560
LINE 2nd	70s	0.576	0.387
node2vec	32s	0.651	0.541
GraRep	19.6s	0.633	0.476

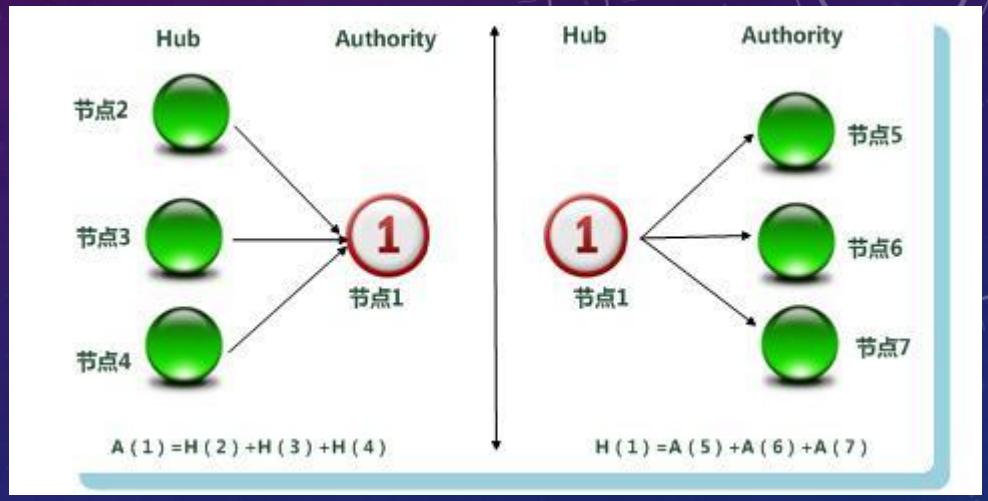




特征工程-关联关系

- 链接分析
  - PageRank
  - HITS
- 度特征
  - 出度、入度、度中心性
  - 以num、weight作为权重求和
  - “互粉关系”->“好友圈”的大小





特征工程-APP

- 计算用户安装APP的数量
- 统计train + valid + test中各个APP的用户数量
- 计算用户安装的APP中用户数的最大值、最小值、中位数、方差等统计量
- 保留用户量最多的前4000个APP，进行onehot编码，再进行降维
  - pca——使得降维后的分布最接近原始分布
  - lda——将每个APP看做单词，一个用户安装的APP构成一篇文档，得到安装APP的主题分布
  - nmf——利用非负矩阵分解，分解用户-APP矩阵，获得用户表示

特征工程-危险行为

- 计算危险行为的总数
- 计算各种危险行为占总数的比例

特征工程-标签类型

- 分别对一级类别和二级类别进行onehot
- 一级类别有24个，二级类别有44个
- 实际训练中发现只使用一级类别效果比较好

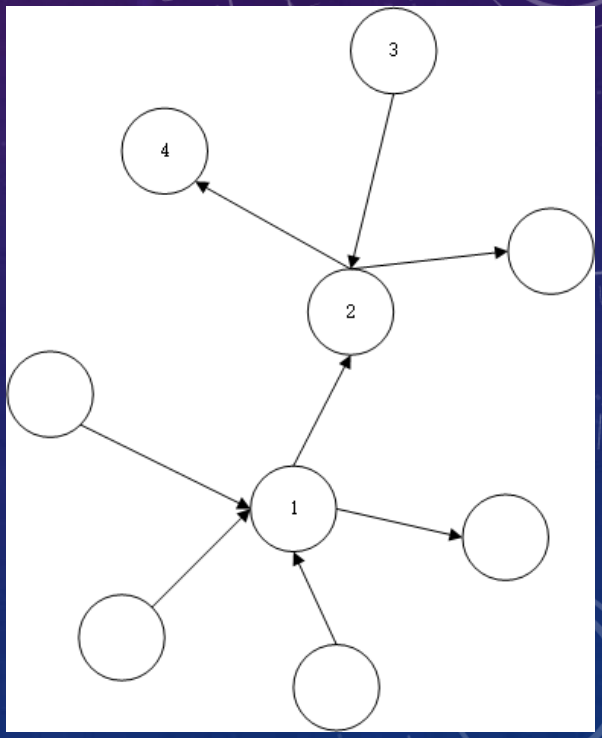
特征工程-一度联系人特征

- 2是1的一度联系人，3, 4是1的二度联系人
- 对于从节点i出发的边(i, j)，按照num加权，特征F 对应的一度联系人特征F'的计算公式如下所示：

$$F_i'\_num\_mean = \frac{\sum_{j \in edge\_from_i} num_{i,j} * F_j}{\sum_{j \in edge\_from_i} num_{i,j}}$$

$$F_i'\_num\_sum = \sum_{j \in edge\_from_i} num_{i,j} * F_j$$

- F ∈ 除embedding以外的自身特征



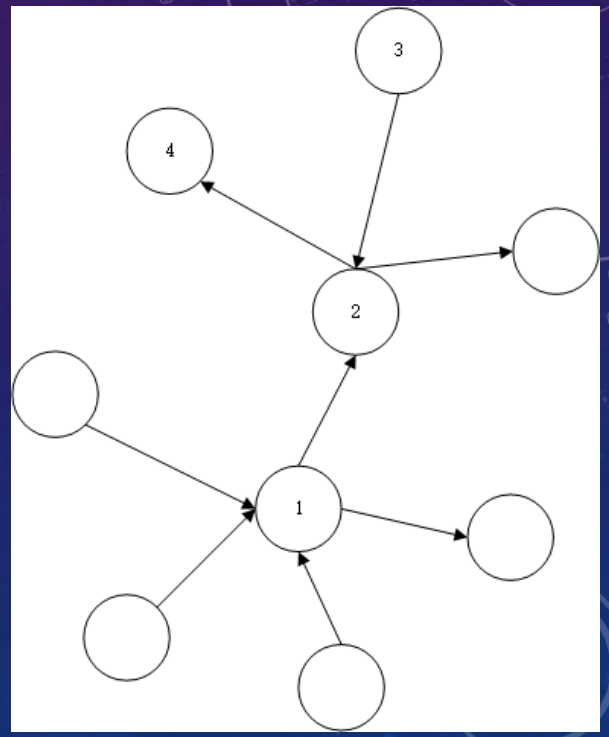
特征工程-二度联系人特征

- 对所有的一度联系人计算一度联系人特征F'
- 按同样的方法，计算二度联系人特征

$$F_i''\_num\_mean = \frac{\sum_{j \in edge\_from_i} num_{i,j} * F_j'}{\sum_{j \in edge\_from_i} num_{i,j}}$$

$$F_i''\_num\_sum = \sum_{j \in edge\_from_i} num_{i,j} * F_j'$$

- F　∈链接分析特征和度特征



特征工程-小结

- 用户自身特征
  - 网络表示学习
  - 链接分析
    - PageRank、HITS
  - 度特征
    - 度中心性、度统计量
- 周围联系人
  - 一度联系人
    - 除embedding之外的自身特征
  - 二度联系人
    - 链接分析、度特征

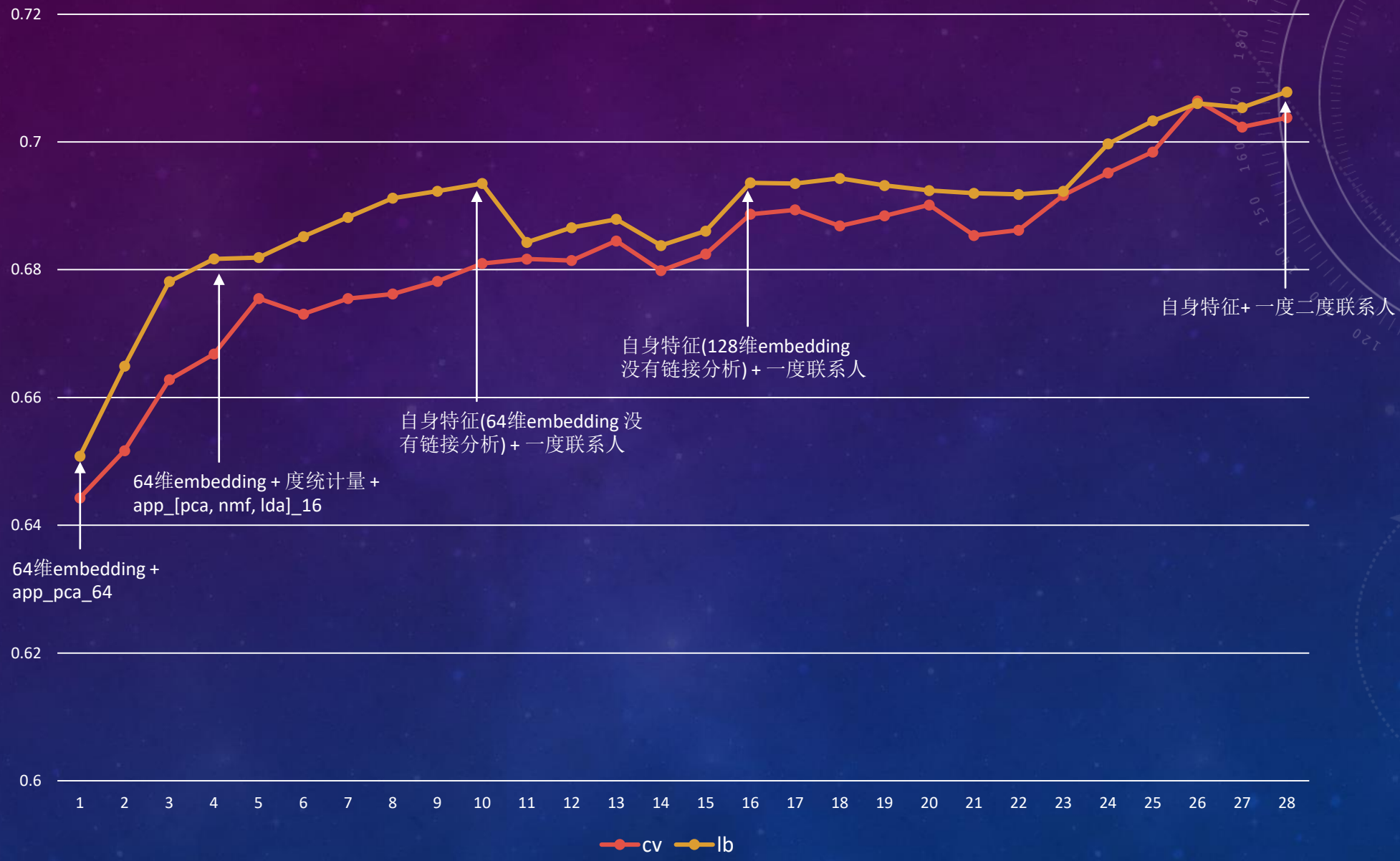
来源	记号	说明
关联关系-网络表示学习	dp_128	
	dp_192	
	dp_256	
关联关系-链接分析	pr	PageRank 值
	hits	HITS 算法计算的权威值和枢纽值
关联关系-度特征	dc	度中心性
	graph_info	出度入度的统计量
危险行为	risk	
标签类型	symbol	
app 安装情况	app_pca_16	
	app_lda_16	
	app_nmf_16	
	app_info	app 的统计信息
一度联系人	pr_1d	pagerank 值
	hits_1d	枢纽值和权威值
	dc_1d	度中心性
	graph_info_1d	
	app_graph_pca	
	app_graph_lda	
	app_graph_nmf	
	symbol_graph	
	risk_graph	
二度联系人	pr_2d	PageRank 值
	hits_2d	枢纽值和权威值
	dc_2d	度中心性
	graph_info_2d	出度入度的统计量

训练和融合

- 训练
  - 采用XgBoost和Lightgbm训练，5折交叉验证结果取平均
  - 使用hyperopt调整参数
  - 6个模型×5组参数
  - 最好单模型：　　CV: 0.7038　LB(valid): 0. 7078
- Stacking
  - 预测概率 、1/rank，共60维特征
  - 特征选择
    - 皮尔逊相关系数、卡方值、RF特征重要度、递归删除特征(REF)
    - 最好的9个特征
  - 二层分类器使用LR　　CV: 0.7157　LB(test): 0.7149

分类器	记号	特征维数
Lightgbm	lgb_572	572
	lgb_585	585
	lgb_628	628
	lgb_692	692
XGBoost	xgb_572	572
	xgb_652	652

lgb单模型上分曲线


0.72
0.7
0.68
0.66
0.64
0.62
0.6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
64维embedding + app_pca_64
64维embedding + 度统计量 + app_[pca, nmf, lda]_16
自身特征(64维embedding 没有链接分析) + 一度联系人
自身特征(128维embedding 没有链接分析) + 一度联系人
自身特征+ 一度二度联系人
cv
lb

参赛感想

- 相信交叉验证的结果，减小线上线下差距
- 被人赶超不要慌，坚持就是胜利
- 感谢融360组织这次比赛，我们学习到了很多知识和经验