基于校园大数据的学生多维画像分析

李特特　厦门大学软件学院软件工程系大三学生　福建省 厦门市　361005
郑龙天　厦门大学软件学院软件工程系大四学生　福建省 厦门市　361005
游侯杰　厦门大学软件学院软件工程系大四学生　福建省 厦门市　361005
李佳声　厦门大学软件学院软件工程系大四学生　福建省 厦门市　361005
杨智涵　厦门大学软件学院软件工程系大四学生　福建省 厦门市　361005

本文系项目“基于大数据的学生多维度画像系统”研究成果之一。

[摘要]本论文研究的主要目标是通过深度学习和数据挖掘技术，利用学生信息对学生个体、群体进行多维画像分析，本系统目前阶段完成了奖学金预测、助学金预测，失联预警、成绩预测等模块，为解决目前数据量有限的问题，我们利用 GAN 网络根据已有数据生成新数据，实现数据增强，本系统主要是使用 Python 语言，利用机器学习和深度学习相关算法进行数据分析，基于 DJango 开发网站，为了更加直观展示，使用 ECharts 和词云技术将数据以词云、图谱等方式进行可视化展示。本系统对于已实现的预测模块都生成了高精度模型融合器，利用学生真实数据使得结果更加有说服力，如我们可以根据往届学生的数据放入模型融合器进行分析，可以快速得到验证，基于 GAN 网络模型生成的数据进行有效性验证，我们的系统未来能用到各个高校，方便学校对学生的管理。

关键词：数据挖掘 学生画像 模型融合 机器学习

分类号：TP319

本文系项目“基于大数据的学生多维度画像系统”研究成果之一。

Multidimensional Picture Analysis of Students Based on Big Data of Campus

Li Tete ,Software School of Xiamen University,Xiamen,361005
Zheng Longtian, Software School of Xiamen University,Xiamen,361005
You Houjie, Software School of Xiamen University,Xiamen,361005
Li Jiasheng, Software School of Xiamen University,Xiamen,361005
Yang Zhihan, Software School of Xiamen University,Xiamen,361005

[Abstract] The main goal of this thesis is to use multi-dimensional portrait analysis of student individuals and groups through student information through deep learning and data mining. The current stage of the system has completed the scholarship prediction, financial aid prediction, lost association warning, performance prediction module. In order to solve the problem of limited amount of data, we use GAN network to generate new data based on the existing data to enhance the data. The system is mainly using Python, machine learning and depth learning, and based on DJango development site. In order to be more intuitive, we use ECharts and word cloud technology to visualize the data in terms of word cloud and atlas. The system has generated a high-precision model fusion for the realized prediction module. For example, we can put the data from the previous students into the model coffer for analysis, which can be quickly verified and validated based on the data generated by the GAN network model. Our system can be used in various colleges and universities in the future to facilitate the management of students in schools.

Keywords: Data Mining; Student Portrait； Model Fusion; Machine Learning

本文系项目“基于大数据的学生多维度画像系统”研究成果之一。

目录

第1章　绪论

随着信息技术的发展，高校以数字化信息和网络为基础，建立集教学、科研、管理、技术服务、生活服务等应用为一体的教育环境。学校的数据库中，保存有学生静态和动态信息。静态信息是学生的相对稳定的信息，包括学生户籍、学号、所属学院等信息；动态信息是随着用户行为不断变化的信息，包括校园卡消费信息、门禁信息、图书馆出入信息等。通过这些信息，可以为学校师生提供更好的服务。

本系统主要目标是将繁杂的学生信息，通过数据挖掘和机器学习技术，从多个维度刻画学生和群体信息，从多角度精确了解学生的相关行为，将学生和群体最大化的“透明化”。然后以此来进行相关预测、预警和提醒，如：“挂科预警”、“失联预警”、“失业预警”、“助学金分析”、“奖学金分析”，满足学校职能部门的管理需求，提高校园的信息化管理水平，促进教育变革与发展。

本文系项目“基于大数据的学生多维度画像系统”研究成果之一。

第2章　相关技术介绍

本系统采用 Django 建站, 使用 Echarts 进行数据可视化，用 sklearn 做预测算法。

2.1 Django

Django 是基于 MVC 思想的 web 开源框架，我们使用该框架主要是为了开发网站和成果的可视化展示的需求。但是 Django 框架不同与 MVC 的是采用了 MTV 模式，与其他框架相比，Django 能够简便、快速的开发数据库驱动的网站，因为他有许多的第三方插件的支持和丰富的 API 供用户使用，甚至还可以添加自身的工具包，使得 Django 有很强的扩展性。

2.2 Echarts

Echarts 是一款便捷高效的可视化工具，对于我们网站而言，最终的结果是分析数据的价值，当然离不开数据的可视化问题，为了让用户对我们的结果看得更加直观，我们选择了 Echarts 可视化，Echarts 实现了饼状图，热力图，云图等，让用户能够多角度看到结果。

2.3 Sklearn

sklearn 是基于 numpy 和 scipy 的一个机器学习算法库，设计的非常优雅，它让我们能够使用同样的接口来实现所有不同的算法调用，借助 SKlearn 库，我们可以简单的实现复杂的机器学习算法的调用、比较，sklearn 封装了特征选择，模型筛选，模型融合等算法的实现，sklearn 还为我们提供了丰富的 API 文档供我们使用，是一款非常适合初学者的 python 学习库。

第3章　系统设计思路

本系统设计思路如图：

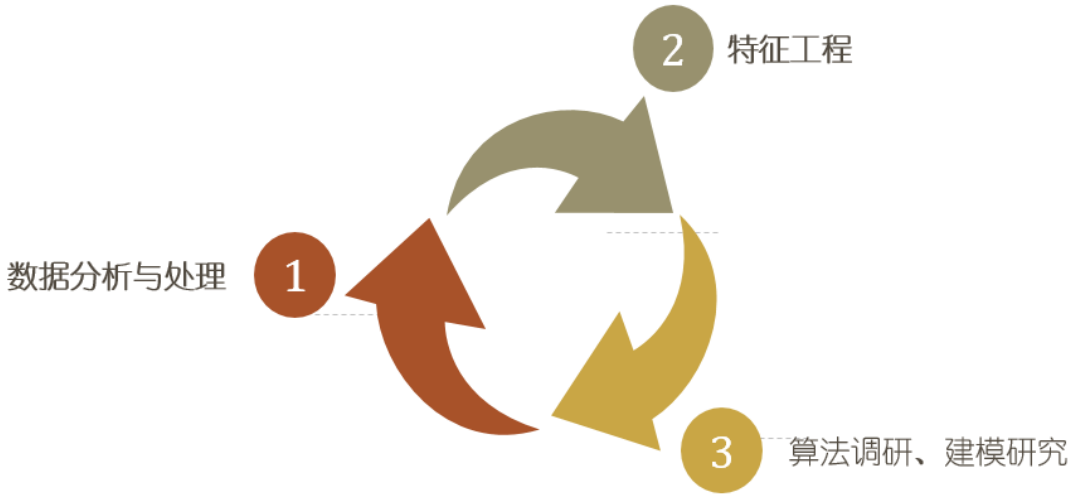

图 1 系统设计思路

3.1 数据分析和处理

本系统主要收集了学生校园卡消费记录、宿舍门禁记录、图书馆门禁记录、图书馆借阅记录、成绩信息，参加科创比赛以及获奖情况等方面的数据，对学生进行多维画像分析，刻画学生的个体画像。

目前由于校方的支持力度有限，所获取的学生真实数据有限，存在数据不平衡、数据缺失等问题，对于数据挖掘技术而言，要让系统的可靠性，真实性增强，必须要有一定的数据量，所以我们采用了分层抽样，过采样分析数据比例以及分布状况，利用 GAN 网络生成可靠真实的数据加入模型的训练，旨在让生成的模型融合器有更强的分析能力和适应性。

数据不平衡：各类别人数比例不平衡，例如：得助学金人数远少于没获奖人数，属于在大量非目标数据中找少量目标数据的问题。通过分层抽样，过采样处理以及 GAN 网络模型处理各个得奖学金类别学生数，使各个类别的学生比例小于 4:1，达到数据平衡，否则会造成最后分析的目标偏向大类，导致模型训练失败。

数据缺失：各个表中都缺失一定比例学生数据，如训练集中 Student_id 有 10783 条数据而 Score 中仅有 9000 条数据。在同一个表中，也存在某些字段缺失。我们对数据进行了清洗，去除表中重复条目，并利用 GAN 网络填充了缺失数据。这样可以使我们真实数据保持有效性。

3.2 特征工程

主要是将数据进行离散化。

离散化的原因其一是众多数据挖掘算法（如决策树、NativeBayes 等）只能处理离散化的数据；其二是使模型结果更稳定：把数据分级，而不是原数据可以破除极端值和异常值影响。

随着学生信息量的增加，可获取的特征也在成倍的递增，为了让最后的模型可靠性以及实用性增强，一方面适当的特征工程可以使得我们筛选出我们需要的特征，另一方面特征数量的减少也更加有利于我们模型训练的速度。

PCA 白化：（特征预处理）

白化的目的是降低输入的冗余性，使输入的特征之间的相关性较低，并且所有的特征具有相同的方差

特征选择方法：

filter method： 通过统计学的方法对每个 feature 给出一个 score，通过 score 对特征进行排序，然后从中选取 score 最高的子集. 这种方法仅仅是对每个 feature 进行独立考虑，没有考虑到 feature 之间的依赖性或相关性. 常用的方法有：卡方检验，信息增益等.

图 2 filter method

wrapper method： 和 filter method 相比，wrapper method 考虑到了 feature 之间的相关性，通过考虑 feature 的组合对于 model 性能的影响. 比较不同组合之间的差异，选取性能最好的组合. 比如 recursive feature selection

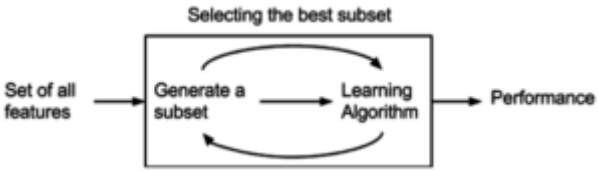

图 3 wrapper method

embedded method： 结合前面二者的优点，在模型建立的时候，同时计算模型的准确率. 最常见的 embedded method 是 regularization methods（简单来说就是通过增加 penalization coefficients 来约束模型的复杂度）.

3.3 算法调研、建模研究

在具体研究过程中，针对不同的学生行为以及种类，我们可以通过构建模型来分析。可以从如下几方面进行考虑：

1）选择适当的模型： 在得到好的属性之后，针对数据的特征，如离散值众多

等特点，选用适当的模型进行预测。在比较各个模型的优劣之后，选择较为合适的模型进行细致的调参。这是有效辨别敏感用户的另一种方式。

2) 多个维度分析建模： 可以从多个角度入手来分析建立模型，例如抽取学生日消费额进行研究，构建月消费分析模型，从中分析学生家庭情况是否存在异常行为，尤其是贫困生，可以针对月消费额变化情况提供是否接收为贫困援助对象等。

3) 不同模型的集成：不同模型具有不同的偏好，当我们从多个角度用不同的模型对问题进行了预测时就可以得到多个具有不同偏好、对结果具有不同偏差的模型。有效的结合这些模型，让他们互相纠正。这不失为提高模型精度的一种好方式。

本文系项目“基于大数据的学生多维度画像系统”研究成果之一。

第4章　系统实现

4.1 数据分析

4.1.1 数据表分析

4.1.1.1 图书借阅数据 borrow

1）内容构成：

字段含义：学生 id，借阅日期，图书名称，图书编号

9708,2014/2/25,″我的英语日记/ (韩)南银英著 (韩)卢炫廷插图″,″H315 502″

6956,2013/10/27,″解读联想思维： 联想教父柳传志″,″K825.38=76 547″

9076,2014/3/28,″公司法 gong si fa = = Corporation law / 范健, 王建文著 eng″

有些图书的编号缺失。字段描述和示例如下（第三条记录缺失图书编号）

对缺失值的处理将在第 4 章描述

2）数据表分析：

”library_borrow”：

该表的一条记录代表了一位学生的借书记录，我们通过主观分析，认为一个学生是否能够获得助学金，以及获得助学金的额度与该学生是否借阅某一些特定的书籍这种具体的数据并没有关联。反之，可能与该学生的总借阅量存在关系。于是，我们通过对该表进行统计分析，生成了一个特征值”library_borrow”。该特征值代表每位学生的图书借阅总量，将作为训练集的特征之一。

4.1.1.2 一卡通数据 card

1）内容构成：

学生 id，消费类别，消费地点，消费方式，消费时间，消费金额，剩余金额

1006,"POS 消费","地点 551","淋浴","2013/09/01 00:00:32","0.5","124.9"

1406,"POS 消费","地点 78","其他","2013/09/01 00:00:40","0.6","373.82"

13554,"POS 消费","地点 6","淋浴","2013/09/01 00:00:57","0.5","522.37"

2）数据表分析：

该表的一条记录代表了一位学生的一次消费记录，通过分析，我们发现消费类别一项的类型包括“charge”,”POS 消费”，“交易冲正”，“卡冻结“等 19 项，我们分析认为对于每一位学生来说，学生卡充值平均金额，即该学生的消费类型为“charge”的消费记录的消费金额之和除以该学生的消费类型为“charge”的消费记录的条数与该学生是否为贫困生有密切关系。于是，我们通过对该表进行统计分析，生成了一个特征值”avg_charge”。该特征值代表每位学生的每次充钱金额的平均值，将作为训练集的特征之一。此外，分析得出“cost_amount”（学生历史总消费金额）、“cost_variance”（学生消费方差）、“cost_avg_day_superMarket”（学生每天超市消费平均值）、“cost_avg_day_laundryroom”（学生每天洗衣房消费平均值）、“cost_avg_day_dinnerHall”（学生每天食堂消费平均值）、“cost_rate_supermarket”（学生花费在超市的消费总额占总消费的比例）、“cost_rate_laundryroom”（学生花费在洗衣房的消费总额占总消费的比例）、“cost_rate_dinnerhall”（学生花费在食堂的消费总额占总消费的比例）、“cost_times_day_supermarket”（学生每天超市消费平均值）、“cost_times_day_dinnerhall”（学生每天食堂消费平均值）、“cost_times_day_laundryroom”（学生每天洗衣房消费平均值）、“cost_times”（学生消费总次数）、“balance_rank”（学生卡内余额值在全体学生中的排名）、“card_days”（学生 card 活跃天数）、“time6_7costs”（学生每日 6 点-7 点的消费总额）、“time7_8costs”（学生每日 7 点-8 点的消费总额）、“totaldinnercosts”（学生该学生日饭堂消费的总额）、“avgdayscosts”（学生的日平均消费）、“consumetimes11_12”（学生每天 11 点 -12 点消费的总次数）、“consumetimes0_25”（学生单次消费金额在 0-2.5 元之间的次数）、“countcost0_10”（学生当日总消费在 0-10 元范围内的总天数）、“cardrecharge”（学生卡充值总额）、“maxcost7_8”（学生 7 点 -8 点间的最大单笔消费的消费金额）、“below10_rank”（学生日消费金额小于 10 天数占其 card 活跃天数的比例）、“below2_5_rank”（学生单次消费金额小于 2.5 次数占其总消费次数的比例）、“consume_rank”（学生的消费排名）、“time7_8consume_avg”（学生 7 点 -8 点间的平均消费）都与该学生是不是应该获得助学金有关联，故将他们都作为训练集的特征。

4.1.1.3 图书馆门禁数据 library

1）内容构成：

字段含义：学生 id，门禁编号，具体时间

3684,"5","2013/09/01 08:42:50"
7434,"5","2013/09/01 08:50:08"
8000,"进门 2","2014/03/31 18:20:31"
5332,"小门","2014/04/03 20:11:06"
7397,"出门 4","2014/09/04 16:50:51"

2）数据表分析：

该表的一条记录代表了一位学生的进入或者离开图书馆的记录，图书馆的开放时间为早上 7 点到晚上 22 点，门禁编号数据在 2014/02/23 之前只有“编号”信息，之后引入了“进门、出门”信息，还有些异常信息为 null。

“library_time_spand”：

根据这一特点，我们统计图书馆门禁数据后发现，有很多看似不合理的数据，例如某一学生在某日内有若干次进出图书馆的记录，且时间相隔较短；又如某学生在某天内有进出图书馆的记录，但最后一次是进门记录而不是离开记录。针对这些情况，我们有意忽视“门禁编号”的信息，对一个学生，将其当日记录时间最早的图书馆门禁记录所对应的记录时间作为该学生当日的进馆时间，将其当日记录时间最晚的图书馆门禁记录所对应的记录时间作为该学生当日的出馆时间，来求出该学生当日的图书馆学习时长。通过对所有记录进行统计，得出每一位学生的图书馆学习总时长“library_time_spand”，并将其作为训练集的特征之一。

“library_times“ ：

由于在计算特征“library_time_spand”时忽略了每天进出图书馆的次数，为了减少这种方式可能产生的副作用，我们又分别统计每位学生进出图书馆的总次数“library_times”,并将其作为训练集的特征之一。

4.1.1.4 学生成绩数据 score

1）内容构成：

字段含义： 学生 id,学院编号,成绩排名

0,9,1
1,9,2
8,6,1565
9,6,1570

成绩排名的计算方式是将所有成绩按学分加权求和，然后除以学分总和，再按照学生所在学院排序

2）数据表分析：

”score”：

该表的一条记录代表了一位学生的成绩情况，还包含了他所属的学院，我们通过分析，认为一个学生是否能够获得助学金，以及获得助学金的额度与该学生的成绩存在联系。于是，我们将原数据中的“成绩排名”作为特征“score”并且作为训练集的特征之一。

“scorerank_divided_by_stunum”:

由于每个学院的学生总数不同，所以如果单独使用“score”作为特征，可能会导致一些问题。所以我们又在原有数据上统计出各个学院的总人数，并用

“scorerank_divided_by_stunum”（学生成绩排名除以所属学院学生总人数）作为特征，来平衡“score”带来的负面影响。

4.1.1.4 助学金数据 subsidy

1）内容构成：

字段含义： 学生 id,助学金金额
10,0
22,1000
28,1000
64,1500
650,2000

2）数据表分析：

该表的一条记录表示已知的某一位学生获得的助学金金额，其值可能去 4 个，分别是 0,1000,1500,2000。这是我们最终需要对测试集进行预测后得到的结果。

“propotion_of_1000”、“propotion_of_1500”、“propotion_of_2000”：

这三个值分别代表某个学生所在学院获得 1000 助学金、1500 助学金、2000 助学金的人数占所有获得助学金学生人数的比例。因为每个学院的情况不一样，学校给每个学院安排的助学金名额也可能存在较大差异，故计算出这三个特征，并将其作为训练集的特征。

4.1.1.5 学生成绩数据 score 联合 一卡通数据 card

“score_rank*consume_rank”：

我们分析认为，一个学生是否属于贫困生，是否应该获得助学金，应该获得多少助学金应该与他的学习成绩，消费情况这两项关系最紧密，为了体现这种关系，我们计算出“score_rank*consume_rank“（学生成绩排名乘以消费排名）”，并将其作为训练集的特征之一。

4.1.2 多表聚合关联分析

本系统的数据表众多，一个学生的数据分散在各个不同的表里头，一个表里头有多条相同学生的不同数据记录。为了更加全面的描述学生，我们建立了一个个模型来描述一个学生的所有特征。我们以学生号为连接属性，将上面提到的各种特征进行连接，得到一张总表，在总表中，每个学生的信息只有一行，其主键为

学生号，其余字段为上述所有特征的值，在行末再添加一个该学生真实助学金情况。

取全学号

由于数据源本身问题，存在部分学生在 subsidy 表中有记录，而在其他表中无记录的情况，这就给最终表的生成造成的影响。如果直接使用学生号作为连接属性去生成最终表，那么会有很多学生因为在某一张表中没有记录而导致其在最终的表里没有记录。为解决这一问题，我们采取的方法是：首先，将 subsidy 表（该表中的学号是全的）中的学生号都提取出来；接下来将所有学号插入到最终表中，至此我们就得到了一张每一行只有学号，而其他信息均为默认值的表。

4.1.3 设置默认值

上面提及了某些学生可能因为记录的缺失而无法计算一些特征值，对此我们采取的办法是给每一个特征项设置默认值，这样在之后的模型训练中，默认值可以作为特殊的一种取值，不会影响模型的训练。

4.1.4 利用 Update 语句更新特征值

在上一部分设置了特征值，之后再以 3.2.1 中得到的完整学生号作为 where 语句连接值，一次计算所有学生的每一个特征，如果该学生在该特征上有值，则用 Update 语句进行更新。

4.1.5 最终表

经过上面的 3 个步骤，至此我们就得到了一张最终表，在对该表进行一定转换后就可以用于模型训练。

4.2 数据处理

4.2.1 语义分析法

由于对贫困生的判断除了是否有贫困证明以外，并没有一个统一、标准的判定流程。不能明确的说通过哪几个特征的值就能判断出结果。因此，我们充分挖掘

特征实际的语义，挖掘其与贫困程度可能存在的关联，我们通过分析出可能与贫困生判定有关的大量特征，再进一步分析学生画像，从而从学生画像（特征）的角度来判断贫困程度。

4.2.2 数据不平衡处理

在所给的数据集中，各种类型学生的人数比例相差比较大，此时用标准算法去解决一定会很困难。传统算法往往偏向于多数类，因为他们的损失函数在没考虑数据分布的情况下优化如错误率等量。最坏的情况是，小类别样本会被认为是大类别的异常值而被忽略，学习算法简单的生成一个平凡分类器，将每个样本都分类为大类别。

各类别人数比例不平衡，获得助学金人数远少于没获奖人数，如下图。

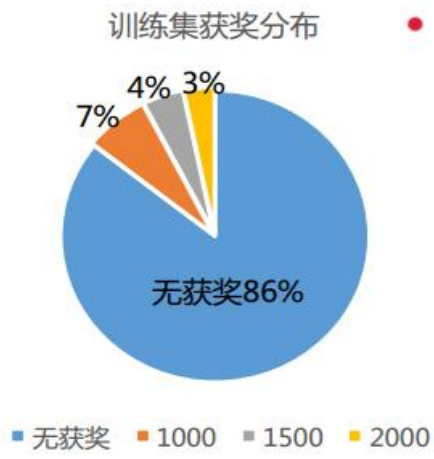

图 4 训练集获奖分布

我们的目标是找出可以获得助学金的学生的模式，并且还要能区分不同等级助学金之间的差异，对初始训练集来说，属于在大量非目标数据中找少量目标数据的问题。如果不解决数据不平衡问题，就必然会使预测结果偏向不获得助学金这一类型，这违背了我们的目的。

故我们采用过采样的方法，在创建训练数据集时，首先记录下各类型样本的数量，接下来比较各类型样本的数量。若出现任意两种类型的样本数量比超过 4:1，则将数量较少的那一类的所有样本复制一遍，再重新比较，直至任意类型样本的数量比都不超过 4:1。

4.2.3 离散化特征

要求离散化的原因：

① 算法需要：众多数据挖掘算法（如决策树、NativeBayes 等）只能处理离散化的数据

② 使模型结果更稳定：把数据分级，而不是原数据可以破除极端值和异常值影响

依次离散化各个特征：

1）选取一个未转化的特征

2）所有记录按照该特征上的值进行升序排序，得到一个有序的结果集合

3）分别取出该集合的 1/4 处的记录，1/2 处的记录，3/4 处的记录，以及集合的最后一条记录

4）用取出的记录在当前选择特征上的取值作为分界的依据，将所有记录的该特征转化为 1,2,3,4 这 4 者之一

5）若还有未转化的特征则返回 1）

6）结束

离散化的优势：

根据上述方法进行特征离散化的处理，可以确保每一条记录在每一个特征上仅有一个唯一的取值，方便后面的模型训练

使得每一特征中各个等级的样本数量大致相同，保证了该特征的可用性

4.3 特征选择

经过之前的一系列操作，最终得到了完整的数据集。该数据集共包含 82 个数据项，数据项 ’ student_num ‘ 为主键，其余 81 项作为特征项。特征项如下：

student_name	avg_stay_out_time	market_min_amount
student_type	canteen_total_amount	other_min_amount
activity_num	market_total_amount	transaction_times
activity_avg_level	other_total_amount	canteen_amount_divide_by_consumption
activity_last_time	charge_total_amount	canteen_times
participation_avg_point	snack_total_amount	Consumption
hornorary_rank	exercise_total_amount	median_of_canteen
hornorary_times	study_total_amount	median_of_market
library_borrow_times	charge_day_max_amount	median_of_charge
library_study_time	exercise_day_max_amount	median_of_snack
library_week_study_time	snack_day_max_amount	median_of_exercise
gpa	study_day_max_amount	median_of_study
score_rank	market_day_max_amount	median_of_other
subsidy_rank	canteen_day_max_amount	mean_of_canteen
subsidy_amount	other_day_max_amount	mean_of_market
failed_num	charge_max_amount	mean_of_charge
failed_pass_num	exercise_max_amount	mean_of_snack
failed_failed_num	snack_max_amount	mean_of_exercise
social_practice_time	study_max_amount	mean_of_study
is_social_practice_great	canteen_max_amount	mean_of_other
in_out_times	market_max_amount	var_of_canteen
student_grade	other_max_amount	var_of_market
scholarship_rank	charge_min_amount	var_of_charge
scholarship_amount	exercise_min_amount	var_of_other
score	snack_min_amount	var_of_snack
avg_out_time	study_min_amount	var_of_exercise
avg_in_time	canteen_min_amount	var_of_study

图 5 特征字段

至此，数据处理的工作已经基本结束，最终数据表体现出特征项数目繁多的特点。此时数据集主要的问题主要体现两个方面：第一，特性项过多很可能导致在后续的模型训练中出现对训练集过拟合的情况，从而使得训练所得的模型在测试集上的表现不尽人意；第二，上述 81 个特征项对不同的预测功能的适用性相差较大，如果直接将所有特征项投入训练很有可能会画蛇添足，导致所得的模型效果不好。因此，有必要进行特征选择，对特征项进行筛选，对特定的预测目标选出特定的特征项进行训练。

特征选择的方法：在 sklearn 提供的算法模型中包含了 score 函数，该函数的会返回每一个特征项的“贡献度”评分，通过比较不同特征项所获的评分，可以为每一个预测功能模块单独筛选出一个特征项集合用于该模块的模型训练。

本文系项目“基于大数据的学生多维度画像系统”研究成果之一。

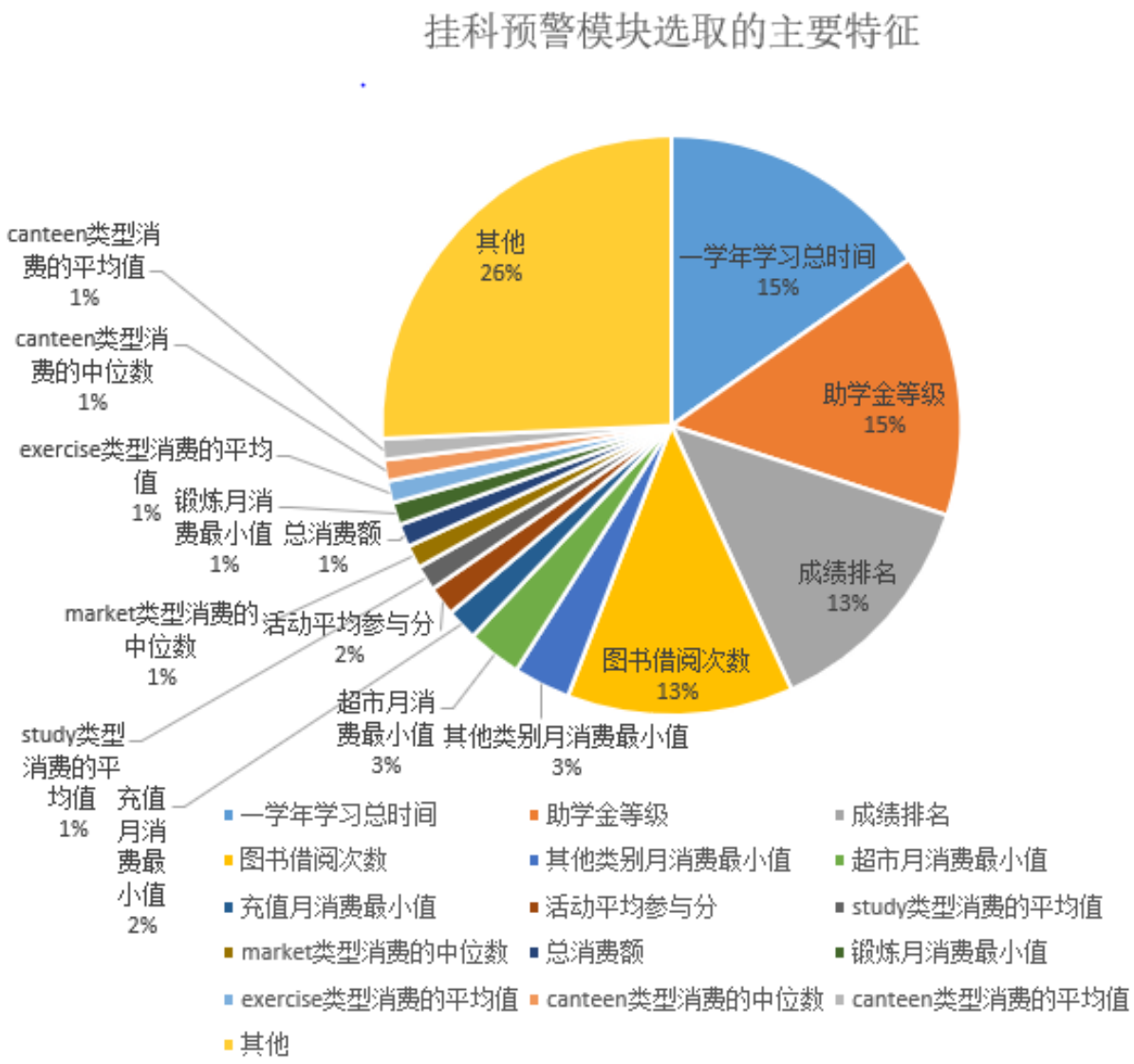

图 6 挂科预警选取的主要特征

4. 4 算法建模

4.4.1 建模流程分析

在对数据进行预处理和特征定义后，我们对已有的各种机器学习算法进行调研，尝试构建不同模型进行训练（主要为适合离散型属性的非线性模型），最后根据定制的模型评估方法（这里使用手动分析）对模型进行评估，剔除对结果影响很小的属性，需要考虑模型的精度、泛化能力等，目的是得出最理想的模型。

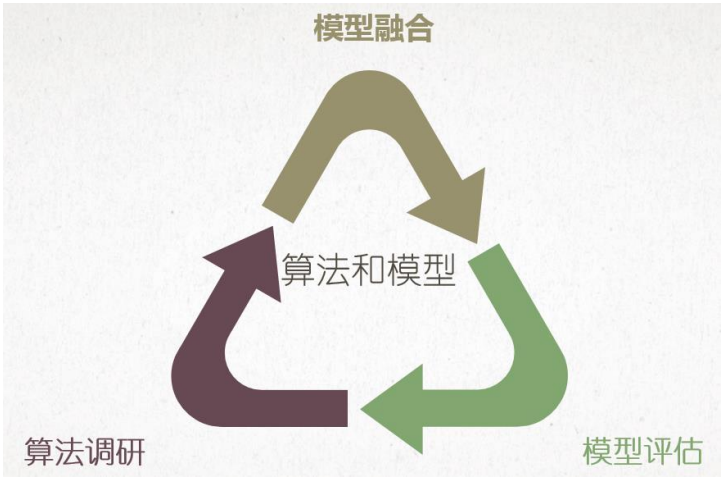

图 7 模型构建流程

每一次迭代都对选取的特征进行不断的增删，经过多次的迭代，使模型更加精准，直至达到预期要求。

4. 5 算法调研

研究各种不同算法的实现原理和它们的适用场景，经过多个模型的相互比较，根据评估指标对比模型在训练集上交叉验证的表现，选择对项目最有利的模型。

4.5.1 算法调研各种适于非线性的分类模型的优缺点

4.5.1.1 C4.5

以信息增益率为衡量标准实现对数据归纳分类

优点：产生的分类规则易于理解，准确率较高

缺点：在构造树的过程中，需要对数据集进行多次的扫描和排序，因而导致算法的低效

4.5.1.2 CART

以基于最小距离的尼基指数估计函数为衡量标准对数据进行递归分类

优点：抽取规则简便且易于理解；面对存在缺失值、变量数多等问题时非常稳健

缺点：要求被选择的属性只能产生两个子节点；类别过多时，错误可能增加的较快

4.5.1.3 ADABOOST

针对同一个训练集训练不同的分类器(弱分类器)，然后把这些弱分类器集合起来，构成一个更强的最终分类器(强分类器)

优点：高精度，简单无需做特征筛选，不会过度拟合

缺点：训练时间过长，执行效果依赖于弱分类器的选择

4.5.1.4 贝叶斯

通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类

优点：算法简单，所需估计的参数很少，对缺失数据不太敏感

缺点：属性个数比较多或者属性之间相关性较大时，分类效率下降

4.5.1.5 KNN

如果一个样本在特征空间中的 k 个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别，则该样本也属于这个类别

优点：简单，无需估计参数，无需训练，适合于多分类问题

缺点：计算量较大；可解释性较差，无法给出决策树那样的规则

4.5.1.6 SVM

建立一个最优决策超平面，使得该平面两侧距离平面最近的两类样本之间的距离最大化，从而对分类问题提供良好的泛化能力

优点：更好的泛化能力，解决非线性问题的同时避免维度灾难，可找到全局最优

缺点：运算效率低，计算时占用资源过大

4.5.1.7 KMeans

输入聚类个数 k，以及包含 n 个数据对象的数据库，输出满足方差最小标准的 k 个聚类

优点：运算速度比 KNN 快

缺点：聚类数目 k 是一个输入参数，不合适的 k 值可能返回较差的结果

4.5.2 在本项目下实现结果较好的几种算法

我们最终发现 ExtraTrees、RandomForest 和 SGD 表现相对其他模型更优。这里面主要原因在于大部分属性是离散性的，且维度较高而且彼此间有相互联系，例如学生各类消费比等级、学生消费方差等属性，这对于构建非线性模型是很有帮助。

4. 6 模型融合

由于单一的分类器在某些方面存在局限性，比如：适于对有高纬度特征的数据进行分类，但是要求各特征间是独立的、和适于对各特征间联系较高的，但对高纬度特征的数据处理困难，这两种算法如果通过某种办法进行优势互补就能得到一个可以处理高纬度且各维度间互有联系的数据的算法模型。

我们使用的是 VotingClassifier 的软投票机制（即不同分类器对分出的不同类所占的权重比值不一样，相比于硬投票，它是不同分类器占不同比值，对所分的类直接作用）。根据对不同分类器设置分类前的权重来使最后的模型偏向某一正确率最高的分类器(坐标数字为按上述的几个算法顺序赋予的权值组合)。

第5章　系统展示

系统整洁大方，数据展示容易让人理解。

5.1 系统主页

主页将各模块的信息部分展示出来，方便即时查看。

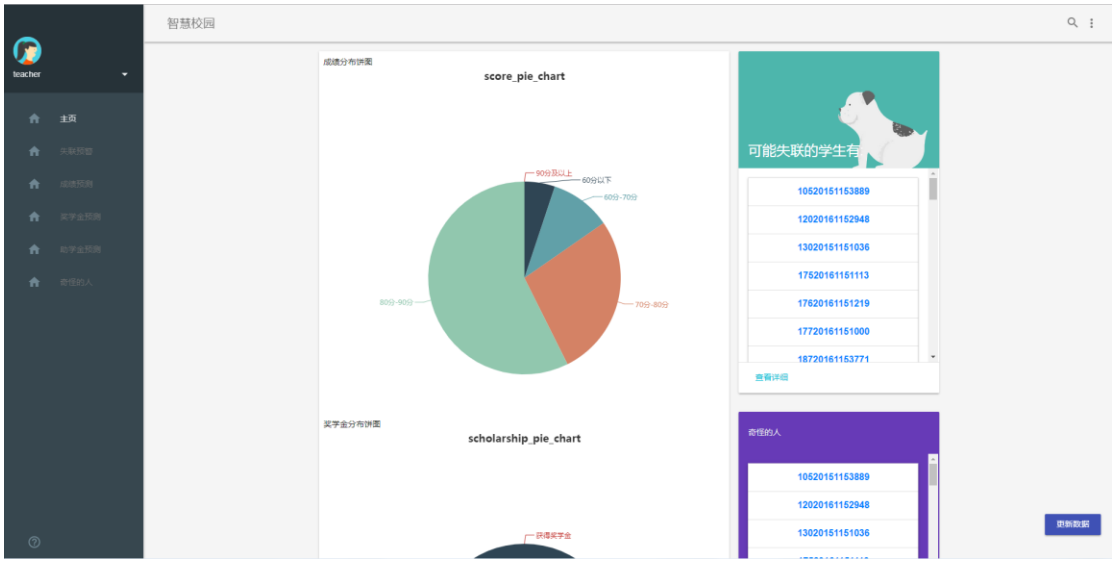

图 8 主页

5.2 各功能模块

5.2.1 失联预警模块

展示出系统预测可能会失联的同学。

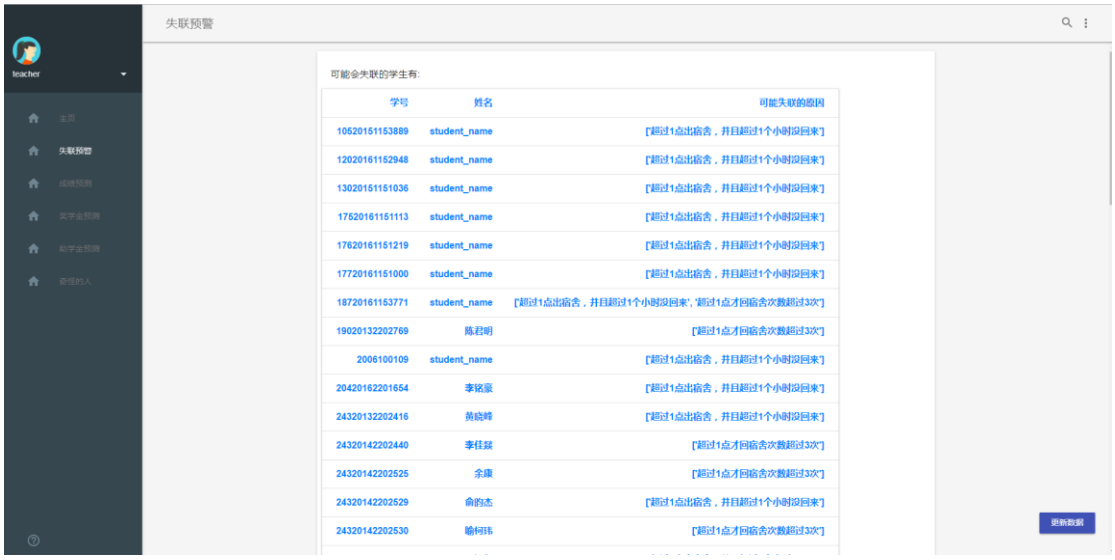

图 9 失联预警模块

5.2.2 成绩预测模块

展示可能会影响成绩的十大特征，并且展示出成绩预测的结果。

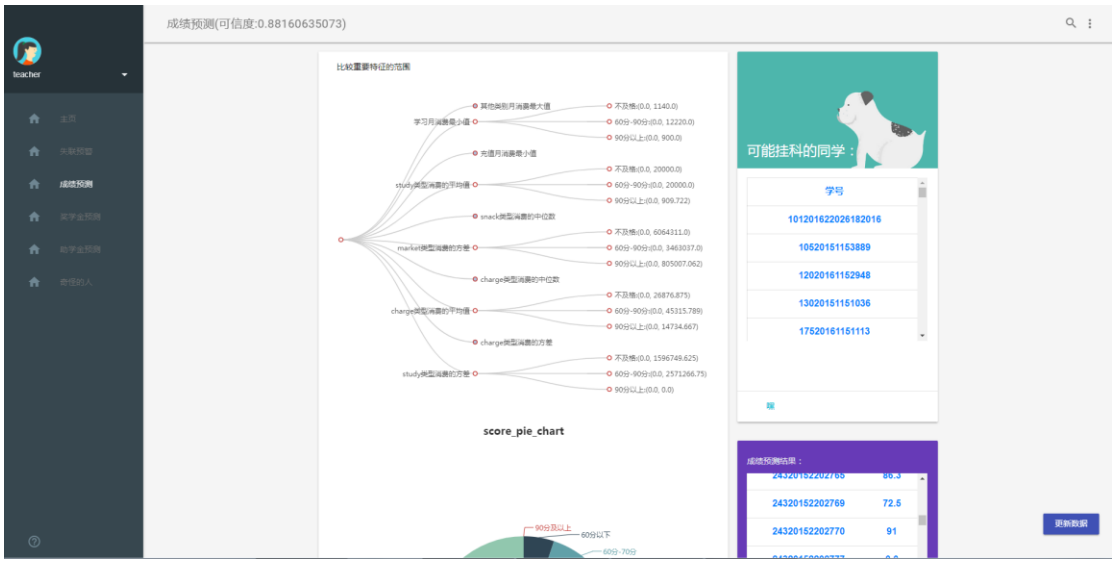

图 10 成绩预测

本文系项目“基于大数据的学生多维度画像系统”研究成果之一。

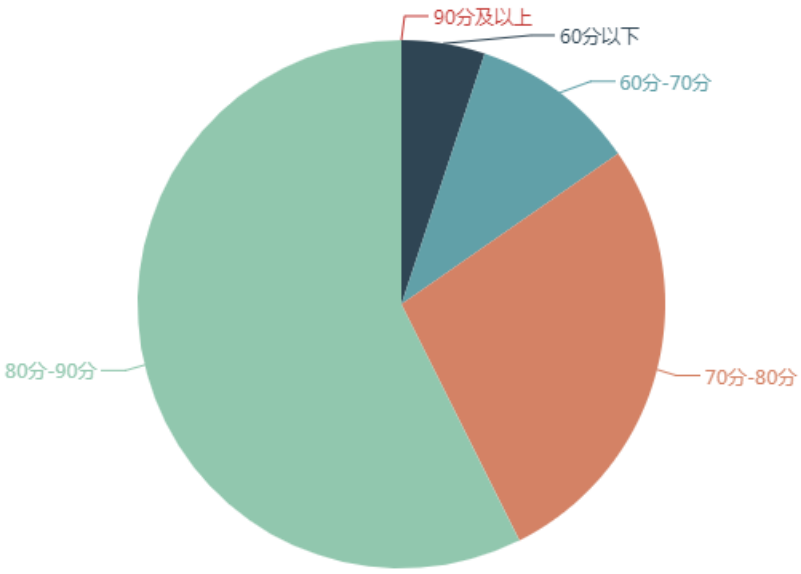

图 11 成绩预测结果分布

5.2.3 奖学金预测模块

展示可能会影响拿奖学金的十大特征。

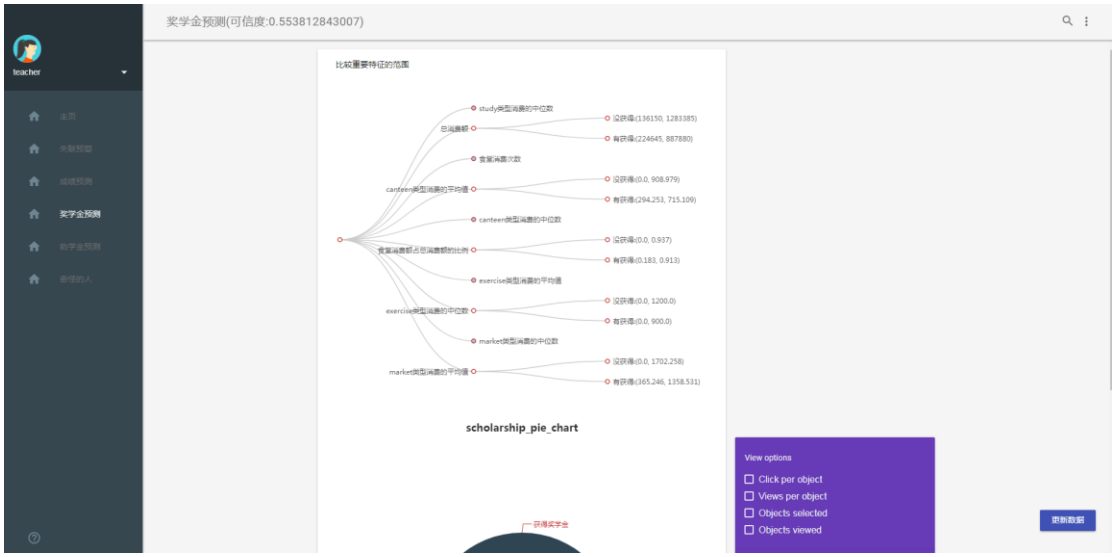

图 12 奖学金预测

本文系项目“基于大数据的学生多维度画像系统”研究成果之一。

5. 2. 4 奖学金预测模块

展示可能会影响拿奖学金的十大特征。

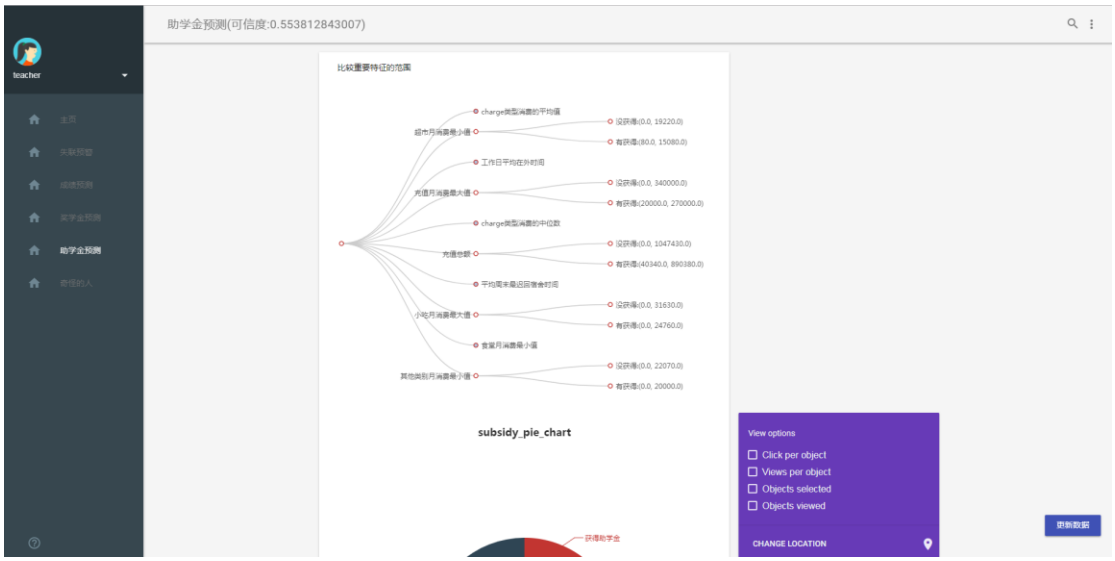

图 13 助学金预测

本文系项目“基于大数据的学生多维度画像系统”研究成果之一。

第6章　总结

此次项目中，我们团队基于 python 编程语言，借助 Sciki−learn, Pandas, Numpy 等开源数据分析模块，采用了多种数据挖掘方法，对学生数据进行深度分析，分别进行了数据预处理，特征工程筛选有意义的特征，预测模块挖掘建模，挖掘结果可视化，并从不同角度进行了统计分析，从而实现了多方位的数据分析挖掘。对所分析的数据，团队通过多次实验、比较分析，寻找并实现了有效的挖掘算法对数据进行建模挖掘。最终，我们是采用 Django 框架进行网站的快速开发，并结合 echarts 灵活的可视化方式，进行多维度的效果展示分析，让用户能够更加直观的看出结果，起初受限于数据量的问题，因为没有足够的数据，我们得出的模型和结果是没有那么强的说服力的，所以我们后期采用的是 GAN 网络通过学习已有数据源的分布特征，来填补缺失值和产生新的数据源，这些数据相对比随机生成的数据，更加真实、有价值，也为我们后期的模型验证起到很大的帮助。经过我们团队数月的努力，本项目目前完成了助学金预测，奖学金预测，失联预警，成绩预测等模块的预测，并为相关的模块生成了较好的模型融合器，对于新的数据，可以把它放入模型融合器快速的进行验证、分析。

后期，我们的系统将往特征选择更加高效智能化方向研究，对于任何的预测模块和任意的特征字段，能够快速筛选出有价值的特征进行训练，生成有效的模型融合器。

虽然只是作为一个大创项目的一小部分，但是这次小项目中我们得到很多有用的知识和手段，这为未来继续帮助大创的进行提供了很好的理论基础和经验。

第 7 章　参考文献

[1]常青.大数据技术在高校智慧校园中的应用[J].信息与电脑,2016,(4):24-25.

[2]刘敏斯,陈少波.大数据时代高校智慧校园建设研究[J].软件导刊,2015,14(8):6-8.

[3]冯玖,李俊玲,张海霞,等.基于数据挖掘的校园一卡通数据应用研究——以石家庄为例[J].石家庄学院学报,2017,19(3):53-58.

[4]王光宏,蒋平.数据挖掘综述[J].同济大学学报,2004,32(2):246-252.

[5]刘卓,崔忠伟.大数据技术在高校智慧校园中的应用[J].软件导刊,2015,14(8):224-225.

[6]李婷,傅钢善.国内外教育数据挖掘研究现状及趋势分析[J].现代教育技术,2010,20(10):21-25.

郑龙天(ORCID：0000-0001-5486-7783)，学生，本科， E-mail:zhenglongtian@outlook.com

李佳声(ORCID：0000-0003-2869-4123)，学生，本科，E-mail:842017615@qq.com;

杨智涵(ORCID：0000-0002-7085-3346)，学生，本科，E-mail：956795145@qq.com；

游侯杰(ORCID：0000-0003-4420-7639)，学生，本科，E-mail：1093831001@qq.com

李特特（ORCID：0000-0002-7048-2944）学生，学生，本科，E-mail：24320152202767@stu.xmu.edu.cn

本文系项目“基于大数据的学生多维度画像系统”研究成果之一。