Study REPORT 2020

学期汇报

陈嘉逸

2021-03-17

性格预测

- 基于学习行为预测学生的性格

基于会话的推荐

- 修改以往的两篇论文

01

02

CONTENTS

性格预测

/01

1.1 问题描述

研究问题

- 性格被证明在学习中起着重要的作用。如何通过学生的学习行为准确地预测学生的性格是值得关注的问题。

已有数据

- 2063名学生的学习行为记录
- 学生性格 (Five-Factor Model) by TIPI 问卷

目标

- 基于学生的学习行为，预测他在Five-Factor Model五个维度的分值/分类
  - 分值预测：直接预测某个维度的分值(1-7分)
  - 分类：根据一定的标准将1-7分化为多个类别
- 对比现有的预测模型，提出更好的模型，更准确地预测性格

Openness to experience

Conscientiousness

Extraversion

Agreeableness

Neuroticism

Low/Neutral/High

1.0/2.0/…/7.0

1.2 此前进展

早期工作

- 预处理数据
- 分类/回归 Baseline实现

提升Accuracy

- 基于行为序列+attention预测性格：关注某个具体的行为
- 例如“在某课程A主动录音X次”与Conscientiousness 有显著的关联

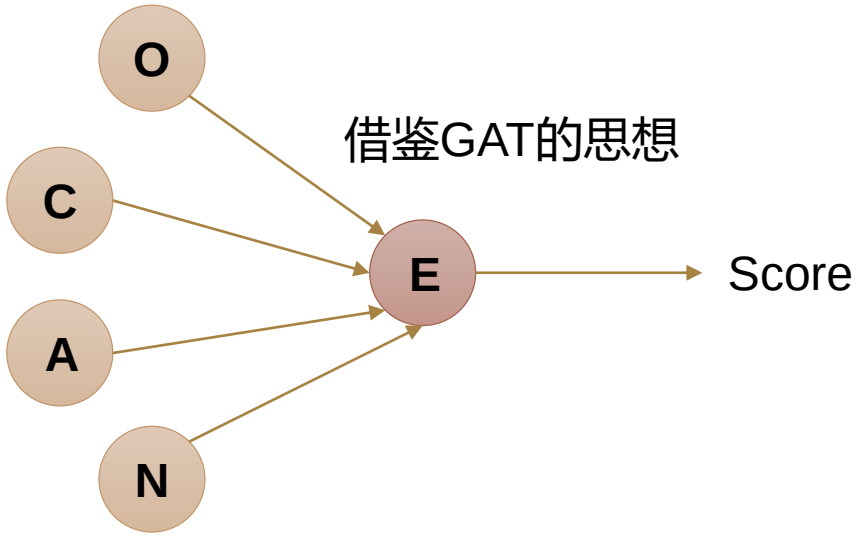
- 性格之间的联系
  - 预测Extraversion的时候考虑其他四个维度
  - 结果有些许提升

Imbalance问题

- 性格分布不平衡导致分类器都将性格预测为多数类——高Accuracy却没有意义
  - 现有的研究很少关注性格中不平衡分布的问题

- **从数据的角度**，尝试Oversampling 和Undersampling的手段
- -Oversampling：提出基于Neighbor的样本生成算法以及将GAN应用于样本生成
- -Undersampling: 尝试最新的模型
- **从模型的角度**：修改损失函数，将Focal Loss/ GHM Loss 用于模型

1.3 标签之间的联系



X	Y	R	P
O	C	0.3185	0
O	E	0.314	0
O	A	0.2627	0
O	N	-0.2579	0
C	E	0.1232	0
C	A	0.3473	0
C	N	-0.4661	0
E	A	0.0463	0.0354
E	N	-0.1974	0
A	N	-0.4406	0

F1

Model	O	C	E	A	N
RF	0.3408	0.3117	0.2743	0.3325	0.3098
SVM	0.3563	0.3461	0.3267	0.3956	0.3345
KNN	0.3485	0.3078	0.2971	0.3645	0.3131
DNN	0.3479	0.3479	0.3297	0.3911	0.3458
DNN-MultiLabel	0.3665	0.3922	0.3364	0.401	0.3592

1.4 Imbalance Data 实验结果

- 相对Baseline提升并不明显，并且不是在每个维度都管用
- 实验结果相对较低
  - G-Mean维持在0.35左右
  - 总体的F、Acc也在0.38左右，达不到较好的预测水平.
- 实验结果不稳定
  - 更换特征后，效果变化巨大

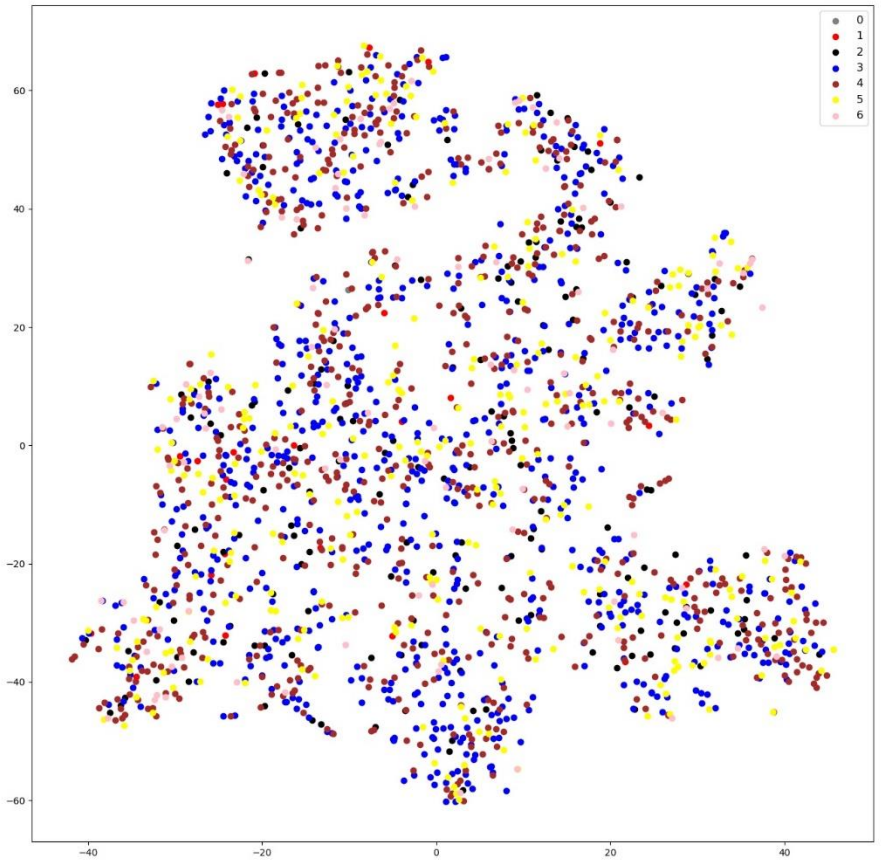
Metric: G-Mean					
Model	O	C	E	A	N
SMOTE	0.3417	0.3751	0.3087	0.3288	0.3439
BDSMOTE	0.3244	0.3705	0.2971	0.3378	0.3213
ADASYM	0.3487	0.3915	0.3111	0.3339	0.3206
SMOTETomek	**0.3586**	0.3821	**0.3285**	**0.3365**	0.3362
KMeansSMOTE	0.3267	0.3163	0.2909	0.312	0.2931
Ours	0.353	**0.407**	0.3282	0.319	**0.3651**

Metric: F					
Model	O	C	E	A	N
SMOTE	0.3624	0.387	0.3247	0.3379	0.3478
BDSMOTE	0.3428	0.3864	0.3339	0.3462	0.3348
ADASYM	0.3664	0.4028	0.3415	**0.3481**	0.3332
SMOTETomek	**0.3755**	0.3913	0.3407	0.3434	0.3423
KMeansSMOTE	0.3572	0.3578	0.3304	0.3315	0.3386
Ours	0.3679	**0.4243**	**0.3442**	0.344	**0.3789**

1.5 问题所在

样本重叠

- 通过T-SNE进行可视化，发现样本重叠严重，因此分类器难以进行分类
- 即使进行重采样，采样后的结果也很糟糕



分类器输出

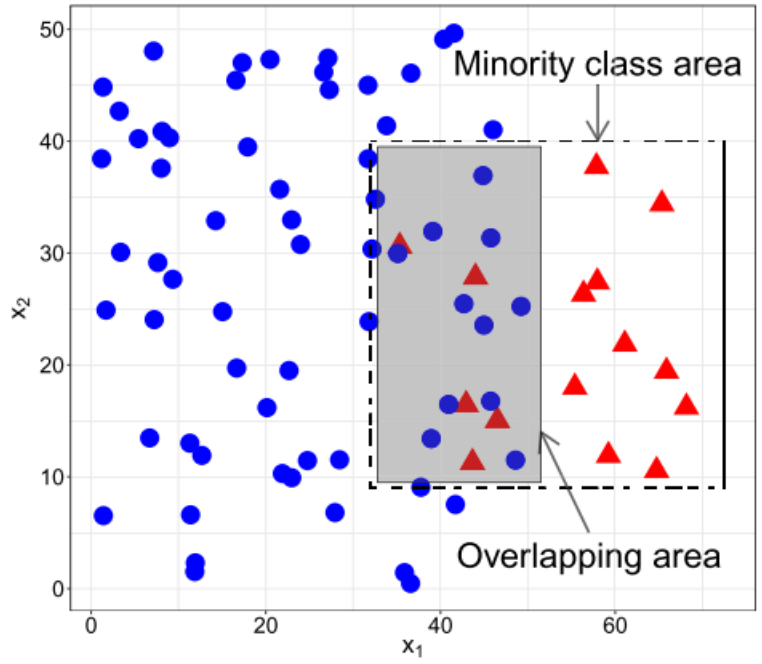
- 由于样本难以区分，分类器对3类输出的概率为0.3/0.4/0.3， 没有显著的区别。
- 因此类似Focal Loss 也无法很好地解决这样的问题

特征选择

- 尝试过多种特征抽取的方式，均无效

1.5 问题所在

理想的重采样



现实的重采样

Classifier	Resampler	O	C	E	A	N
DNN	RUS	0.3289	0.3618	0.34	0.3271	0.3371
DNN	NearMiss	0.3252	0.3244	0.3044	0.3138	0.3269
DNN	NB_Tomek	0.3058	0.2788	0.2277	0.1565	0.3256
DNN	NB_Common	0.2915	0.2612	0.1691	0.1694	0.3067
DNN	SMOTE	0.3204	0.3766	0.2829	0.3347	0.2955
DNN	BDSMOTE	0.3305	0.3834	0.3181	0.3258	0.3438
DNN	ADASYN	0.2916	0.3721	0.2874	0.3481	0.3345
DNN	SMOTETomek	0.3224	0.3643	0.2714	0.3419	0.3031
DNN	KMeansSMOTE	0.3111	0.3834	0.2717	0.3243	0.3237
DNN	AdaOBU	0	0.1284	0.2224	0.0752	0.1696
DNN	BoostOBU	0.3239	0.3563	0.3045	0.3289	0.3519

基于会话的推荐

/01

2.1 当前进展

修改以前的论文

01.推荐 + 可解释性

- 考虑Sequential Pattern、重复点击以及Item Similarity
- 已经完成修改并投出.

Why recommend this item

02.Accuracy + Long Tail

- Accuracy 和 Coverage的权衡
- 目前正在收集Baseline的实验结果
- 以及自己的模型修改.

Recommend long tail items

3.1未来工作

更好的序列建模

- Graph：充分利用全局图信息和单个序列的信息

更多的评价指标

- 针对Long tail的问题，将Imbalance Learning的思想应用到基于会话的推荐

基于会话的推荐

学习资源推荐

序列 & 推荐

个性化课程推荐

- 结合知识追踪、认知诊断，以及学生的画像，制定个性化的学习路径/推荐课程

Q&A

陈嘉逸

2021-03-16