文章编号：2095-2163(2020)07-0227-07 中图分类号：TP391 文献标志码：A

基于 TensorFlow 框架的可视化大学生行为分析系统设计

周 锐，鲍沛泽，孔 钦，万 凯
（南京大学金陵学院 信息科学与工程学院，南京 210089）

摘 要： 学生生活学习信息的采集、管理与分析工作需要耗费学校、教师、学生大量的时间精力，其中涉及的环节繁琐耗时，同时存在大量无用信息占用资源，而一些有效信息未得到充分利用的情况。为促进高校教育信息化发展，本文设计的基于 TensorFlow 框架的可视化大学生行为分析系统能够在一定程度上满足此需求，解决相关问题。其目的在于利用机器学习算法研究教育数据挖掘，进行学生数据分析，进而为学生自我提升、教师改良教育方法提供指导性依据。通过对本校学生问卷调查数据进行机器学习训练，并利用随机森林算法预测，其预测准确度可达 90%以上，具有实践价值。

关键词： TensorFlow；数据分析；随机森林；机器学习；教育数据挖掘

Design of visual behavior analysis system for college students based on TensorFlow

ZHOU Rui, BAO Peize, KONG Qin, WAN Kai
(School of Information Science and Engineering, Nanjing University Jinling Conlleng, Nanjing 210089, China)

【Abstract】 At present, the information work of the business process of the university management system is relatively perfect, but the collection, management and analysis of students´ life and learning information need to take a lot of time and energy of schools, teachers and students, and the involved links are tedious and time-consuming. At the same time, there is a lot of useless information occupying resources and some effective information has not been fully utilized. In order to promote the development of educational informatization in colleges and universities, the visualized behavior analysis system of college students based on tensorflow framework can meet this demand and solve related problems to a certain extent. Its purpose is to use machine learning algorithm to study and analyze students´ data, and to provide guidance for students´ self-improvement and teachers´ improvement of education methods. Through the machine learning training of the questionnaire data of the students in our school and use of the Random Forest algorithm, the prediction accuracy can reach more than 90%, which has practical value.

【Key words】 Tensorflow; Data analysis; Random Forest; Machine Learning; Educational Data Mining

0 引 言

随着当代高校信息化建设的不断推进，信息科技在学校教学，管理所占的比重也越来越大。大数据分析技术飞速发展及对信息治理的重要性与迫切性的更深入理解，各个高校已经基本建设了基于其不同部门的业务系统，解决了基本业务的管理需求。但是在高校信息化建设中，存在许多问题，如缺少全面综合的管理信息化系统、缺少以人为本的指导思想等[1]。

以教育信息化带动教育现代化，已成为中国加快从教育大国迈向教育强国的重大战略抉择[2]。学生行为数据可视化分析是实现教育现代化的重要步骤。目前国内大多数高校对于学生数据都只是使用 Excel 表格或数据库进行单纯的存储，而无再加工处理，很多拥有潜在价值的数据也因此被尘封，失去了利用这些数据创造价值的可能性。

从满足高校信息化服务建设需求的角度考虑，基于 TensorFlow 框架的可视化大学生行为分析系统（以下简称为“系统”）能在一定程度上解决学生数据的采集整理、处理分析和个性服务。

1 相关概念及技术介绍

1.1 相关概念

数据挖掘（Data Mining）是随着互联网、计算机通信、存储技术发展而产生的一种对大量数据处理分析的需求，本质是在数据库中发现隐藏的知识内容，是当今社会人工智能和数据处理领域的重点研究的课题[3]。教育数据挖掘（Educational Data Mining，EDM）是数据挖掘技术在教育领域的应用[4]。在过去几年中，教育领域和信息领域都发生了革命性的变化，在线学习系统、智能手机应用和社

基金项目： 2019 大学生创新训练计划项目校级（136462019007X）。

作者简介： 周 锐（2000-），男，本科生，主要研究方向：机器学习；鲍沛泽（2001-），男，本科生，主要研究方向：机器学习；孔 钦（1983-），女，博士，讲师，主要主研方向：计算机应用、数据分析、数据质量度量；万 凯（1988-），男，硕士，工程师，主要研究方向：计算机应用。

通讯作者： 孔 钦 qinkong@ vip.126.com

收稿日期： 2020-04-26

交网络为 EDM 研究提供了大量的应用和数据[5]。如今教育数据数量急速膨胀,亟需高效的数据分析处理方法。

预测建模(Predictive Modeling)是通过所收集的数据建立一个模型,利用机器学习等一些算法进行模型训练,从而实现对未来进行预测,是数据挖掘常用的一种手段。

1.2 本文采用的相关技术介绍

机器学习(Machine Learning)是指计算机模仿人的学习行为,对现有数据进行学习,进而改善和完善自身性能的方法,是数据挖掘的常用手段。随机森林算法(Random Forest Algorithm, RFA)是机器学习算法的一种,其实现方法是利用多棵决策树进行分类和预测[6]。

TensorFlow 是谷歌开源的深度学习框架,TensorFlow 的 Python API 较为丰富的实现了反向传播,提供了较为丰富的功能实现方法,且支持 Keras。本系统采用 TensorFlow 2.0 进行大学生行为分析系统的数据分析算法设计。

Django 是一个开源的 Web 应用框架,由 Python 来实现,采用 MVT 的软件模式,拥有严谨的维护,干净的设计, 其所有的代码使开发的 Web 程序遵循最佳实践[7]。

1.3 开发环境介绍和搭建

可视化大学生行为分析系统的开发过程中采用了 Anaconda3 作为 python 项目开发环境。Anaconda 是一个开源的 Python 发行版本,包含了 conda、Python 等其它许多科学包及其依赖包。为实现系统功能,还额外安装了 Django、Tensorflow、scikit-learn、apscheduler、django - apscheduler、djangorestframework、drf-extensions、pillow、coreapi 等多个第三方库。

1.4 研究方法

大学生行为分析系统是高校教育发展需求与人工智能数据分析新技术结合的产物。大学生行为分析系统通过调查问卷提取关键词,刻画学生形象剪影,将学生个性化行为抽象为张量进行分析。以 BP 神经网络模型为基础,构建学生行为分析模型,通过 Tensorflow 2.0 开源框架绘制机器学习算法对样本数据进行训练、分类并预测。利用随机森林算法在对数据进行训练分类的同时,还可以给出各个变量对分类影响的重要性评分,来深度挖掘出每一位学生的知识掌握水平、学习方式偏好、娱乐生活等个性化情况,从而为实现更加个性化的校园管理和服务提供数据支持和科学依据。

利用问卷调查收集的学生学习、生活等各方面的行为相关数据,通过训练得出学生行为分析模型,能够根据新输入的学生行为参数给出预测结果或评估建议。例如:对于如下的学生数据集——“年级、专业、听课是否认真、课上坐第几排、毕业后目标、……、几点就寝、专业兴趣”等一系列特征值,见表 1。通过整数编码生成特征值张量矩阵。并以“绩点”为输出层,系统可以通过线性回归模型训练,达到输入除“绩点”之外的特征信息,自动预测用户绩点的效果。

表 1 绩点影响因素数据集截取

Tab. 1 Data Set Capture of Performance Point Influencing Factors

年级	专业	性别	听课是否认真	上课坐第几排	毕业后目标	……	几点就寝	专业兴趣	绩点
四	电子自动化	男	一般	1-4 排	找相关专业工作	……	12 点前	有兴趣	3.5 以下
二	计算机大类	男	认真	5-8 排	进一步深造	……	12 点前	有兴趣	4.0 以上
一	电子自动化	女	一般	1-4 排	找相关专业工作	……	12 点前	有兴趣	4.0 以下
四	计算机大类	女	认真	1-4 排	进一步深造	……	更晚	一般	4.0 以下
二	计算机大类	男	一般	5-8 排	还未考虑	……	凌晨 1 点前	有兴趣	4.0 以下
二	计算机大类	男	认真	1-4 排	进一步深造	……	12 点前	没有兴趣	4.0 以下
…	……	……	……	……	……	……	……	……	……
四	电子自动化	女	一般	5-8 排	另有打算	……	12 点前	一般	4.0 以下
四	电子自动化	男	认真	1-4 排	进一步深造	……	12 点前	有兴趣	4.0 以下

2 系统功能设计

2.1 需求调研

通过问卷、访谈等多种形式,对学生学习生活、娱乐购物等行为相关方面的分析和研究,最终得到本校学生的相关行为数据集,最终得到系统的总体需求。

系统总体目标是实现一个可视化的大学生行为分析系统。学生可以通过该系统更好把握自己在校

的各种行为，为改善生活学习习惯提供依据；教师则可以实时关注学生数据变动，以方便针对不同学生改进教学方案。

2.2　系统功能设计

系统主要功能：

（1）注册登录功能。用户输入用户名和密码进行注册或登录，将这些数据信息存储到数据库中。

（2）训练数据上传功能。用户可以上传学生数据集用于训练模型，服务器在收到数据后开始同步训练。

（3）模型训练功能。系统对上传的学生数据集进行整数编码转换，生成机器学习所需张量矩阵，利用 TensorFlow 框架构建的模型进行训练。学生数据将通过线性回归梯度下降算法自动进行分类，从而得出一个较为完善的深度学习模型。

（4）数据可视化功能。通过随机森林算法进行模型训练后的结果将以图表的形式展现出来，供用户查看。

（5）特征预测功能。用户上传数据集经过训练后，其模型可用于对未来数据的预测。用户可在系统上通过问卷形式填写模型参数，模型根据所填参数为用户提供预测数据。

2.3　系统功能介绍

用户在注册和登录后可进入系统主页，如图 1。主页左侧栏分有“成绩因素分析报告”、“网购调查分析报告”、“就业意向分析报告”三个已完成的数据分析功能模块。每个功能模块都有负责相应报告提交、数据分析和预测的子模块。

图 1　系统登录界面

Fig. 1　System Login Interface

图 2 为报告提交功能界面。系统在此处提供了样品数据，供下载试用，选择文件后点击“新建报告”即可开始数据集训练。数据集及其训练进度可在下方查看。同时可对数据集进行下载或删除。

图 2　数据集上传功能模块

Fig. 2　Data set upload module

图 3 为数据分析功能界面。雷达图直观显示了就业积极性占比，下方统计了数据集各特征值比例。右上角折线图为随机森林算法计算得出的特征值重要性评分。

图 4 为预测功能模块。系统通过随机森林算法预测功能对自己的特征值进行预测。以“成绩因素分析报告”为例。用户以调查问卷形式填入自己的行为数据，预测后可获得预测结果。（如此次显示结果为“你的预测绩点为：3.5-4.0 左右”）

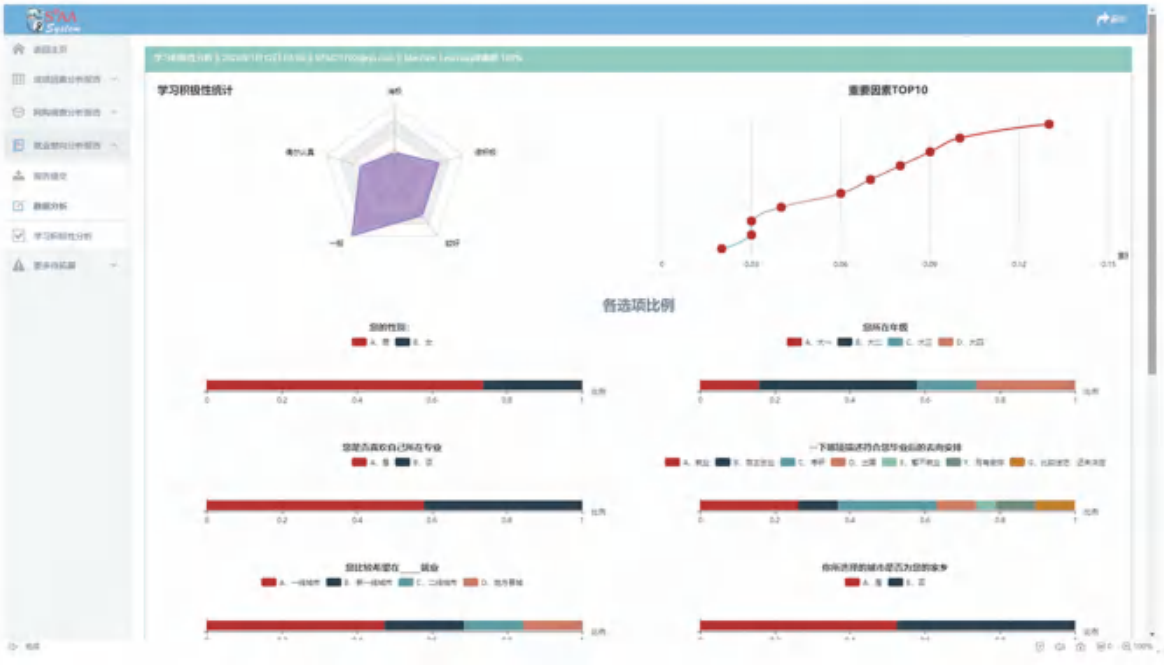

图 3 数据可视化功能模块

Fig. 3 Data visualization function module

图 4 预测功能示例

Fig. 4 Forecast function example

3 系统架构设计

3.1 系统整体设计

如图 5 所示,用户从开始访问网站系统,通过注册模块,将账号信息录入数据库,通过登录模块与数据库交互信息后进入系统。随后可对已实现的三种功能模块进行选择。选择某一模块后,通过报告提交功能,将数据集提交到数据库,同时系统数据分析模块从数据库中提取数据集开始训练。完成训练后,将所得的分析信息反馈给各个模块。

3.2 数据分析功能模块详细设计

如图 6,用户选择具体某一数据分析功能模块后,可选择 3 个子功能模块——报告提交、数据分析、模型预测。选择报告提交后,将数据集提交至系统数据库,进行机器学习训练,训练完成的模型也存

入数据库；用户选择数据分析模块时，可看到可视化数据，如：占比统计、特征值重要性等；用户若选择模型预测功能，以问卷形式填写完成个人行为特征信息，可获得根据训练模型预判的特征值结果。

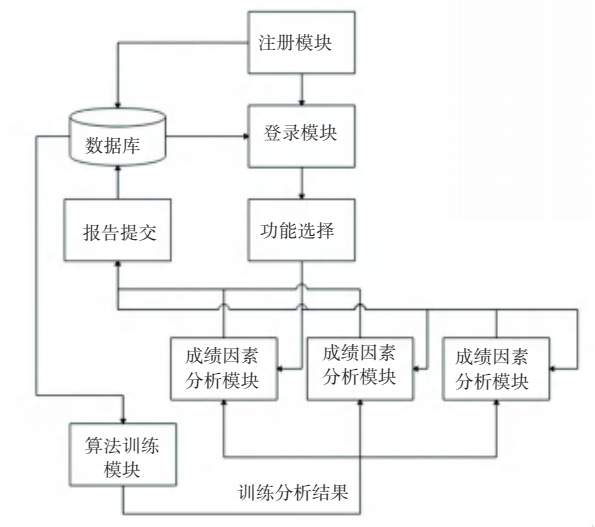

图 5　系统整体功能结构示意图

Fig. 5　Schematic diagram of overall function structure of the system

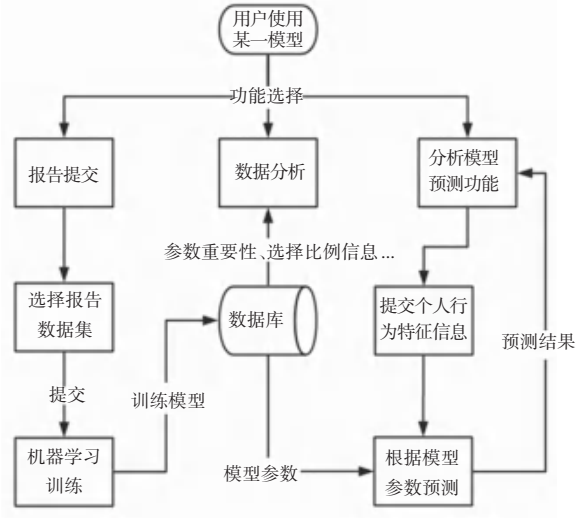

图 6　数据分析模块数据流示意图

Fig. 6　Dataflow chart of data analysis module

3.3　Web 框架结构设计

本项目使用 Django 网页框架来进行开发，如图 7 所示。

views.py 中设计有多个 API 函数，包括训练模型 createModel(obj, filePath, modelPath)、删除模型 deleteModel(request)、预测特征值 predictScore(request)等。

(1) createModel(obj, filePath, modelPath)可以将数据集 obj 存入 filePath，并对其进行训练，完成训练后将训练模型存入 modelPath。

(2) deleteModel(request)根据 request.GET['id']获得用户请求，删除模型的 id，并对 id 所对应的模型进行删除。

(3) predictScore(request)根据用户在前端提交的表单数据，即个人行为特征信息，生成张量矩阵 data，调用函数 predict(data)，获得预测值，以 json 返回给前端。

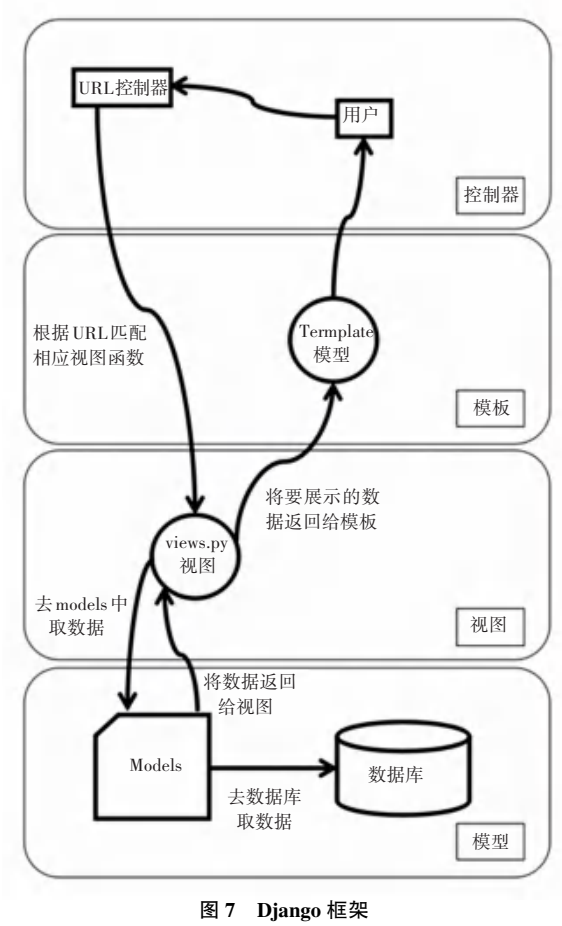

图 7　Django 框架

Fig. 7　Django framework

4　系统核心功能设计

4.1　机器学习算法结构设计

系统核心功能是利用 TensorFlow 框架搭建神经网络。通过随机森林这一机器学习算法进行模型训练、预测和特征重要性评分。如何搭建神经网络以及如何选用激活函数等，将对算法实现效率产生重要影响。

经过实际测试，设置迭代次数为 400 次，训练比例为 80%，学习率为 1e-5，损失函数为交叉熵损失函数(categorical_crossentropy)。在此条件下，分别构建 1 至 4 层神经网络，对每一层使用 BatchNormalization()函数正则化，传入参数 momentum = 0.8，记录下各激活函数训练准确度。

表 2 激活函数-层数-正确率测试结果

Tab. 2 Activationfunctions-layer depths-accuracy test results

层数 正确率	激活函数				
	softmax	*relu*	*selu*	*sigmoid*	*tanh*
1	0.940 2	0.030 5	0.179 5	0.902 3	0.247 9
2	0.968 3	0.030 5	0.199 0	0.941 4	0.221 0
3	0.989 0	0.030 5	0.268 6	0.978 0	0.340 7
4	0.986 6	0.030 5	0.152 6	0.974 4	0.304 0

通过表 2 可以看出几个常用的激活并不都能符合系统的条件。Relu 函数在此条件下无法正常训练。Selu 和 Tanh 函数也无法达到高准确度。测试过程通过后台观察发现,Softmax 相比 Sigmoid 函数计算更快,且更快达到高准确度。相同条件下,Softmax 正确率比 Sigmoid 高。通过对比几个常用激活函数,决定使用 Softmax 作为系统机器学习算法的激活函数。同时通过表 2 发现,在迭代 400 次的条件下,超过 3 层的神经网络已经不能带来更多的效益。因此,出于对算法计算效率的考虑,系统初步考虑采用构建 3 层全连接神经网络层的结构。为了进一步优化算法效率,将针对迭代次数与神经网络层数进行调整测试,得到相关数据,见表 3、表 4。

表 3 层数-精确度/时间-迭代次数关系测试结果

Tab. 3 Layerdepths-acc/time-iterations test results

迭代次数	层数			
	1	2	3	4
100	71.68%	90.96%	91.33%	92.67%
150	86.43%	92.80%	94.51%	94.51%
200	87.91%	93.04%	95.60%/25.92 s	95.24%/26.22 s
250	89.74%	94.99%	96.95%/23.76 s	97.80%/27.14 s
300	91.33%	95.77%/24.70 s	97.19%/24.78 s	97.19%/25.29 s
350	92.43%	95.73%/27.65 s	97.83%/26.78 s	97.68%/23.79 s
400	91.58%	96.95%/25.21 s	98.53%/23.63 s	98.41%/24.14 s

注:正确率指训练结束时正确率;时间指迭代过程中首次连续 5 次正确率超过 95%所用时间,若未达到则不填。

表 4 神经网络层数与所需迭代次数关系表

Tab. 4 Relationship between layer depths of neural network and the number of iterations needed

层数	1	2	3	4	5
达到 95%正确率平均迭代次数	\	255 次	183 次	157 次	146 次

根据表 3 测试结果,可以发现在该项目实验环境下,若设置的迭代次数减少,则可能导致无法获得较高的准确度。根据表 4 的实验结果可以看出,当设置的层数仅为 1 层时,正确率始终无法到达一个较高的数值。根据表 3、表 4,层数超过 3 层,增加网络层数所获得的所需迭代次数减少效益开始急剧减少。由于网络层数的增加导致迭代速度减慢,但每次迭代的正确率提高,复读增加。两种因素叠加后的效果见表 3,在实验环境下,达到 95%的准确度所耗费时间基本都处于 25s,属于误差范围内。

经过实验验证及综合考虑,决定将迭代次数设置为 300 次,而神经网络层数设置为 3 层,以满足高准确、高效率的算法结构优化要求。

4.2 随机森林算法判断重要性

为了提升数据模型的精确度,系统采用随机森林算法来有效提高数据决策树深度和抑制数据噪声。随机森林(RF)是一种统计学习理论,是基于决策树的一种集成学习算法,是广泛应用的一种树状分类器,在树的每个节点 通过选择最优的分裂特征不停地进行分类,直到达到建树的停止条件[8]。它利用 bootsrap 重抽样方法从原始样本中抽取多个样本,对每个 bootsrap 样本进行决策树建模,组合多棵决策树的预测,通过投票得出最终预测结果。

随机森林算法能够将训练过程的数据重要性进行定量化的定义,与此同时,其对线性回归模型中数据关联重要性的定量化定义结果,同样可以作为可视化数据呈现出来,使得其更加符合数据分析的预期要求。

系统利用 sklearn.ensemble 的 RandomForest Regressor( )函数,传入参数 n_estimators = 300, n_jobs = -1, oob_score = True, bootstrap = True,对训练集进行特征重要性判断。

5 结束语

通过对教育大数据的获取、存储、管理和分析,可以构建学习者学习行为相关模型,分析学习者已有学习行为,并对学习者的未来学习趋势进行科学预测[9]。基于 TensorFlow 的可视化大学生行为分析系统,立足于大学生这一社会联系信息丰富的群体,融合了新的机器学习技术,并尝试构建了一套更加科学高效的大学生行为分析服务。与传统的学生数据管理方式相比,具有独特的优势:该系统以师生为中心,表现出学校对学生和教师群体的人文关怀;在具体功能方面,可视化大学生行为分析系统除了具有传统功能外,还结合了高效的机器学习算法,为广大学生提供定制服务,满足了用户多方面的功能需求;可视化数据分析拥有直观性、可读性。

目前的学生数据管理尚未实现高效的智能化,存在巨大的数据价值挖掘空间,可视化大学生行为

分析系统有着很大的潜力。将来可以基于更多数据价值和挖掘手段，利用新技术、新方法将其价值更高效、更合理地挖掘并呈现出来，以减少人力消耗、数据价值浪费，为高校信息化建设中的数据利用提供更多的有效、新颖的解决方案。

参考文献

[1] 吴艳. 高校管理信息化建设中存在问题与对策研究[J]. 科技与创新，2020(2)：114-115.

[2] 张武威，曾天山，黄宇星，等. 我国高校教育信息化重心转移：从硬技术向软技术创新[J]. 高等工程教育研究，2014(5)：102-107，138.

[3] 张凯萍. 大数据时代背景下数据挖掘技术的应用探讨[J]. 赤峰学院学报(自然科学版)，2018，34(8)：52-54.

[4] 陈子健，朱晓亮. 基于教育数据挖掘的在线学习者学业成绩预测建模研究[J]. 中国电化教育，2017(12)：75-81，89.

[5] 周庆，牟超，杨丹. 教育数据挖掘研究进展综述[J]. 软件学报，2015，26(11)：3026-3042.

[6] 陈凯，朱钰. 机器学习及其相关算法综述[J]. 统计与信息论坛，2007(5)：105-112.

[7] 王冉阳. 基于 Django 和 Python 的 Web 开发[J]. 电脑编程技巧与维护，2009(2)：56-58.

[8] 王奕森，夏树涛. 集成学习之随机森林算法综述[J]. 信息通信技术，2018，12(1)：49-55.

[9] 徐鹏，王以宁，刘艳华，等. 大数据视角分析学习变革——美国《通过教育数据挖掘和学习分析促进教与学》报告解读及启示[J]. 远程教育杂志，2013，31(6)：11-17.

(上接第 226 页)

图 8　专业管理界面布局说明图

Fig. 8　Professional management interface layout diagram

上海轨道交通 17 号线列车为 6 节编组，每侧共 30 扇屏蔽门、2 扇端头门、2 扇作为应急疏散门可打开的固定门。屏蔽门关闭且正常的状况下，为蓝色，如图 9；全部打开且正常的情况下，为绿色，如图 10；黄色为中间态，如图 11。

图 9　屏蔽门关闭

Fig. 9　Closed screen door

图 10　屏蔽门开启

Fig. 10　Open screen door

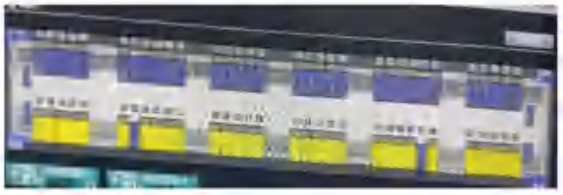

图 11　屏蔽门中间状态

Fig. 11　Intermediate status of screen door

4　结束语

轨道交通在给乘客带来舒适、便捷的乘车环境的同时，需要合理的规划建设、成熟的运营管理以及健全的安防系统，来保证乘客、员工的人身安全。本文也在智慧地铁车站的基础上提供了智慧化安防系统的几种方式，为智慧地铁车站增加了更多的科技含量和安全防护。就目前现状而言，安防系统处于较为完善的阶段，但与国际先进领域仍存在一定的差距，随着科技的不断进步，城市轨道交通的管理也将更加智能化和多元化，安防系统的智慧化也将为整个交通行业乃至社会治安提供坚实的安全保障。

参考文献

[1] 赵云.智能视频监控系统在城市轨道交通的应用与发展趋势[J].科技创新与应用，2016(34)：250.

[2] 黄德贤. 安防系统的发展趋势[J] .硅谷，2013(5)：12-13.

[3] 石杨. 切实加强城市公共交通安保工作[N]. 人民公安报，2014—08—26(1).

[4] 韩巍巍. 轨道交通综合安防系统的应用探讨[J]. 中国新通信，2018，20(12)：155-156.

[5] 郑春晓. 云计算在轨道交通安防系统应用方案[J]. 电子技术与软件工程，2019(6)：190-191.