落后生预警模型

摘要

越来越多的大学生因为在一学年内挂科太多或者没有获得足够的学分而留级甚至退学，在认识到这个严重的现状之后，我决定着手建立一个分析某位学生是否有留级或者退学风险的模型。这种预警模型出来的结果有两种，一种是有风险，另一种是没有风险。采用机器学习中的监督学习方式，由于预测结果只有两种所以是一个二分类问题，分类算法有很多，我们需要做的就是挨个尝试各种算法并最终找到相对比较准确的预测模型。该模型不仅可以针对在校学生的学业情况，还可以扩展到商业等其他领域，诸如可以预警公司员工是否有被解雇的危险等等。

关键词：落后生预警，机器学习，监督学习，数据挖掘

A Model for Prediction of At-risk Students by Using Educational Big Data

Abstract

More and more College students face the danger of failing to go up to the next grade or dropping out because they have failed too much subjects during a term or failed to get enough credits.Building a model which can predict whether a student will drop out the school has a

Keywords:Prediction Model,Machine Learning ,Supervised Learning,Data Mining

目录

1.Introduction

1. 简介

人工智能在当今迅猛的发展，各行各业都在迅速向人工智能靠拢，谷歌的‘AlphaGo‘，无人驾驶技术以及小米的智能家居等等，都向我们证明了未来是属于人工智能的。2018年的谷歌开发者大会也向人们展示了谷歌未来的人工智能研究方向。然而机器学习是目前最接近人工智能的系统，如果在没有机器学习的情况下创建人工智能，那么这个过程将会是复杂耗时的。所以机器学习是人工智能的基础。如今机器学习的应用更是家常便饭，例如垃圾邮件的分类，图像识别技术，某商品的销售趋势预测等等数不胜数。可以说机器学习无处不在，然而机器学习必须建立在对大量数据的分析之上，通过对已有数据的分析和处理加之学习之后就能对新的数据作出相对正确的预测。

如今在有了学生智能卡系统和教务处系统的条件下，大量的学生数据可以被收集起来，研究者们认为这些数据可以反映学生学习的努力程度（Campbell, DeBlois, & Oblinger, 2007）和学生的学习成绩（Macfadyen & Dawson, 2012; Yin et al.,2015）。

为了降低大学生的挂科率和留级率，许多学者利用学生在校表现等方面的数据来识别他们是否有学业预警的危险。他们认为可以成功预测学生在一门课程上的最终成绩（Lackey, Lackey, Grady, & Davis, 2003; Olani, 2009; Jin, Imbrie, Lin, & Chen, 2011）。预警模型可以提前提醒那些濒临挂科的学生来降低学校留级率。但是上述研究中存在普遍的问题就是预警模型只针对单独的一门课程而不是对所有的选修课程进行综合分析。另一个问题就是之前许多研究都是为网上课程设计的，现在缺乏对传统课程的合适研究模型.

所以在这篇论文中，预警模型是针对在校大学生的传统课程（非线上授课），找到学业落后的学生。使用中国海洋大学2013级本科生的数据，经处理后得到样本集，样本分为落后生和普通生两类，将其转变为二分类问题，（分类标准参见之后）对多种算法运用交叉检验，针对落后生寻找符合要求的模型，然后展示了实验结果，并进行分析，最后对模型进行一定的改进。

1. 实际问题分析

在实际的学习生活中，学生的学习成绩的优异与否往往和他的一些行为有关，好的学生偏向于上自习次数多一点，早起的次数多一点。学生的专业同样也对学业成绩有一定的影响，例如某些专业的课程难度偏大，或课程任务偏重等等都会导致学生无法在最后取得理想的成绩。学生的个人信息也可能是一个值得挖掘的地方。学生的户籍反映了他是来自哪里的学生。众所周知，来自江苏，山东，河北等省份的学生素来有着良好的学习习惯（可能和高中的教育方式有关），所以他们可能在步入大学殿堂之后学习可能要比其他地区的学生要轻松一点，相反来自那些偏远教育落后地区的学生可能就不占太大的优势，有可能成为他们学业落后的一项原因。综上所述，现在的任务就是从数据中分析出可能影响学生成绩的特征，然后再找出这些特征的关联程度，根据数值特征进行机器学习建立理想的模型从而达到预测落后生的效果。

1. 数据处理

3.1准备数据

从中国海洋大学数据中心取出数据，经分析2013级学生数据最为完整，于是以2013级学生为研究对象，表3-1介绍了提取的原始数据，包括其记录数和具体描述。2

表3-1：数据描述

|  |  |
| --- | --- |
| 表名 | 描述 |
| 基本信息表 | 基本信息表包含了入学年份、入学方式、专业类别、生源地、考试类别等信息。 |
| 成绩表 | 成绩表包含学年学期、年级、专业、课程名称、学分、成绩等信息。学年学期指的是学生在哪个学期选择了这门课；课程通过后，学生可得到相应的学分，学分可以衡量课程的重要程度；成绩字段指的是学生在课程中取得的最终成绩，包含平时课堂表现以及期末考试成绩，平时课堂表现占最终成绩的30%，期末考试占70%，成绩大多为百分制。 |
| 入馆记录 | 记录的是学生进入图书馆的时间，包括学号和时间两个字段，图书馆是本校学生学习的主要场所，进出图书馆的频率很大程度上表示学生在学业上的投入程度。 |
| 借书记录表 | 记录的是学生借阅和归还图书的时间，包括学号，时间和借阅或归还的动作。借阅图书的次数也能在一定程度上反映学生的学习热情 |
| 智能卡打卡记录 | 智能卡打卡记录包含学生智能卡的使用记录，比如在餐厅、超市的消费等记录，主要字段有时间、终端编号，终端编号代表其消费地点。 |

3.2 分析数据

3.2.1 学生成绩表分析

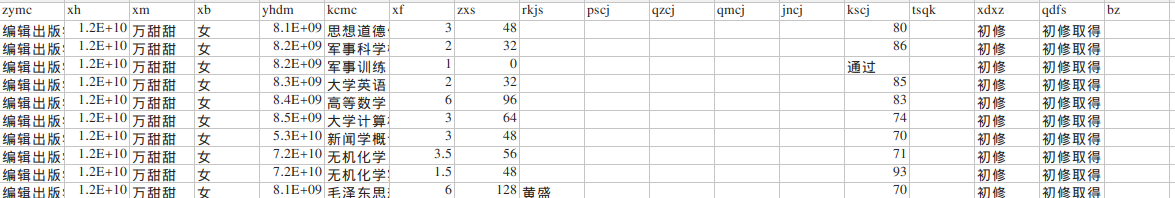


图3.1

在’‘grades.csv‘’文件中存储的是学生成绩信息，图3.1是表中的详细信息，注意到图中有很多信息，包括专业、学号、姓名、性别、选课号、课程名、课程学分、课程学时、课程分数等等。我们将用这张表来判断学生是否为落后生，我的判定依据就是依照中国海洋大学相关标准：在中国海洋大学，学生在学期末不足12学分会受到“学业警示”，在学年末没有满足学分和成绩要求则会受到留级处分。根据成绩表，提取有用信息，包括学生学号，选修课程学分，选修课程成绩得到一个精简的表之后再删除除2013-2014学年之外的学期数据行，注意到课程成绩列中包含非数值型的分数（如“通过”），根据实际情况一般这些课程并不能正确反映学生的学习状况，所以过滤掉这些成绩。之后对学号进行分类，计算每个学生每学期的平均分，2013-2014学年的加权平均分。最终得到的数据如图3.2所示。

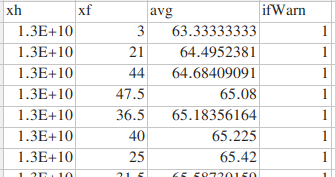


图3.2

加权平均分这个信息能够反映很多，如何评判一个学生的成绩好坏，一方面看他是否没有修够足够的学分，还有一方面是看他的加权平均分。年纪排名以及奖学金的评定都是以加权平均分为依据的。所以判定一个学生是否落后生可以参照他的加权平均分。

3.2.2入馆记录表和借书记录表

入馆记录表（如图3.3）借书记录（如图3.4）

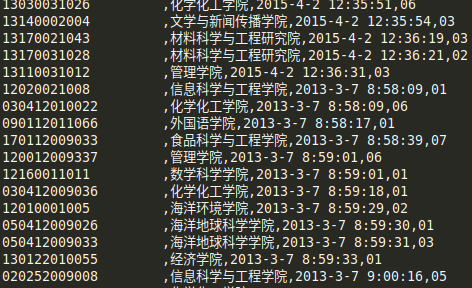


图3.3

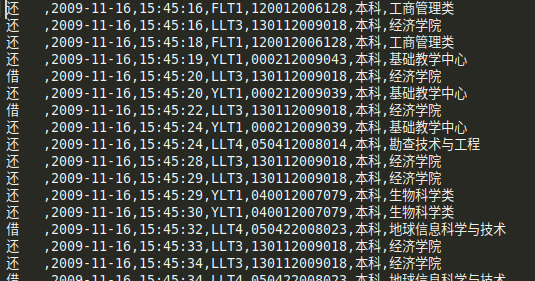


图3.4

对于入馆记录，我最终要得到的是每名学生在2013-2014学年内去图书馆的次数，所以仍然需要删除所属学院这列无用信息。再看时间数据，他是精确到时分秒，其实没有必要，于是除保留日期之外删除时间。最终在新表的基础上仍然根据学号分类得到学生的入馆次数。如图3.5.

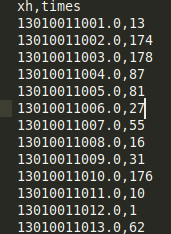
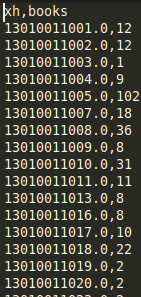
 

图3.5 图3.6

由于借书表与图书馆记录表差不多，进行相似操作得到最终数据表，图3.6

3.2.3学生信息表

这个表的信息量有点大，所以我选择不展示该表。对信息表中提取我所需要的数据，包括学生学号，专业，民族，生源地如图3.7

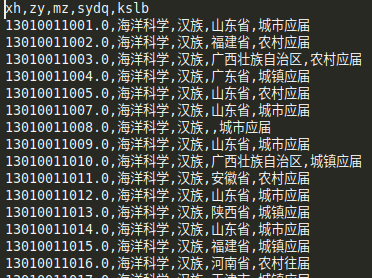
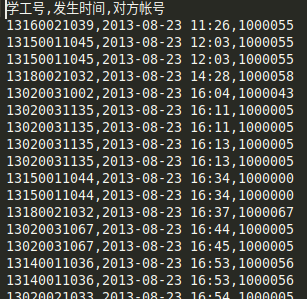
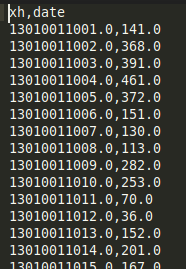


图3.7

3.2.4智能卡打卡记录（如图3.8）

是否每天坚持迟早饭是一个很好的习惯，身为一个大学生的我感同身受。那些很少吃早饭的学生一般偏向于睡觉到很晚因此错过了迟早饭的习惯，而那些坚持吃早饭的学生偏向于早起去上课或者自习。久而久之，行为体现在成绩上的差距就越来越明显。我把九点之前有过打卡记录的数据规定为吃早饭行为，统计每位学生在2013-2014学年之内的吃早饭总数如图3.9

图3.8 图3.9

3.2.5合并数据

经过上面的数据清洗和数据聚合操作，得到了5张新表，这五张表中描述了学生的特征如表3-2

|  |  |
| --- | --- |
| Features | Definitions |
| 专业 | 分为法学、工学、教育学、经济学、理学等 |
| 生源地 | 表示学生来自哪个省份 |
| cityOrTown | 学生是城镇户口还是农村户口 |
| totalCount | 图书馆入馆总次数，即学生在大一学年总的入馆次数 |
| totalBorrow | 借书总次数，即学生在大一学年总的借书次数 |
| totalBreakfast | 学生在一学年之内的吃早饭次数 |
| avg | 学生在一学年之内的加权平均分 |
| totalCredits | 学生在一学年之内所修的总学分 |
| 上学期是否受到学业警示 | 在中国海洋大学，学校会在每学期末予以学业警示，根据上学期的成绩和学分计算每个学生的学业警示情况。 |
| 上学期期末成绩 | 在成绩记录中有少数课程为等级制，这些课程大多无法体现学生学业表现，所以删除这些课程，保留百分制的课程，计算每个学生的期末成绩。在求得成绩后，考虑不同专业的平均成绩不同，所以需要对成绩进行标准化，将所有学生的成绩划归到同一个分布。 |

由于这些数据存在与不同的表格之中，所以我们对各个表进行合并操作，合并依据是相同的学号。

3.3数据预处理

机器学习里有一句名言：数据和特征决定了机器学习的上限，而模型和算法的应用只是让我们逼近这个上限。这个说法形象且深刻的提出前期数据处理和特征分析的重要性。这一点从我们往往用整个数据挖掘全流程60%以上的时间和精力去做建模前期的数据处理和特征分析也能看出。

3.3.1数据归一化

经过数据分析之后，得到许多特征，由于他们的单位以及数值范围的不一致，如果不提前进行数据预处理可能会破坏预测模型的准确性。所以对原始数据进行归一化（又称规范化）操作。归一化的任务主要有两个：

1 把数变为（0，1）之间的小数  
        主要是为了数据处理方便提出来的，把数据映射到0～1范围之内处理，更加便捷快速，应该归到数字信号处理范畴之内。  
2 把有量纲表达式变为无量纲表达式  
        归一化是一种简化计算的方式，即将有量纲的表达式，经过变换，化为无量纲的表达式，成为纯量。

经过归一化之后，有两个好处：

1. 提升模型的收敛速度

2. 提升模型的精度

常用的归一化方法有min-max标准化 和 z-score 标准化，这里选取第一种方式。min-max标准化（也称为离差标准化，是对原始数据的线性变换，使结果值映射到[0 - 1]之间）的公式：newvalue=(oldvalue-min)/(max-min)

3.3.2字符特征编码

在前面的章节里面，对连续型的数值特征进行了预处理，但是数据集中还存在其他的一些特征，例如专业类别，生源地，民族，农村还是城市，这些都是字符串型的数据，不能被我们机器所学习，所以想要利用这些数据，首先就必须给这些数据进行编码。

scikit-learn库里面提供了一种比较好的编码方式”one-hot”编码。首先定义一下独热编码：独热编码即 One-Hot 编码，又称一位有效编码，其方法是使用N位状态寄存器来对N个状态进行编码，每个状态都有它独立的寄存器位，并且在任意时候，其中只有一位有效。

3.3.3划分测试级和训练集

由于手上处理好的数据是2013级学生的，这是全部数据，我们在根据这些数据建立模型之后要对模型进行正确率的分析，所以必须将全部数据划分为训练集（training set）和测试集（test set）。为避免数据划分带来的不平衡的影响，这里采用交叉验证法。交叉验证法的定义：”交叉验证法”先将数据集D划分为k个大小相似的互斥子集每个子集都尽可能保持数据分布的一致性，即从D中通过分层采样得到。然后，每次用k-1个子集的并集作为训练集，余下的子集作为测试集；这样就可以获得k组训练/测试集，从而可以进行k次训练和测试，最终返回的是k个测试结果的均值。这种划分避免了划分不当带来的偶然性，使算法生成的模型更加的一般化。在这次落后生的模型当中，为了确定k的最佳值，依次枚举k从5到10带入交叉算法中验证，看效果最好的是哪一个。最终决定k选取5.

3.4 特征选择

前面的章节我们已经将所有的特征确定下来，接下来要做的就是如何利用这些特征进行学习。然而在学习之前，我们还有一项重要的事没有做，那就是特征选择，又称特征工程。

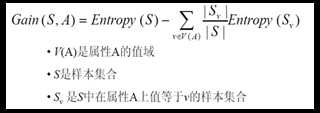
3.4.1为什么要特征选择

我们能用很多属性描述一名学生，例如加权平均分，借书次数，吃早饭次数等等，但有经验的人往往只需要看看他的加权平均分，和他一学期所修的总学分就可以大致知道他是否是一名学习优异的学生。换言之，对一个学习任务来说，给定属性集，其中有些属性很关键，另一些则可能没什么用，相关性弱的特征会给数据集带来冗余信息甚至噪声，干扰模型训练，对最终结果产生不利影响，同时过多的属性会造成工作冗余和资源浪费。所以我们必须进行特征选择，将高维空间的样本通过映射或者是变换的方式转换到低维空间，达到降维的目的，然后通过特征选取删选掉冗余和不相关的特征来进一步降维。

3.4.2如何进行特征选择

特征选择算法有好多，例如过滤式选择和包裹式选择。我采用了第一种方式。过滤式的主要思想是：对每一维的特征“打分”，即给每一维的特征赋予权重，这样的权重就代表着该维特征的重要性，然后依据权重排序。这里的权重称为信息增益，信息增益可以表示各个特征对分类任务的作用大小。接下来的任务就是计算每个特征的信息增益（见表3-3），增益的值越高，说明它和分类结果关系越大。

信息增益的数学公式：



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Feature | InfoGain | Feature | InfoGain | Feature | InfoGain |
| 上学期期末成绩 | 0.201 | meanM | 0.017 | totalCount | 0.003 |
| 专业排名 | 0.082 |  | 0.016 |  | 0.002 |
| 学业警示 | 0.034 |  | 0.013 |  | 0 |
| 专业类别 | 0.029 |  | 0.011 |  | 0 |
| 学分 | 0.026 |  | 0.006 |  | 0 |
| 生源地 | 0.021 |  | 0.003 |  | 0 |

表3-3

1. 模型建立
2. 总结
3. 参考文献