菜鸡组合队《大数据分析方法》课程创新性试验报告

学生成绩预测报告

队 名： 菜鸡组合

队 长： 汪东启

队员姓名： 滕云鹏

专 业： 计科一班

完成日期： 2018.06.20

**摘要**

通过本学期对大数据分析方法课程的学习，学习了很多数据处理及数据挖掘及分析方面的知识，预测问题是一类很常见的问题。本次项目选择的是一个对学生成绩进行预测的问题，本问题来源于DataCastle，通过此次项目实现了对学生成绩的预测。

关键词：数据分析 数据挖掘 分类预测

目录

[1.问题描述...............................................................................................................................1](#_Toc516740169)

[2.数据说明...............................................................................................................................1](#_Toc516740170)

[3.解决思路解决方案...............................................................................................................2](#_Toc516740171)

[3.1解决思路........................................................................................................................2](#_Toc516740172)

[3.2数据预处理.....................................................................................................................2](#_Toc516740173)

[3.3模型建立.........................................................................................................................2](#_Toc516740174)

3.4模型评估...............................................................................................................................3

[4.总结与展望...........................................................................................................................3](#_Toc516740175)

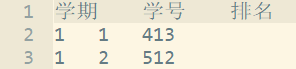
# 1.问题描述

我们希望通过借助大数据相关的挖掘技术和基础算法，从学生的校园行为数据中，根据学生出入图书馆的次数，以及借书和消费情况等，挖掘用户作息规律、兴趣爱好等，精准地预测学生之间的相对排名。通过对这些日常行为的建模来预测学生的学业成绩，发现学生成绩和日常校园行为之间的潜在关系，可以实现提前预警学生的异常情况，并进行适当的干预，因而对学生的培养、管理工作将会起到极其重要的作用。

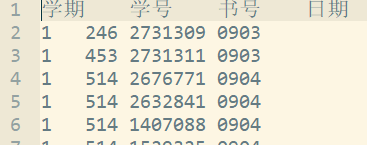
# 2.数据说明

本次竞赛中，我们将从某高校的某个学院随机抽取一定比例学生，提供这些学生在三个学期的图书馆进出记录、一卡通消费记录、图书馆借阅记录、以及综合成绩的相对排名。这一部分数据将作为训练数据。我们从另外的某学院随机抽取一定比例的学生，然后提供他们在三个学期的图书馆进出记录、一卡通消费记录、图书借阅记录、以及前两个学期的成绩排名。第三学期的成绩排名作为预测目标

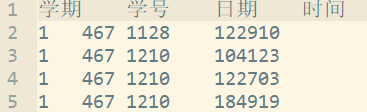
训练\成绩.txt。训练集的成绩文件，包含学期、学号、以及相对排名，共1614条数据



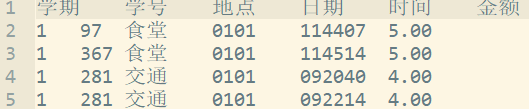
训练\借书.txt。训练集的图书借阅信息，包含学期、学号、书号、日期，共16297条数据



训练\图书门禁.txt。 训练集的图书门禁进入，包含学期、学号、日期、时间，61029条数据



训练\消费.txt。训练集的消费数据，包含学期、学号、地点、日期、时间、金额，共458091条数据



测试\成绩.txt。 测试集的成绩文件。字段同上

测试\借书.txt。 测试集的图书借阅信息。字段同上

测试\图书门禁.txt。 测试集的图书门禁进入。字段同上

测试\消费.txt。 测试集的消费数据。字段同上

# 3.解决思路解决方案

## 3.1解决思路

读取训练目录下的所有数据，统计每个学生的门禁、消费、借书等每个学期的总次数以及的相对离散程度，为了减少数据集的维度，用平均、最大、最小、方差四个维度来描述数据的离散程度，并把最后得到的训练集放在不同的训练模型中进行训练，找出最优结果的训练模型，并根据此结果对测试集进行与训练集相同的预处理放入该训练模型进行测试，得出最终预测出来的测试集的成绩。

## 3.2数据预处理

首先读取训练集中的成绩文件，把得到的学期、学号和排名信息存入数据集中，并对数据集中学号为主key、学期为次key进行排序，使得之后的数据集的更新的时候，可以通过(学号-1)\*3+(学期-1)为索引查找到该行。

之后读取借书文件。先根据图书类别文件，使用一个字典，储存每个图书id对应的图书类别号。使用字典统计所有种类图书每个学期的分别借书总数，作为该学期的46维数据，后来根据借书的时间，统计了借书总数的相对离散程度，也即分别统计了每个月、每天的借书次数，计算其最大、最小、方差、平均，以及统计了24小时借书的总次数，作为其离散特征，将得到的46维数据和离散特征全部存储到数据集中的对应行。

读取图书门禁文件，只统计了门禁次数的相对离散程度，同上，最后将离散特征存储到数据集中的对应行。

读取消费文件，将所有的消费信息统计到一个字典中，按字典编号查找所有需要的信息，统计每个学期在不同地点的消费总额以及5点到7点，12点到14点，晚8点到晚10点的消费次数及消费额度，每月的消费次数及消费额度。

最后将得到的数据集把每个学号的三个学期拉伸为一行，即将每个学生三个学期的所有数据视为一个整体，提出数据集中第三学期的成绩作为标签并在数据集中删除。

测试集同上得到。之后对测试集和训练集进行归一化，并把归一化后的训练集及其标签放入SVM、BP、随机森林、线性回归等模型中分别进行训练，使用评估函数对最后的模型预测结果与实际标签进行打分，选则评估值最高的模型作为最优模型，并把归一化后的测试集放入该模型进行测试，得到最后的预测结果。

### 3.3模型建立

导入sklearn包中的RandomForestRegression包，引入随机森林的相关函数及算法，创建一棵树，树的颗数及每棵树的最大属性数均可变化，通过在不同颗数及不同属性数下建立不同的随机森林，找到最优解决方案。

**3.4结果评估** 算法通过衡量预测排名和实际排名的Spearman相关性，为[0,1]之间的值，值越大，表示越相关，排名的预测就越准确。若要考虑n个学生的排名，学生i的预测排名为pi，而实际的排名为ri，di = pi - ri，那么



通过此函数将预测结果与实际结果作比较，结果范围为[负无穷,1]，通过计算得出当前最优方案下模型评分为0.8993，排名可以达到17名。

# 4.总结与展望

从训练数据及训练结果可以看出，学习成绩不但和个人本学期去图书馆门禁、消费、借书等的总次数有关，并且往往他的规律性越强，学习成绩越高。往往一个学生他的门禁时间在早上6点到7点，晚上10点左右，在早7点及中午12点等就餐高峰就餐时，他的成绩一般会越高。而学习成绩不理想的学生的作息习惯往往不太规律。作息的规律性一般会对成绩有较大的影响。

本次实验与这个项目中优秀的结果相比仍然有部分差距，可以通过多种不同算法结合来进一步提升精度，达到更优的效果。通过本次实验，掌握了很多数据分析及挖掘相关的知识，了解了相关的思想，在今后的学习中可以用这种方式解决一些实际问题，同时也会对今后的学习有很大的帮助。在如今的大数据时代，大数据的分析方法将会成为生活中不可或缺的一部分。我们要紧跟此次大数据的潮流，学习更多的知识和方法。

# 