基于学生行为数据的学习异常（异常表现）监测与预警

摘要（背景与简介）

一卡通提供了丰富的数据，刻画用户的行为特征

数据挖掘在教育领域的广泛应用：奖学金、贫困生预测、MOOC预测等

本文的目标在于根据一卡通系统记录的数据，刻画用户的多维行为，以实现：

* 学生挂科情况进行预测；
* 提前预测以预警并提供人工帮助；
* 对于不同维特征分别建模，分析有效的影响因子；
* 对于学期的不同阶段进行建模，分析影响时段。（\*）
* 以周为单位进行多个模型演化与结合（？是否可行呢）

1 相关工作

（成绩预测的相关工作）

* 学生按期毕业预测graduate
* 期末成绩预测final grade
* 在线课程弃课率预测dropout

（成绩预测利用的主要属性特征）

* CGPA
* 中期测评
* 人口统计学信息

2 先验知识

学习异常：挂科现象，影响毕业与工作

学生行为，可对正反例样本学生的刷卡记录在各类行为的分类做一个雷达分布图，进而分析

特征提取的合理性（非监督与监督学习结果是否一致）

行为与学习的关系

3 特征

|  |  |
| --- | --- |
| feature | detail |
| 基本生活行为 | 0平均吃早饭频率/周（6:00-10:00）+ breakfast  1平均食堂（在校）吃饭频率/周 + meal  2平均周末于食堂（在校）吃饭频率/周 + meal\_weekends  3平均运动频率/周（不完全统计）+ sports  4平均8:00AM前离开宿舍频率/周 （离开通过一天最早的室外刷卡记录进行判断？5:00-8:00）+ leave\_b\_8  5平均11:00PM后回宿舍频率/周 （晚上11点至第二天5点前）+ back\_a\_11  6平均洗澡频率/周（刷卡记录极少，不知道为何）+ shower（去掉）  7平均就医频率/周 + treatment |
| 消费行为 | 8平均消费金额/周 + consum  9平均充值金额/周 +recharge  10平均消费记录/周 + records  11平均食堂（吃饭）消费金额/周 +consum\_canteen  12平均食堂（吃饭）消费记录/周 + records\_canteen  13平均就医金额/周 + treat\_consum |
| 学习行为 | 14平均图书馆进出频率/周（刷卡记录极少）+ lib\_entrance（去掉）  平均图书馆借阅频率/周（记录太少）  15平均打印（购买）资料频率/周 + materials  16平均资料消费金额/周 + materials\_consum |
| 上网行为 | 平均网教时间/周（小时）（拿不到）  平均网教频率/周 |
| 增加行为特征 | 17日常购物频率/周 shopping  18 日常购物消费 shopping\_consum  19 交通频率 school\_bus  20 一卡通业务办理频率card\_operation |

暂时取18周的向量作为每一个频率特征向量，则每一个用户的特征为一个特征d\*18矩阵

对于多次刷卡行为如何定义为同一次吃饭行为（考虑地点类别、位置距离、刷卡时间距离）可以对每一个用户周消费分析得到一个密度图

如果不考虑周期性，那么消费特征还应加入充值持续时间等等。

增加规律性特征！！！

|  |  |  |
| --- | --- | --- |
| 位置类型p | 消费类型c | 行为类别b |
| 1 食堂 | 01 02 04 33 充值 | 1 食堂吃饭(p in {1,2},c=20) |
| 2 面包房 | 20 消费 | 2 食堂吃早饭(p in {1,2},c=20,t in {5:00-10:00}) |
| 3 浴室 | 其他 | 3 周末食堂吃饭(p in {1,2},c=20,t in {sat,sun}) |
| 4 宿舍 |  | 4 购物(p in {10},c=20) |
| 5 健身与活动 |  | 5 运动(p in {5},c=20) |
| 6 图书馆通道 |  | 6 洗澡(p in {3},c=20)去掉 |
| 7 图书馆借还处 |  | 7 就诊(p in {8},c=20) |
| 8 医务室 |  | 8 打印/购买资料(p in {9},c=20) |
| 9 书店、资料室 |  | 9 图书馆学习(p in {6,7}) |
| 10 超市等消费场所 |  | 10 充值(c in {01,02,04,33}) |
| 11 一卡通业务 |  | 11 一卡通业务(p in {11}) |
| 12校车 |  | 12 交通 (p in {12}) |
|  |  | 13 早8点之前离开宿舍(p not in {3,4},t in {05:00-08:00} ) 无法得到 |
|  |  | 14 晚11点之后回宿舍(p in {4},t in {23:00-05:00}) |

重新整理特征：

|  |  |
| --- | --- |
| 提取行为 | 相应特征 |
| 1.食堂吃饭 | 1 meal\_fre 01,0-2  2.meal\_amo 23,3-5 |
| 2.食堂吃早饭 | 3.breakfast\_fre 45,6-8  4.breakfast\_amo 67,9-11 |
| 3.周末于食堂吃饭 | 5.meal\_weekends\_fre 89  6.meal\_weekends\_amo 1011 |
| 4.购物 | 7.shopping\_fre 1213  8.shopping\_amo 1415 |
| 5.运动 | 9.sports\_fre 1617 |
| 6.就诊 | 10.treat\_fre 1819  11.treat\_amo 2021 |
| 7.打印/购买资料 | 12.materials\_fre 2223  13.materials\_amo 2425 |
| 8.图书馆学习 | 14.lib\_fre（已去掉） |
| 9.充值 | 15.recharge\_fre 2627  16.recharge\_amo 2829 |
| 10 一卡通业务 | 17.card\_service\_fre 3031 |
| 11 校车交通 | 18.bus\_fre 3233 |
| 12 早8点之前离开宿舍 | 19.leave\_b\_8 3435 |
| 13 晚11点之后回宿舍 | 20.back\_a\_11（已去掉） |
| 用户活跃度（行为序列长度） | 21Activeness 3637 ,54 |
| 多样性（行为序列中不同行为转移的次数） | 22Diversity 3839, 55 |

4解决方案

框架

模型与方法

评价指标

ROC

MCC

对于异常指数列表的top-k准确率

是否可以提出新的更准确的评价指标呢？对于二分类问题可能有点困难！

5 实验

数据描述

实验设置

实验结果与分析

6 结束语

7 未来工作

构建语义轨迹，并分析该特征对异常学习特征

观察人工干预机制对于学生异常学习状态的反馈

2017.3.1

更新

1. 行为集（将吃饭细粒度化，分为早中午，无法区分周末吃饭、离宿时间）

|  |  |  |  |
| --- | --- | --- | --- |
|  | 行为 | Entropy\_P | Entropy\_O |
| 1 | 早饭 | 55 | 67 |
| 2 | 中饭 | 56 | 68 |
| 3 | 晚饭 | 57 | 69 |
| 4 | 购物 | 58 | 70 |
| 5 | 运动 | 59 | 71 |
| 6 | 就诊 | 60 | 72 |
| 7 | 购买资料 | 61 | 73 |
| 8 | 图书馆 | 62 | 74 |
| 9 | 充值 | 63 | 75 |
| 10 | 一卡通业务 | 64 | 76 |
| 11 | 校车 | 65 | 77 |
| 12 | 宿舍行为 | 66 | 78 |
| 0 | 其他 |  |  |

Csv文件中的行为标签含义（已去掉图书馆学习与离宿）：

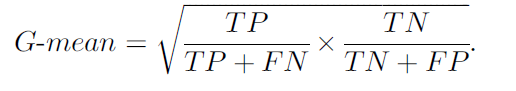
|  |  |  |
| --- | --- | --- |
|  | 属性 | 特征 |
| 1 meal\_fre  2.meal\_amo | 1  2 | 1,2,3  4,5,6 |
| 3.breakfast\_fre  4.breakfast\_amo | 3  4 | 7,8,9  10,11,12 |
| 5.meal\_weekends  6.meal\_weekends\_amo | 5  6 | 13,14,15  16,17,18 |
| 7.shopping\_fre  8.shopping\_amo | 7  8 | 19,20,21  22,23,24 |
| 9.sports\_fre | 9 | 25,26,27 |
| 10.treat\_fre  11.treat\_amo | 10  11 | 28,29,30  31,32,33 |
| 12.materials\_fre  13.materials\_amo | 12  13 | 34,35,36  37,38,39 |
| 15.recharge\_fre  16.recharge\_amo | 14  15 | 40,41,42  43,44,45 |
| 17.card\_service\_fre | 16 | 46,47,48 |
| 18.bus\_fre | 17 | 49,50,51 |
| 19.leave\_b\_8 | 18 | 52,53,54 |
| 21Activeness |  | 79 |
| 21diversity |  | 80 |

特征18\*3+12+12+2

1. 对一定时间段内、相同位置、相同消费类型的一组连续事件合并为活动，构造“活动序列”（O）注：待解决充值造成一次吃饭活动被分裂为两顿的情况；对时间与金额离散化，得到有限活动空间；t（1h为一个单元片）,p（有效地点11，去除浴室3）,c（15中刷卡类型）,金额三档, (24 \* 11\*15\*3)，以索引序列代替活动序列

金额离散化否？[0,10],[10,50],[50+]，活动为一组事件合并后，时间选择最初时间

1. 为每一个活动赋予行为标签并构造“行为标签序列”（I），忽略行为标签为0的行为与活动
2. 构建隐马模型并求参（以周为周期还是整个学期为周期）
3. 评价指标：



1. Synthetic Minority Oversampling Technique(SMOTE)