摘要

随着“智慧校园”的普及和信息化技术的发展，智能化分析技术也在校园中广泛应用。其中，教育数据挖掘的兴起，为教育行业的发展带来了新的方向，即借助大数据技术并基于海量的学生行为数据找到学生日常生活潜在的变化规律，从而辅助教育工作者或者学生个体更快更方便地发现自身在日常学习和生活中存在的问题，并针对性地制定或者改进教育方案。

本文主要研究基于数据挖掘的学生画像关联性分析与可视化建模。该研究是以校园管理平台日常产生的日志数据为研究对象，并对其进行数据预处理，将处理后的数据与数据挖掘算法相结合，挖掘学生行为数据背后的规律和价值，从而引导学生的行为轨迹。本文的主要研究内容概括如下：

一、处理学生行为数据。本文针对校园日志数据进行数据预处理。首先对原始数据进行数据清洗，包括缺失值、异常值处理，数据集成，即对不同源德数据进行二次整合，构建德、智、体、美、劳和消费评价数据指标表，数据规约，即对不同量级的进行[0,1]区间映射，为聚类算法提供分布更加合理的数据。

二、为通过德、智、体、美、劳和消费维度划分学生群体，本文针对预处理后的数据进行聚类分析。根据不同指标的贡献程度设置相应的权重，与DBSCAN密度聚类算法结合，划分出不同群体的学生。最后，根据划分出来的不同簇设置相应的符号标签，构建学生行为画像与相应的标签体系集合。

三、针对学生行为画像进行模型融合分析。以学生各类行为的聚类类别为基础，构建不同学生行为的离散化标签集合，并进一步利用多频繁模式树的分类算法模型，发现学生行为数据的强关联规则识别并描述学生的各类行为特征。

**关键词：**教育数据挖掘，关联性分析，学生画像，可视化建模

ABSTRACT

With the popularization of "smart campus" and the development of information technology, intelligent analysis technology is also widely used in campus. Among them, the rise of educational data mining has brought a new direction for the development of the education industry, that is, using big data technology and based on massive student behavior data to find the potential changes in students' daily life, so as to assist educators or individual students to better Find out your own problems in your daily study and life quickly and easily, and formulate or improve education programs in a targeted manner.

This paper mainly studies the correlation analysis and visualization modeling of student portraits based on data mining. This research takes the daily log data generated by the campus management platform as the research object, and preprocesses the data, combines the processed data with the data mining algorithm, and mines the laws and values behind the student behavior data, so as to guide the students' behavior. behavioral trajectory. The main research contents of this paper are summarized as follows:

1. Processing student behavior data. This paper preprocesses the campus log data. Firstly, data cleaning is performed on the original data, including missing value and outlier processing, and data integration, that is, the secondary integration of data from different sources, and the construction of moral, intellectual, physical, beauty, labor and consumption evaluation data index tables, data specifications, That is, the [0,1] interval mapping is performed on different magnitudes to provide the clustering algorithm with more reasonably distributed data.

2. In order to divide the student groups through the dimensions of morality, intelligence, physique, beauty, labor and consumption, this paper conducts cluster analysis on the preprocessed data. The corresponding weights are set according to the contribution of different indicators, and combined with the DBSCAN density clustering algorithm, different groups of students are divided. Finally, according to the different clusters, the corresponding symbol labels are set, and the student behavior portrait and the corresponding label system set are constructed.

3. Conduct model fusion analysis on student behavior portraits. Based on the clustering categories of various student behaviors, the discrete label sets of different student behaviors are constructed, and the classification algorithm model of multi-frequent pattern tree is further used to find the strong association rules of student behavior data to identify and describe various behaviors of students feature.

**Keywords**: Educational data mining, correlation analysis, student portraits, visual

modeling

目录

[摘要 1](#_Toc103247916)

[ABSTRACT 2](#_Toc103247917)

[第一章 绪论 5](#_Toc103247918)

[1.1 研究背景及意义 5](#_Toc103247919)

[1.2 研究历史和现状 6](#_Toc103247920)

[1.3 研究内容和章节安排 7](#_Toc103247921)

[第二章 理论基础和相关技术 9](#_Toc103247922)

[2.1 引言 9](#_Toc103247923)

[2.2 教育数据挖掘 9](#_Toc103247924)

[2.2.1 教育数据挖掘概述 9](#_Toc103247925)

[2.2.2 数据挖掘预处理 10](#_Toc103247926)

[2.3.3 数据挖掘算法 11](#_Toc103247927)

[2.3 学生画像 12](#_Toc103247928)

[2.2.1 学生画像概述 12](#_Toc103247929)

[2.2.2 学生画像构建方法 13](#_Toc103247930)

[2.4 本章小结 14](#_Toc103247931)

[第三章 基于密度聚类的学生行为画像分析 15](#_Toc103247932)

[3.1 引言 15](#_Toc103247933)

[3.2 日志数据采集与预处理 15](#_Toc103247934)

[3.3 行为指标细分方法 19](#_Toc103247935)

[3.4 基于DBSCAN的行为聚类分析 21](#_Toc103247936)

[3.5 实验结果与分析 23](#_Toc103247937)

[3.6 本章小结 29](#_Toc103247938)

[第四章 基于关联规则的学生行为融合分析 30](#_Toc103247939)

[4.1 引言 30](#_Toc103247940)

[4.2 基于FP-Tree的行为融合分析 30](#_Toc103247941)

[4.3 实验结果与分析 31](#_Toc103247942)

[4.4 本章小结 35](#_Toc103247943)

[第五章 总结和展望 36](#_Toc103247944)

[5.1 工作总结 36](#_Toc103247945)

[5.2 工作展望 36](#_Toc103247946)

[参考文献 38](#_Toc103247947)

# 绪论

## 研究背景及意义

百年大计，教育为本，随着教育普及率越来越高，当代青年受教育水平也大幅提升。2021年中国高等教育毛入学率已达到57.8%，比十年前翻了一倍还多，2022年高校应届毕业生超过1000万，规模创造历史新高[1]。尽管国民教育水平大幅度提升，然而在众多的学生数量下如何对学生个体进行针对化培养，也成为了校园管理者所面临的难题。此外，如何及时地发现学生个体所存在的一些问题并加以纠正，如成绩下滑、沉迷游戏、透支消费等，也是现代教育工作者所面临重要难题。当前，以大数据为代表的信息技术，正与教育行业深度融合，即在海量的校园数据下挖掘出有价值的信息，辅助教育工作者展开教学任务。一方面，将教育数据向智能化、精细化和可视化方向转变，减少了教师的人力成本；另一方面，构建多维度的科学评价体系，有助于提高教育评价精准性、客观性，并促进“智慧教育”的建设进程。

2018年，教育部印发了《教育信息2.0行动计划》[2]，以更好地推动高校智慧校园的建设，指出数字化校园建设将覆盖全校，并在“互联网+教育”的大平台上为校园提供智慧化工作、学习和生活一体化的环境。在当前大数据时代背景下，传统的依托于纸质化记录、分散的电子存档的大学管理模式难以整合所有数据，获得教师和学生的教学反馈，已经无法满足教育2.0行动计划的需求[3]。随着人工智能、机器学习和数据挖掘技术的发展，现有教务系统已经可以实现管理者、教师和学生的智能辅助教学任务。通过将学生学习、生活产生的行为数据以及各类电子存档下的学生的学习结果、个人发展趋势和个人特质等数据整合挖掘并可视化，学生可以从中获取到自身学习情况的定量评估，并对个人情况监督，实现自我提升。教师从中可以及时了解学生的教学反馈，从而优化自身的教学方式。

大数据信息技术在教育学中的广泛使用，也促进了教育数据挖掘（Education Data Mining，EDM）学科[4]的发展，即建立基于大数据的智能模型，对教育基础数据进行直观且可靠的分析处理，并为教育决策提供超过经验判断的结论和预测。此外，为实时反映学生的在校状态，通过搜集学生的专属属性特征，建立学生画像。学生画像是根据用户的数据维度，例如自然特征、社会属性、消费行为、生活规律等所有的相关信息构建的一个标签化、精准化以及可视化的学生模型，辅助相关教师和管理人员掌握学生的动态，从而加强学生的监管，对提高高校学生的整体素质具有重要意义[5]。

## 研究历史和现状

数据挖掘的产生可以追溯到20世纪90年代所提出的数据库中知识发现（Knowledge Discovery in Databases, KDD）[6]这一概念，并在过去的20多年间，数据挖掘技术被生物医疗，商业分析，科学研究，智能制造业等领域广泛应用，并逐步拓展到教育教学领域。2005年，在人工智能国际会议（AAAI’05）[7]上学者们第一次提出了教育数据挖掘EDM这一研究主题。并在此之后，各计算机应用国际会议都会设立有关EDM的研讨会[8]，研究此类课题的学者在此之后组成了国际教育数据挖掘协会，并于2008年在加拿大召开了第一届“教育数据挖掘国际会议”。从图1-1中相关领域的发文量可以看出，08年之前EDM相关主题的发文量相当少，自08年之后每年发文量逐步增加，教育数据挖掘不但成为了一个新的学术研究领域[9]，也随着学者们的深入研究其领域价值也愈加突出。

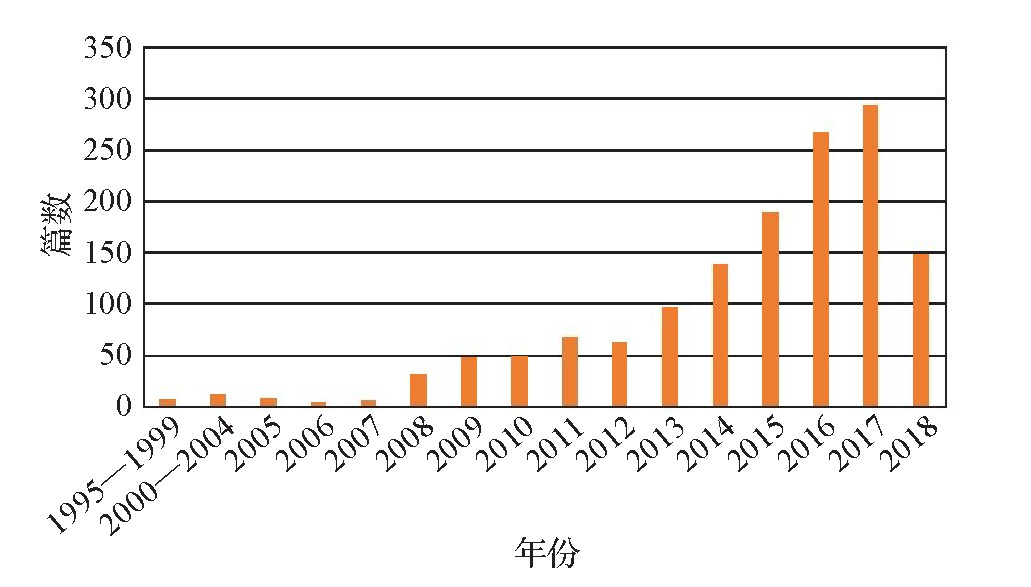


图1-1 相关文献的发文量

国外对于EDM的研究起步较早，早在1995年就有一系列相关综述发表[10-13] 。2008年，Cristo等人就以Moodle系统[10]作为案例分析了数据挖掘在课程管理系统中的应用；2009年，Lee等人[11]提出了利用数据挖掘分析学习者的喜好，结果表明独立型学习者在阅览电子书籍时，花费在书本内容上的时间更多，而现场依赖型学习者花费在书籍目录或者书本选择上的时间更多，并且具有更多的重复访问次数；2010年，Chow等人[12]利用统计学习研究了护理专业相关人员对于对象化和可视化生物医学系统的满意度，结果表明信息技术拥有改善临床学科学习环境的潜力；2012年，美国教育部发布了《通过教育数据挖掘和学习分析存进教与学》[13]的数据报告，对美国国内的教育大数据的研究进展、应用现状和存在的问题加以详细介绍。在此之后，越来越多的学者对EDM展开了更深入的研究。

国内对于EDM的研究起步较晚，大致可以分为三个阶段，首先是萌芽阶段（2002年—2012年），2002年国内有关数据挖掘教育应用的研究才陆陆续续展开，并随着国内外网络课程的大量出现，如国外的Coursera、国内大学的MOOC等在线教育平台的发展以及免费公开课的大量出现和广泛使用，国内的教育数据挖掘研究规模迅速壮大。其中，张媛等人[14]在基于数据挖掘的选修课成绩分析系统上通过数据关联挖掘分析学生成绩优良的相关因素，唐言利用数据挖掘技术实现了远程教学系统中的个性化教学。然而此阶段，研究人员背景较为单一，都为计算机专业背景，且研究内容大多以综述和单一信息分析为主。第二阶段为兴起阶段（2013年—2014年）[15]，在2013年大数据时代到来后，国内的教育专家开始关注这一领域，并在技术应用和技术革新上开展研究工作，其中陈池等人设计了面向在线教育领域的大数据模型，为EDM提供了新的思路。第三阶段为快速发展阶段（2015年—至今），2016年《中国基础教育大数据发展蓝皮书》发布会在京召开，正式发布了国内首份基础教育大数据蓝皮书[16]，以支撑和引领国内教育大数据研究和实践。柴艳妹等人[17]从Web of Science数据库收集08年到17年的文献进行了统计和可视化分析，于方等人[18]介绍了近10年EDM的发展后，提出了以用户为中心的EDM应用研究框架。

而用户画像这一概念最早由Alan Cooper提出[19]，用户画像被定义为描述用户特征的虚拟形象；在此基础上，Amato等人[20]对画像开展了更深入的研究，包含用户兴趣爱好和日常行为习惯等；D.Travis等人[21]给出了较为全面的7个基础条件：基本性、移情性、真实性、独特性、目标性、数量性和应用性，即用户画像的七个基本特性。国外学者Alex Sundt等人[22]指出用户画像可以帮助交流用户体验，Cynthia等人[23]通过创建用户认知与障碍心理模型，巧妙地将用户画像应用到数字图书馆的设计工作中。国内学者王宪朋[24]提出了三个部分，即用户业务特征、用户数据以及用户信息价值共同构成了用户画像，马安华等人[25]则充分考虑了用户的个人偏好，建立了更为细致的用户画像模型。

## 研究内容和章节安排

根据上述内容，教育数据挖掘和用户画像对于学校管理者教育教学，方案制定以及学生培养等方面具有非常重要的现实意义。为实现学生在校行为分析与预警，本文结合现有的数据挖掘算法对多模态的学生校园数据进行处理，并提炼数据潜在的特征和规律，对学生进行针对性的教学指导。研究内容包括教育挖掘技术的基础理论概述，用户画像技术在学生教育中的应用概述，即学生画像技术；校园日志数据的采集、数据预处理和特征选择，以及基于DBSCAN的学生行为画像分析研究；基于多频繁模式树的学生组合特征识别。具体的研究内容可分为以下五个章节：

第一章，绪论。首先介绍了本文的研究背景与意义，接着介绍了教育数据挖掘和用户画像技术的研究历史和现状，最后简单介绍了本文的主要研究内容和结构安排。

第二章，相关技术介绍。本章对教育数据挖掘的工作流程和学生画像的构建方法做出整体分析，并着重介绍了分类、聚类、关联分析和时间序列预测等挖掘方法。

第三章，基于密度聚类的学生行为画像分析。本章对采集到的校园日志数据首先进行了数据预处理，整合并规约了影响学生行为的指标，并从德、智、体、美、劳和消费六个方面刻画学生行为特点。基于此，为更好地适应数据样本的分布，本章通过密度聚类DBSCAN算法对学生行为进行聚类分析，并根据最优聚类k值输出的聚类类别对学生各类行为进行类别标记，并分析了学生的行为数据。

第四章，基于模型融合的学生分类研究。本章基于多频繁模式树的分类算法建立了学生的行为模型，以学业成绩为研究对象，关联分析了影响成绩的各因素，对优秀、普通和不合格学生的行为特征做出进一步分析。

第五章，总结和展望。本章首先对本文所做工作进行总结，然后针对基于数据挖掘的学生画像研究的不足进行了探讨和分析，并对未来的研究方向进行了展望。

# 理论基础和相关技术

## 引言

近年来，人们越来越关注使用数据挖掘技术来解决教育研究领域中的科学问题，教育数据挖掘EDM也应运而生，与传统的数据挖掘不同，EDM更加专注于捕获在教育环境中收集的信息层次性质，从而更好地优化学生服务，改进教育方案。实际上，EDM也可以被理解为数据挖掘在教育领域的应用，而数据挖掘的目的一般可分为两个方面，即描述型和预测型。描述型数据挖掘通过寻找海量复杂数据下的变化规律，简洁高效地提炼和总结数据中的主要特征和信息。本文引入行为画像技术，通过挖掘分析校园日志数据，从德、智、体、美、劳和消费水平六个维度精准刻画学生画像，全面描述和评估学生在校园中的表现。另一方面，预测型数据挖掘则是根据历史数据推断和预测未来数据的趋势走向，在本文中主要基于学生地历史成绩预测未来成绩变化，从而更好地帮助老师了解学生的成绩动向并积极干预。

为利用教育数据刻画学生画像并预测未来的学业成绩，本章首先对教育数据挖掘和用户行为画像技术做了整体概述，接着介绍了相关的数据挖掘算法，为后续的学生画像分析和学业成绩预测铺垫了理论基础。

## 教育数据挖掘

### 2.2.1 教育数据挖掘概述

数据挖掘，也称为数据库中的知识发现（Knowledge Discovery in Databases，KDD），基于机器学习、专家系统和模式识别可挖掘海量数据中潜在的信息，已广泛应用于电子商务、生物信息以及教育研究等领域。其中，数据挖掘在教育领域的深度应用，也使教育数据挖掘EDM成为一门由计算机科学、统计学和教育学组成的新兴交叉学科，通过对在教育环境下所产生大量的校园数据进行深层处理和分析，提取一些有价值的信息以及知识，并从多个维度出发向教育工作者展示学生的学习和生活状态，及时为学生提供学业预警和学习方案调整等功能。在现代信息技术飞速发展的背景下，EDM可分为以下两个研究方向，其一是融合多模态数据分析，即利用可穿戴的感知设备，实时采集学生的时空轨迹和学习过程，利用统计分析和机器学习算法挖掘上述因素与学生学业成绩、心理健康等方面的关联程度，并为各学生制定个性化的学习计划；其二是基于现有数据针对学生的未来的学业成绩、入学率等指标进行预测，并按照指标水平进行等级划分，为处理风险级的学生群体设置各项预警功能，并帮助学校管理者制定恰当的教育决策。

EDM工作流程如图2.1所示，与DM的流程高度相似，均包含数据准备、数据挖掘、分析与评价三个阶段。



图2.1 EDM处理流程图

在本文研究中，EDM主要用于分析校园信息处理系统的日志数据，对日志数据进行预处理和二次加工后，获得影响学生德智体美劳和消费方面表现的数据指标，通过聚类分析提供模型标签，结合事实标签，从而构建出全面的学生画像。通过关联分析方法获得了不同学生群体的行为特征变化规律，识别优秀学生群体和普通学生群体之间的特征差异。此外，基于学生的历史成绩和时序模型预测了学生未来成绩的变化趋势。

### 2.2.2 数据挖掘预处理

EDM的主要研究对象是教育环境中所产生的数据，这些未处理的数据通常包括大量的结构化数据（学生后台信息记录数据库），半结构化数据（日志文件、邮件、XML文档等），非结构化数据（教学音频、视频、综合评价文本等），数据挖掘的第一步就是将这些杂乱无章的数据整合优化成清晰可辨的规则数据，也称为数据预处理。数据预处理包括数据清洗、数据集成，数据加工，数据转换等。

数据清洗：根据某些规则去除原始数据中的缺失值、异常值、冗余重复值和噪声数据，提高数据质量和一致性。对于缺失值，我们可以根据缺失值所在属性的重要程度和分布情况进行处理。若缺失率少且属性重要程度低，可以采用平均值、中位数、以及全局常量分别对数值型和类别型数据进行填充；若缺失率较高且属性重要程度低时，可以直接删除该属性，但是会一部分潜在信息；若缺失率高且属性重要程度高时，可用通过插入法和建模法进行填充。插入法有随机插补法和拉格朗日插入法等。随机法是从总体中抽取某几个样本代替缺失样本；拉格朗日插值法则是找到一个多项式，可以给出一个恰好穿过二维平面上已知离散点的多项式函数，缺失值由所得函数值代替。建模法则是利用回归，贝叶斯，随机森林等模型根据已有的值预测缺失值，准确率较高，但是属性之间的相关性也随之变大。对于异常值，即在原始数据中明显偏离其余数据的离群点，可以通过可视化的箱装图观测并剔除，也可以根据数据集的描述性统计量与数据点比较，发现是否存在不合理的值。对于重复项的判断，基本思想是“排序+合并”，先将数据按照一定规则进行排序，接着比较邻近记录是否相似来检测记录是否重复。

数据集成：经过清洗后的数据一般分布在不同的存储节点上，而数据集成则可以将多个文件或多个数据库中的异构数据进行合并，存储在统一的数据表或者数据库中并针对特定的挖掘任务进行数据重组。在进行集成的过程中，一般需要考虑实体识别和冗余重复问题。实体识别需要考虑来源于不同的信息源的等价实体，可能存在属性名称、内容或单位的不一致问题，往往通过元数据来避免此类问题。另外，经过集成后的实体可能会出现属性内容重复，属性相互包含的关系，可通过相关性分析检测到。

数据变换：集成后的数据可能存在量纲不一致，导致属性数值之间的差异过大，数据的特征及分布情况不明显，导致数据挖掘算法难以有效地发挥其价值。因此数据变换通过对数据进行规范化处理，使数据更加规整，例如标准化、Z-score变换、对数变换、Box-cox变换。标准化主要是将数据缩放到[0, 1]范围内，避免数据的分布过于广泛，但是该方法极易受到异常值的影响。Z-score变换通常用于将数据变化为正太分布，主要基于一般的统计分析方法都立足于正太分布数据上的规律。Box-cox是统计建模中常用的数据变换方法，用于连续响应变量不满足正态分布的情况，可以在一定程度上减小观测的误差和预测变量的相关性。

### 2.3.3 数据挖掘算法

EDM的第二阶段便是数据挖掘，运用知识发现和数据挖掘算法提取用户感兴趣的知识模式，并利用数据可视化进行展示。在经过数据预处理后，便可以确定规范化后的学生行为数据和属性，进而实现学生行为画像建模。在建模过程中，可以选择常见的数据挖掘算法，可以使用统计分析方法。在本文中，首先对预处理后的学生校园数据进行汇总和统计分析，得到可以评价学生消费水平，学业水平等统计指标，例如月平均消费金额，月消费峰值，每学期平均学业成绩等。接着，分析各学生属性指标和统计指标的重要程度，选择有助于刻画学生画像的数据，结合数据挖掘算法，对学生德智体美劳和消费等方面进行行为建模和画像分析。数据挖掘算法主要有分类、聚类、关联分析和时间序列预测等。

分类是数据挖掘中一种重要的方法，即通过学习海量的数据构建分类模型，并使用模型预测给定数据的类标号。在学习阶段，分类算法通过分析从训练集中“学习”数据规律构造分类器，训练集提供每个训练元组的类标号，这一阶段也称为监督学习。常见的分类算法有决策树、朴素贝叶斯等方法。最后，依据各类的准确定来评估模型性能的好坏。

聚类是一种无监督学习的挖掘方法，基于不同实体之间的相似度，将其划分为多个类别。相似度度量一般有两种方法，一种是对所有对象做特征投影，一种是距离计算。前者可从直观的图像上反映对象之间的相似度关系，而后者通过衡量对象之间的距离来反映对象之间相似度。聚类分析按照聚类思想可分为基于分层、密度、网格和模型等多种聚类方法。基于分层的聚类是在训练集合的基础上对样本进行过层次分解，直到满足预设条件。基于密度的聚类通过寻找被低密度区域分离的高密度区域来划分不同的类簇，此类方法基于数据对象分布的密度着手，可以形成不同形状的类簇。基于网格的聚类是把对象空间量化为有限数目的单元，这些单元形成网络结构，所有的聚类操作都在该结构上进行。基于模型的聚类是给每一个类簇都假定为概率统计模型，然后找到能够很好拟合此类模型的数据。

关联分析旨在大量数据中挖掘数据集之间的关联关系，进而对某些数据的出现提出预测。一般而言，我们可以根据两个度量指标支持度和置信度分析关联规则。即当一条关联规则的支持度大于当前设定的最小支持度以及它的置信度大于当前设定的最小置信度时，可以表明该关联规则是一条强关联规则。定义关联规则由表示，和是事务中的一项和多项，则可定义支持度为事务和同时出现的概率，即。置信度定义为事务X发生时事务Y发生的概率，即。目前，关联分析规则算法可分为层次算法、搜索算法、数据集划分算法和抽样算法。  
 时间序列预测算法是基于历史一段时间的变化情况，预测未来一段时间内的变化情况。时间序列预测可大致分为：传统的线性时间序列预测，通过经典的数学方法建立随机模型，如自回归模型、自回归滑动平均模型、求和自回归滑动平均模型和季节调整模型等。非线性时间预测，即采用嵌入空间法或神经网络法，特别是混沌时间序列预测、基于神经网络的预测。

## 学生画像

### 2.2.1 学生画像概述

用户画像是对目标用户的具体表示，即根据已获得的用户的个人属性、行为习惯等数据，生成一系列标签来对用户进行具体的，可视化的描述，通过这些标签可以准确地认知用户，直接或者间接地抽象出一个虚拟用户的特征全貌。

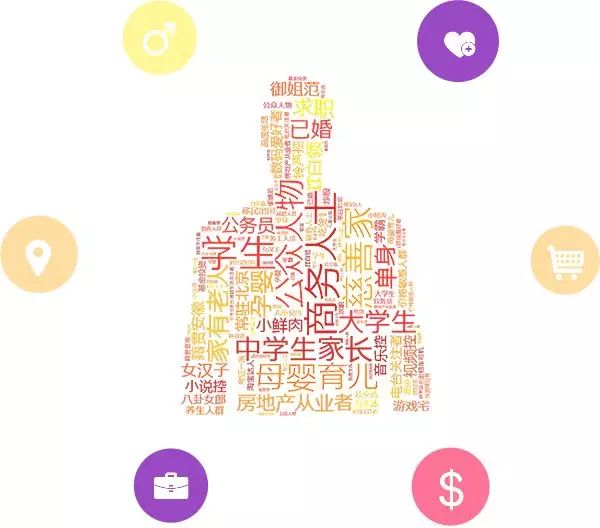


图2.1 用户画像

用户画像的核心作用就是给用户打标签,而标签则是经过高度凝练和提取出来用于标识人物的特征全貌,如年龄,性别,爱好,特长等等,如图2.1所示展示了描述该用户的评价标签,每个标签分别描述了用户的一个维度,每个维度相互联系,共同构成了对用户全貌的具体描述.

### 2.2.2 学生画像构建方法

用户画像是建立在已知数据的基础上，发现用户画像的构建一般可分为目标分析、标签体系构建和画像构建三个步骤。目标分析是在业务数据分析的基础上，构建合适的目标画像，进而实现精准营销、产品改进等目标。针对构建目标画像目标的不同，用户画像的数据来源也不同，此外，数据采集的方式有爬虫采集、平台系统抓取以及问卷调查等方式。对采集到的数据进行预处理和分析之后，我们需要制定合适的标签体系，用于全面地描述用户画像。如图2.2所示，展示了经过层次化后的标签体系，处于最底层的基础数据用于刻画用户的事实标签，事实标签主要以人物的静态信息存在，例如姓名、性别，使用次数、使用历史趋势等，这些标签构建难度低，实际含义明确。模型标签是标签体系的核心，大多数用户的核心标签都是模型标签，即以事实标签为基础，建立与实际业务问题相关联的模型表征，并进行模型分析得到，例如学生兴趣爱好，喜爱的地点等。高级标签基于事实标签和模型标签进行统计建模得出，可以是简单的数学模型，也可以是复杂的机器学习模型，倾向于对学生群体进行智能分类、学业水平、身体素质水平等方面预测。最后在画像构建阶段，利用可视化技术，词云和统计图等，展示用户画像构建的成果。

其中，如何提取标签可成为了学生画像构建的核心内容。根据不同的提取方法，标签构建方法可分为基于本体的构建、基于统计的构建，基于规则匹配的构建，基于主题模型的构建和基于机器学习的构建。基于本体的构建方法通过对特定领域中的属性和关系高度抽象和形式化表达，并将具有鲜明层次结构的本体概念融入到用户画像标签体系中，从而构建清晰的画像标签。基于统计的构建方法是依据统计学手段对用户数据进行分析整理，找出用户的兴趣偏好和行为规律。基于主题模型的构建方法对一系列文档建模，挖掘用户文本内容的数据，然后用维度较低的主题特征作为用户画像标签。基于机器学习的用户画像构建方法，例如聚类、回归和朴素贝叶斯等，主要应用于对用户某些属性的预测。

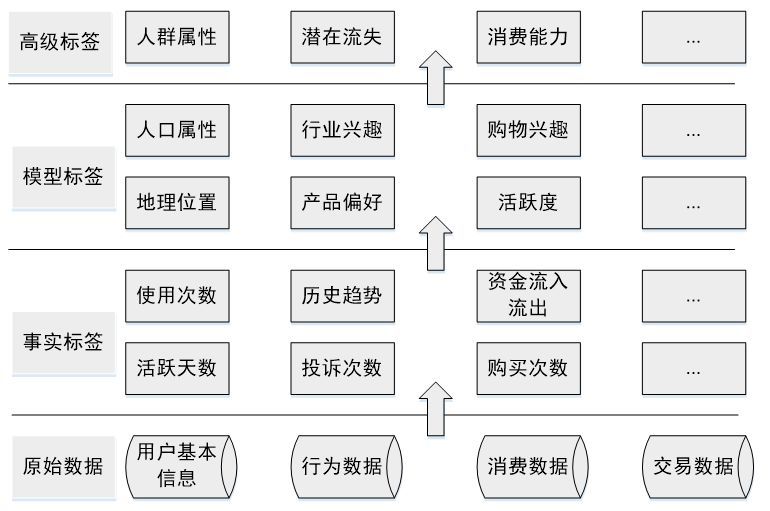


图2.2 层次化的标签体系

## 本章小结

本章介绍了数据挖掘教育和学生画像构建的理论基础和相关技术，首先对教育数据挖掘的概念、研究方向以及处理流程做了详细介绍，接着阐述了数据挖掘预处理以及相关的挖掘算法原理与工作方式，最后介绍了学生画像标签的提取方法和画像的构建方式。

# 基于密度聚类的学生行为画像分析

## 引言

想要对学生特征进行准确又全面地描述，就需要基于海量的校园数据进行数据挖掘，发现多维度多模态数据内潜在的变化规律，从而提取对应的特征。数据挖掘的第一步便是数据预处理，用于消除原始数据噪声和纠正不一致性，提高数据质量。在本章中，预处理具体分为数据采集，数据清洗，数据集成，数据转换四个步骤。数据采集为构建学生画像提供基础信息，数据清洗、集成和转换则是将数据转化为易于分析和算法处理的数据格式。进一步，对预处理后的校园数据进行聚类划分，提出基于聚类的学生细分方法。实验结果验证了本章方法的有效性和可行性。

## 日志数据采集与预处理

随着信息技术的飞速发展，学校针对学生的信息化管理平台也日趋完善，例如学生学籍管理模块，学科课程管理模块，课题研究管理模块，日常考勤管理模块等，不仅方便了学生的日常学习的计划安排，也利于校方对于全体学生的管理。立足于校园信息化管理平台，每时每刻都会产生大量的学生数据，如何有效地挖掘和分析此类数据，对学生日常行为表现进行全面评估，基于评估结果辅助学校和老师及时调整教育方案，探索新的人才培养模式，成为本文的研究重点。因此，本文通过模拟学校各管理模块、校园一卡通、门禁系统等日志数据构建学生信息数据采集库，如表3.1所示。

表3.1 学生信息数据采集库

|  |  |  |
| --- | --- | --- |
|  | 类别 | 描述 |
| 基本信息数据 | - | 当前在校学生的详细信息 |
| 课外活动数据 | 德、体、美、劳 | 学生参与课外活动的记录 |
| 学科成绩数据 | 智、体、美 | 每学期学生的成绩记录 |
| 课题研究数据 | 智 | 每学期学生课题研究信息 |
| 考勤数据 | 劳 | 每学期学生的考勤数据 |
| 消费数据 | 消费水平 | 每学期学生的消费信息 |

本文的日志数据均来源于学校的学生信息管理平台，然而在抓取数据的过程种由于存在的不稳定因素，包括采集数据时的网络中断延迟、系统数据结构混乱以及学生信息状态更新不及时、缺考、休学等都会造成数据部分缺失。一卡通和门禁系统等日志数据也会学校教职工、校外人员等历史数据，造成数据冗余。此外，课外活动等信息通常借助人为登记，可能造成数据偏差，造成不合理的数据分布。此类原始数据需要进行数据预处理，对于缺失值，本文采用数据插补的方法。例如，学生某月的消费金额视为缺失值，再利用拉格朗日乘子法对其插补。首先，求解已知该学生n次某门考试成绩的n-1次多项式，计算方法如(3-1)所示：



将n个学生的消费金额坐标代入公式，得到公式(3-2)：



其次，解出拉格朗日插值多项式，如公式(3-3)所示：



最后，应用拉格朗日多项式，代入缺失值对应的，进而得到对应的消费金额。

对于异常值，首先分析异常值产生的原因，再判断是否将含有异常值的学生数据进行剔除，若发现含有异常值的是无效数据类型，则进行删除或者重新填充。若是有效数据类型，本文可以利用箱装图识别异常值。箱装图中有四分位点内距（Inter-Quartile Range，IQR），指上四分位数与下四分位数之差，其值越大，表明数据对象行为突变程度越大，相反说明数据对象行为数据的突变程度越小，异常值则定义为大于或者小于的数据点。

如图3-1所示，对某个班级的学生各科成绩进行异常值分析，橙色点代表异常值点，可以看到只有语文和数学两个科目有少量的异常值，分布于箱装图的上下边缘之外，我们将其剔除或者填充；箱子中的直线为各科成绩的中位数，中位数大概位于中心位置，数据分布较为对称。

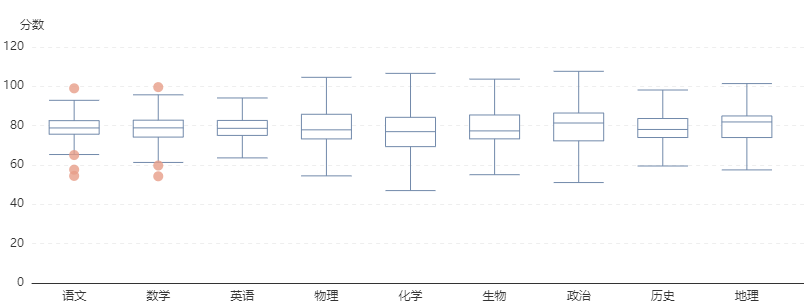


图3.1 成绩绩点分析

我们将清洗后的日志数据进行归类和整合，通过德、智、体、美、劳以及消费水平六个方面刻画学生行为画像，从表3.1可以发现学生的不同行为由多方面数据所决定，因此我们提取并整合每类日志数据中的量化指标用于刻画学生的不同行为，进而得到学生行为画像。首先，我们根据经验分布模拟生成各类量化指标，并使用CSV（Common Spearated Values，逗号分隔值）文件格式统一存储为表格数据。

学生的学科成绩是反应学生在学校的学习态度和知识掌握程度的重要指标，一般而言，成绩越好的学生，学习态度越端正，知识掌握越牢固。本文收集了学生每学期每次考试的文化课成绩以及课题研究数据，用于评价学生的知识学习水平。文化课包括语文、数学、英语、物理、化学、生物、政治、历史和地理，每门课都有对应的学分。此外，每位学生都有相应的研究课题，我们记录每位学生课题的研究时长。我们基于上述指标对学生的知识学习水平进行聚类分析并评定相应的标签，如表3.2所示。

表3.2 学生知识学习的评价指标

|  |  |  |  |
| --- | --- | --- | --- |
| 指标名称 | 类型 | 取值范围 | 描述 |
| 考试学科 | 枚举型 | 语文|数学|政治|英语|… | 课程名 |
| 考试成绩 | 数值型 | -2-100 | -2作弊，-1缺考 |
| 学科学分 | 数值型 | 1-3 | - |
| 课题研究时长 | 数值型 | 0-240 | 每学期的投入时长 |
| 课题评分 | 数值型 | 50-100 | 每学期课题考核成绩 |
| 知识学习水平 | 枚举型 | 优秀|良好|合格|  不合格 | 评价知识学习  水平标签 |

表3.4列出了对学生的德/体/美的评价指标，即学生的思想品德、身体素质和才艺能力。本文采集了对应的课外活动数据和课程考核成绩作为上述三个维度的评价指标。活动数据包括活动类型、活动级别、活动次数，活动表现，考核成绩包括体育、音乐、美术等对应的学科考核。在课外活动中，思想品德方面设置了思政比赛，志愿服务两种活动类型，身体素质方面设置了体育比赛类型，才艺能力方面设置了艺术活动类型。学生的德/体/美水平是指在此类指标的基础上对学生进行聚类分析，分别将学生的思想品德、身体素质和才艺能力划分为不同的类簇，根据类簇给学生评定不同的水平标签。

表3.4 学生德/体/美的评价指标

|  |  |  |  |
| --- | --- | --- | --- |
| 指标名称 | 类型 | 取值范围 | 描述 |
| 考核成绩 | 数值型 | 50-100 | 每学期学科考核分数 |
| 活动类型 | 枚举型 | - | - |
| 活动次数 | 数值型 | 1-10 | 每学期参与活动次数 |
| 活动级别 | 数值型 | 1-5 | 级别越高活动越重要 |
| 活动表现 | 数值型 | 1-10 | 每次活动的得分情况 |
| 德/体/美水平 | 枚举型 | 优秀|良好|合格|  不合格 | 评价德、体、美  水平标签 |

学生在学校的勤劳程度也是刻画学生画像的重要维度。在日常的校园生活，一方面我们可以通过校园考勤管理系统得到学生进校、离校时间，规定正常的上下学时间（8:00-9:00入校，18:00-19:00离校），可以得到学生迟到早退次数（8:00之后入校，18:00之前离校），早起次数（8:00之前入校）；另一方面，在校参与实习实践创新创业项目也能反映学生的勤劳程度，详细的评价指标如表3.5所示。

表3.5 学生勤劳程度的评价指标

|  |  |  |  |
| --- | --- | --- | --- |
| 指标名称 | 类型 | 取值范围 | 描述 |
| 学生平均到勤时间 | Datetime | 2019/09/01 08:00:00 –至今 | - |
| 学生平均离勤时间 | Datetime | 2019/09/01 19:00:00 –至今 | - |
| 迟到早退次数 | 数值型 | 0-180 | 每学期缺勤次数 |
| 早起次数 | 数值型 | 0-180 | 每学期早起次数 |
| 实习实践次数 | 数值型 | 0-10 | 每学期实习次数 |
| 单次实习实践表现 | 数值型 | 50-100 | 单词实习得分 |
| 创新创业项目次数 | 数值型 | 0-10 | 每学期参与次数 |
| 单次创新创业表现 | 数值型 | 50-100 | 项目得分 |
| 勤劳程度 | 枚举型 | 优秀|良好|合格|  不合格 | 评价勤劳程度  标签 |

此外，为掌握学生的消费规律，表3.6列出了能够刻画学生消费规律的评价指标，包括每学期的消费金额，月消费频次，月消费金额，消费时间，消费习惯以及消费水平。一般而言，学生在校园里的消费主要是以食堂消费为主，因此分析消费习惯是否规律也能间接推断出学生的饮食习惯是否规律，消费水平同样是根据基于以上指标聚类分析而得。

表3.6 学生消费水平的评价指标

|  |  |  |  |
| --- | --- | --- | --- |
| 指标名称 | 类型 | 取值范围 | 描述 |
| 学期消费金额 | 数值型 | 1-8000 | 每学期学生的消s费金额 |
| 月消费频次 | 数值型 | 1-200 | 每月的消费次数 |
| 月消费金额 | 数值型 | 0-30 | 月消费金额 |
| 消费时间 | Datetime | - | 每次的消费时间 |
| 消费习惯 | 枚举型 | 规律|不规律 | 评价学生消费习惯是否规律 |
| 消费水平 | 数值型 | 高消费|中等消费|低消费 | 评价学生消费水平的标签 |

由此，我们通过分析校园日志数据的可用性，评价学生在校行为的指标，构建了上述可以描述学生行为的“画像”特征库，进而从德、智、体、美、劳和消费六个方面展示学生的行为特点。

## 行为指标细分方法

本节依据数字化校园系统中的数据，构建能够反应学生各方面行为价值的相关指标，重点介绍所需指标权重的设置细节，进而采用数据挖掘中的聚类算法对学生进行划分，最后根据聚类结果分析不同群体学生的行为特点，为学校教育工作者提供更加清晰的学生画像，并全面准确地制定与改进学生地个性化管理方案。

在学生知识学习水平评价方面，我们首先利用每次考试各科学分乘以相应课程成绩，再除以所有课程学分总和计算每个学生的加权平均成绩，作为学生的学业成绩数据。由于存在缺考、作弊等行为，我们直接将违反校规的学生学科成绩置为零。课题的研究时m,a（小时）和课题评分也将作为学生评分指标。本文根据上述指标对学生知识学习的贡献程度分配如下权重，如表3.7所示：

表3.7 学生知识学习的评价指标

|  |  |  |
| --- | --- | --- |
| 指标名称 | 取值范围 | 指标权重 |
| 加权平均学期成绩 | 1-100 | 0.7 |
| 课题研究时长 | 1-240 | 0.1 |
| 课题评分 | 50-100 | 0.2 |

在学生的思想品德、身体素质和才艺能力水平评价方面，通过课程成绩和课外活动得分指标进行评价，课程活动得分通过计算活动等级乘以活动得分所得，

参与活动的次数反映了学生的学习积极性，上述指标对学生德、体、美方面的贡献程度分配如下权重，如表3.8所示。

表3.8 学生德/体/美的评价指标

|  |  |  |
| --- | --- | --- |
| 指标名称 | 取值范围 | 指标权重 |
| 课程学期考核成绩 | 1-100 | 0.7 |
| 活动平均得分 | 1-10 | 0.2 |
| 活动参与频次 | 1-5 | 0.1 |

在学生的勤劳程度评价方面，我们根据学生实际的到勤、离勤时间与规定时间做对比，可以得到学生的按时到勤情况，以月平均到勤频次为衡量学生劳动程度的标准之一，同时还记录了学生的早起频次，即比规定时间早道勤情况，实习实践和创新创业项目表现通过总得分除以参与次数得到。如表3.5所示为上述指标对学生勤劳程度的贡献分配情况。

表3.5 学生勤劳程度的评价指标

|  |  |  |
| --- | --- | --- |
| 指标名称 | 取值范围 | 指标权重 |
| 到勤频次 | 1-180 | 0.2 |
| 早起频次 | 1-180 | 0.2 |
| 实习实践平均得分 | 1-10 | 0.3 |
| 创新创业平均得分 | 1-10 | 0.3 |

在学生的消费评价方面，我们根据学生的学期消费总金额，月平均消费金额，学期消费频次和消费峰值对学生的消费水平做进一步评价。如表3.6所示为上述指标对学生消费水平的贡献分配情况。

表3.6 学生消费水平的评价指标

|  |  |  |
| --- | --- | --- |
| 指标名称 | 取值范围 | 指标权重 |
| 消费峰值 | 1-100 | 0.1 |
| 消费频次 | 1-30 | 0.2 |
| 月平均消费金额 | 1-1000 | 0.3 |
| 消费总金额 | 1-10000 | 0.4 |

## 基于DBSCAN的行为聚类分析

聚类分析是一种能够挖掘数据深层信息与知识的有效方法，以“簇内相似，簇间相异”准则将给定的样本划分为不同的类簇，使研究者能够更好地归纳和发现数据样本的特征。其中，K-means是常用的基于欧式距离的距离算法[26]，认为两个目标的距离越近，相似度越大。虽然K-means易于实现且算法复杂度低，但依旧存在样本形状包容性差，易受异常值影响等问题。

为克服常规K-means算法的上述问题，本文采用基于密度聚类算法DBSCAN（Density-Based Spatial Clustering of Applications with Noise，具有噪声的基于密度的聚类方法），该算法[27]将具有足够密度的区域划分为簇，并将簇定义为密度相连的点的最大集合，将密度低的数据点作为分类的边界。DBSCAN基于一组邻域来描述样本集的紧密程度，它具有两个输入参数：邻域半径和密度阈值minPts，通过这两个参数，我们便可以区分出高密度点和低密度点，即在某个数据点的邻域范围内的数据点数目超过了最小包含点数阈值minPts即为高密度点，而DBSCAN的主要思想就是将满足高密度点的数据点聚类成一个簇。

给定样本集，我们可以得到邻域、密度阈值minPts、核心点、边界点、离群点等定义。

1）邻域：对于，以为中心，为半径的超球体区域称为邻域，其中所有的样本点满足



2）密度阈值minPts：某样本邻域中包含样本点个数的阈值，需要人为提前设定。

3）核心点：对于给定的数据点，在其以为半径的邻域内存在至少minPts个样本点，则称为核心点。

4）边界点：对于给定的数据点，在其以为半径的邻域内个数比密度阈值少，但是该点位于核心点的邻域内，则称为边界点。

5）离群点：样本S中既不是核心点，也不是边界点的样本点，又称为噪声点。

根据以上定义，DBSCAN中数据点之间关于密度存在如下三种关系：密度直达、密度可达和密度相连。

密度直达：如果位于的邻域内，且为核心对象，则称和密度直达。

密度可达：对于和，若存在样本集合，且满足，，且由密度直达，则称由密度可达。此时样本集合中的均为核心点。

密度相连：给定数据点和，若存在核心点，使得由密度可达以及由密度可达，则可以称和密度相连。



图3.2 DBSCAN原理图

如图3.2所示，图中有三个密集区域，因此也有三个聚类簇。均为核心点，为边界点，由密度直达，和密度相连。

算法的主要流程如下：输入为包含有n个数据点的数据集，邻域半径以及密度阈值minPts。首先标记所有的点都为未访问状态，并随机选择一个未访问的数据点，标记该数据点为已访问状态。如果数据点的邻域中至少有minPts个数据点，创建一个新簇，并把数据点添加到中，接着找出附近邻域中所有的样本点，将其加入到类簇中，并以这些样本点为核心点继续寻找，直到不满足密度阈值，最后将这些点连接为聚类簇。将形成聚类簇的点都标记为已访问状态，并在未访问数据点中继续上述步骤寻找聚类簇，最后得到聚类划分后的结果。

DBSCAN算法可以对任意形状的数据集进行聚类，可适用于非凸和凸数据集，并在聚类的时候发现离群点，从而将其排除在外，对异常点并不敏感，聚类结果也受参数设定的影响较小。

为定量地评估DBSCAN的聚类结果，本文采用轮廓系数（Silhouette Coefficient）。轮廓系数同时评估了样本数据点之间的簇内的内聚度和簇间的分离度，可以在未标注的样本数据集下评估聚类效果。

轮廓系数的基本原理如下：首先计算样本到其他样本的平均距离。越小，说明样本越应该被聚类到该簇，又称为样本的簇内不相似度。计算样本到距离它最近的簇中所有样本的平均距离。越大，说明样本越不属于该簇，又称为数据样本与其他类簇的簇间不相似度的最小值。根据簇内相似度和簇间不相似度，定义数据样本的轮廓系数为：

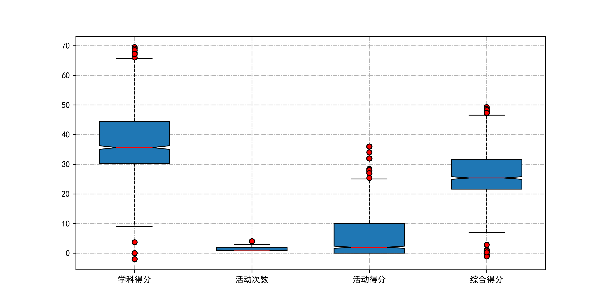
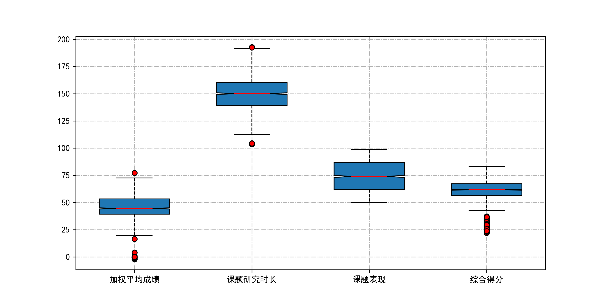


从上式可以看出，的取值范围为，当越接近于1，则代表内聚度和分离度越高，聚类效果越好；当越接近于-1，则代表内聚度和分离度越低，聚类效果越差。

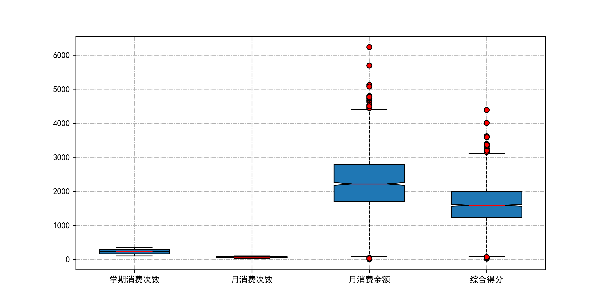
## 实验结果与分析

本文的实验数据通过某高校的数据化校园的管理系统中的日志数据采集和预处理得到，经过预处理后的数据包含了学生基本信息表，学生各项活动参与表，学生的消费信息表以及学生的考勤打卡信息表。时间跨度从2013到2018年，统计了五个学年1753名学生的各项指标，其中不乏有缺失的学生ID，学生记录和各种异常值，在数据预处理中通过箱装图分析，并结合剔除、填充等方法进行数据矫正，进而建立学生评价特征库，随后的效果和性能分析都是基于特征库进行。

针对本章学生行为描述的六个方面：德、智、体、美、劳和消费水平，我们建立相应的指标并根据指标的贡献程度提出不同权重，最终得到描述学生行为的各类评价表。如图3.3所示为学生智、体和消费三方面的箱装图，直观地展示了学生各类行为数据的分布原貌，红色点代表异常值，可以看到小部分学生的评价表有异常点，由于异常点较少，我们将经过数据清洗、集成后的数据视为正常值进行分析。



（a）知识水平 （b）体育水平



（c）消费水平

图3.3 各类行为数据的箱装图

考虑到所涉及到的学生行为数据之间的量级差别较大，为了使得不同量级数据分布在统一的范围内，我们用最大-最小值规范化。即根据相应的指标权重，计算各类行为数据的加权综合得分，包括知识水平、体育水平、品德水平、艺术水平、勤劳水平和消费水平的加权综合得分，并将综合得分进行规范化。图3.4为学生的各行为得分之间的相关性分析矩阵，从对角线处可以发现行为数据基本都符合标准的正太分布，并且从下三角或者上三角的分布图可以观测到学生某一行为数据与其他行为数据的相关分布都趋于一致，说明学生各类行为数据之间的分布变化不明显，因为我们对学生的数据做进一步的分析研究。

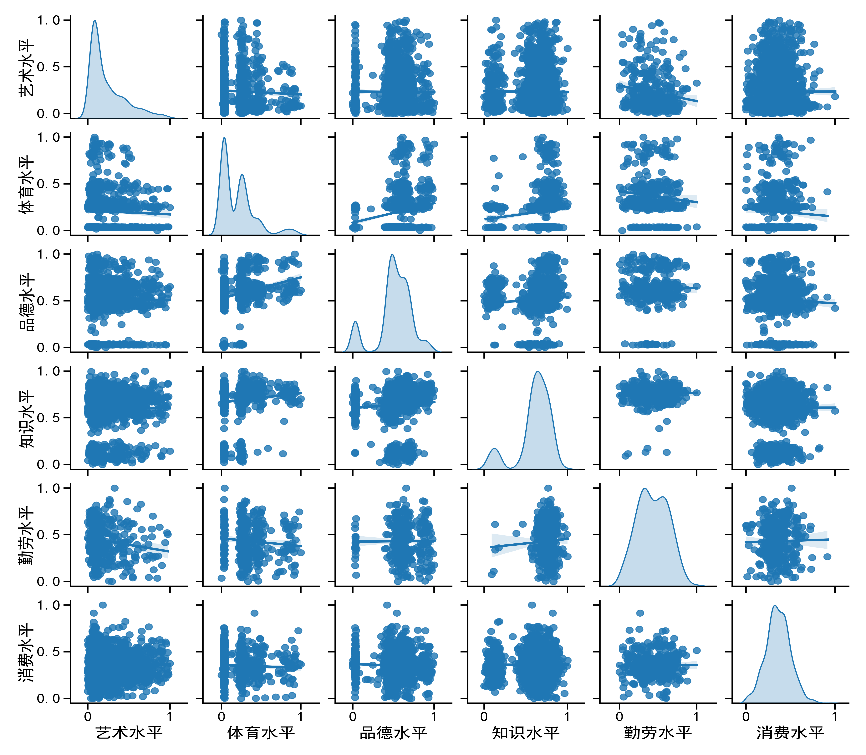


图3.4 各类行为数据的散布矩阵

接着我们基于经过数据规约后的学生行为特征库进行DBSCAN聚类分析，并且通过轮廓系数评价聚类结果。为了达到最优的聚类结果，我们通过对比轮廓系数随着密度阈值minPts和邻域半径的变化情况，确定最佳的算法参数。在最佳minPts确定过程中，由于知识水平和消费水平聚类结果在minPts>4的情况下聚类标签为0，因此在此不展示。

如图3.5所示为不同minPts下轮廓系数的变化情况，轮廓系数曲线变化趋势都是随着minPts的增加先上升再下降，我们取最高轮廓系数对应的minPts即为最优的密度阈值。同理，我们可以得到最佳的邻域半径。

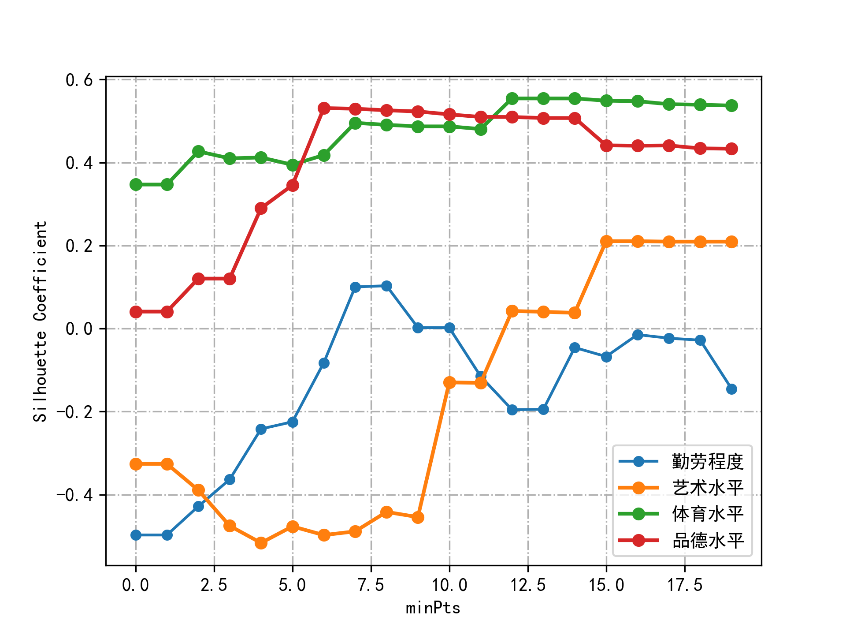


图3.5 不同minPts情况下的轮廓系数

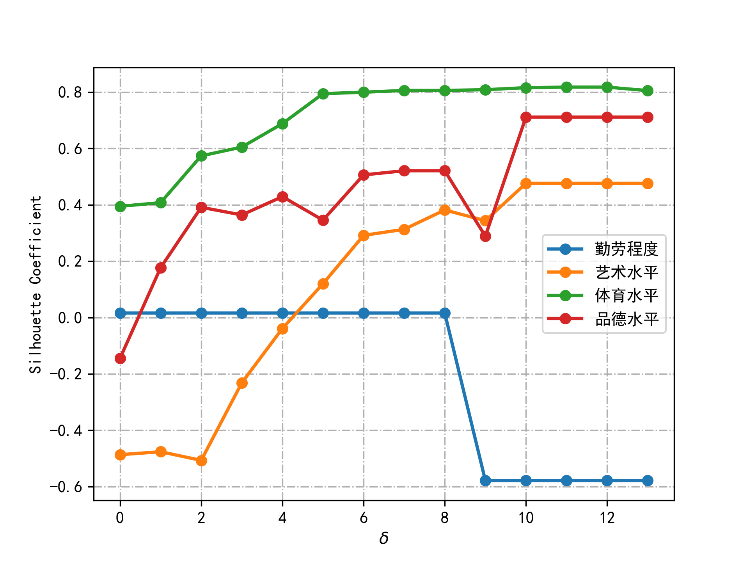
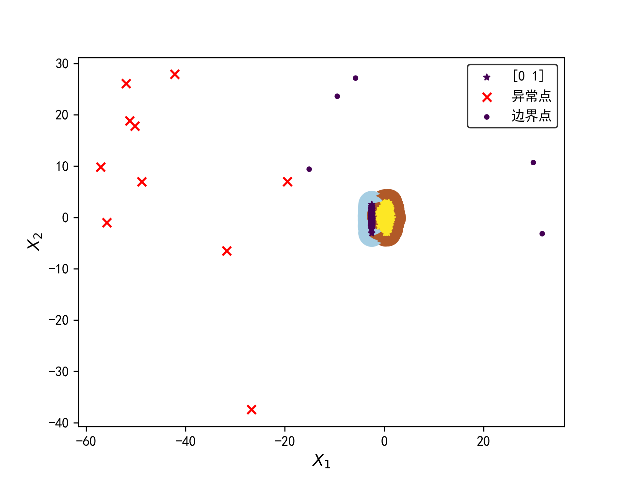
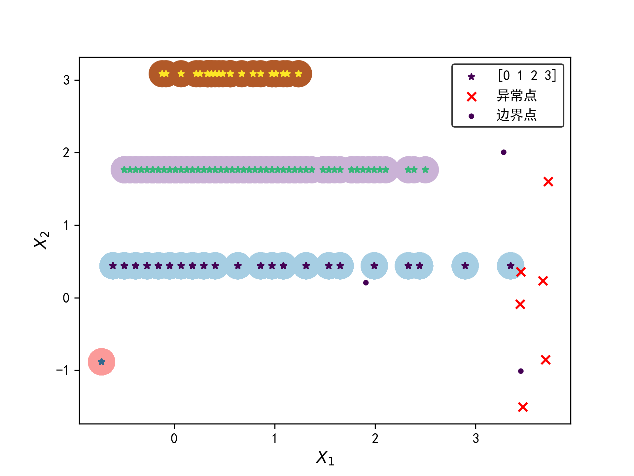
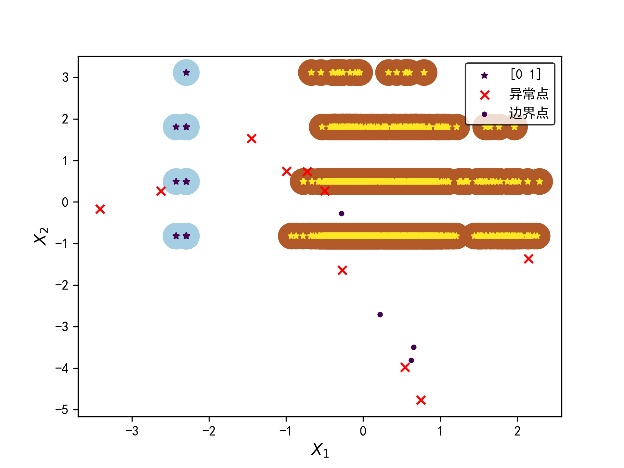
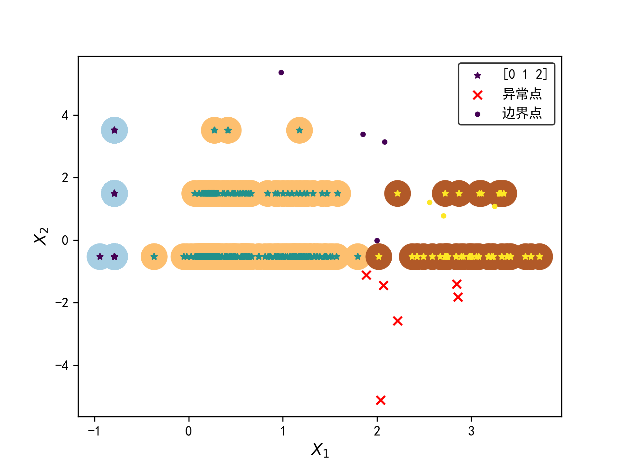


图3.6不同情况下的轮廓系数

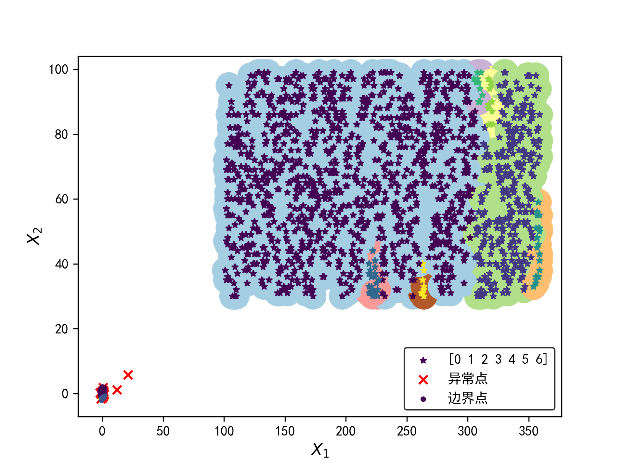
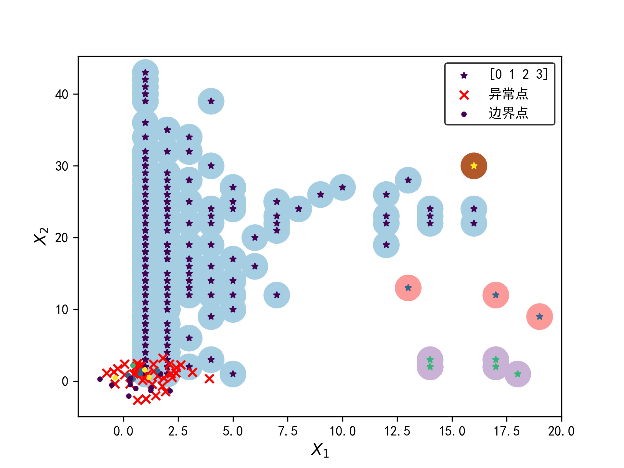
当我们确定了最优的DBSCAN的算法参数时，我们便可以对学生各类行为数据进行聚类分析。根据图3.7聚类分析的结果可知，学生的知识水平被划分为2个类别，艺术水平被划分为4个类别，体育水平被划分为3个类别，品德水平被划分为2个类别，劳动水平被划分为4个类别，消费水平被划分为7个类别。基于此，我们根据不同的类别符号对聚集到同一个簇内的数据做相同的符号标记。

****

1. 知识水平 (b) 艺术水平



(c) 体育水平 (d) 品德水平

****

(e) 劳动水平 (f) 消费水平

我们根据上述聚类结果就能得到以下五个统计表，分别展示了每个簇中综合

得分的均值、人数比例、行为标签和类别符号。表3.7-3.12统计分析了各类行为数据的标签，并重新生成了对应的聚类结果表。具体如下所示：

表3.7 艺术水平聚类结果

|  |  |  |  |
| --- | --- | --- | --- |
| 类别符号 | 综合均分 | 人数占比 | 行为标签 |
|  | 57.91 | 0.88 | 优秀 |
|  | 30.15 | 0.11 | 普通 |

表3.8 知识水平聚类结果

|  |  |  |  |
| --- | --- | --- | --- |
| 类别符号 | 综合均分 | 人数占比 | 行为标签 |
|  | 15.05 | 0.11 | 优秀 |
|  | 12.43 | 0.01 | 良好 |
|  | 10.35 | 0.38 | 中等 |
|  | 6.65 | 0.48 | 差 |

表3.9 体育水平聚类结果

|  |  |  |  |
| --- | --- | --- | --- |
| 类别符号 | 综合均分 | 人数占比 | 行为标签 |
|  | 10.35 | 0.48 | 优秀 |
|  | 7.55 | 0.05 | 一般 |
|  | 6.65 | 0.44 | 普通 |

表3.10 品德水平聚类结果

|  |  |  |  |
| --- | --- | --- | --- |
| 类别符号 | 综合均分 | 人数占比 | 行为标签 |
|  | 10.35 | 0.11 | 优秀 |
|  | 6.65 | 0.88 | 普通 |

表3.11 劳动水平聚类结果

|  |  |  |  |
| --- | --- | --- | --- |
| 类别符号 | 综合均分 | 人数占比 | 行为标签 |
|  | 13.05 | 0.01 | 优秀 |
|  | 11.40 | 0.02 | 良好 |
|  | 10.35 | 0.83 | 中等 |
|  | 6.95 | 0.02 | 差 |

表3.12 消费水平聚类结果

|  |  |  |  |
| --- | --- | --- | --- |
| 类别符号 | 综合均分 | 人数占比 | 行为标签 |
|  | 19.15 | 0.009 | 极高 |
|  | 10.65 | 0.006 | 高 |
|  | 10.35 | 0.006 | 较高 |
|  | 9.6 | 0.002 | 一般 |
|  | 7.35 | 0.006 | 普通 |
|  | 6.65 | 0.887 | 低 |
|  | 5.85 | 0.006 | 极低 |

根据以上聚类结果表，我们便可以构建描述学生行为的离散化标签集合，用于刻画学生画像，具体形式如表3.13所示

表3.13 学生各类行为数据的离散标签集合

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 学生编号 | 知识 | 艺术 | 体育 | 品德 | 劳动 | 消费 |
| 1 |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |
| 3 |  |  |  |  |  |  |
| … | … |  | … |  | .. | … |
| … | … | … |  | … | … | … |
| … |  | … |  |  |  |  |
| … | … |  | … | … |  | … |
| 1753 |  |  |  |  |  |  |

## 本章小结

本章在第二章的理论基础上，首先通过对采集的校园日志数据进行数据清洗、数据集成；其次结合轮廓系数的评价指标利用DBSCAN算法对学生的各类行为数据进行聚类分析，同时确定最佳的聚类算法参数；最后根据聚类结果，通过类别符号建立描述学生行为数据的离散化标签体系，为下一章的关联分析做铺垫。

# 基于关联规则的学生行为融合分析

## 引言

学生的艺术成绩是评价艺考学生在校表现的重要指标，本章以艺术水平为主要研究对象，根据上一章节对学生各类行为数据构建的标签化集合，结合多频繁模式树算法对学生行为数据中频繁模式进行分类，进一步确定影响学生艺术综合得分的因素。最后，经过实验研究，验证了所提方法的有效性。

## 基于FP-Tree的行为融合分析

模型融合是指基于机器学习的集成算法的表达形式，一般通过构建并融合多个模型来完成学习任务。因此，模型融合主要包含两个阶段：构建若干个单模型和模型融合。为分析影响学生艺术成绩的因素，本文应用多频繁模式树（Frequent Pattern-Tree，FP-Tree）分类算法[28]的融合模型对学生的德、智、体、美、劳和消费六个方面进行关联分析，从而找出对学生艺术成绩影响最大的影响因素。FP-Tree的主要思想是将代表频繁项集的数据压缩在频繁模式树中，每一项之间的关系被保留在频繁模式树中，然后将频繁模式树按照条件模式基拆分成一组条件FP子树，并分别挖掘这些子树所包含的频繁项集。FP-Tree作为基于树形结构的模式增长类算法，采用分而治之的策略，对基于主树建立的子树进行快速挖掘，相比于基于候选集的关联规则挖掘算法，例如Apriori、IAA算法[29,30]等，搜素效率大幅提升。

FP-Tree由建立项头表、建立FP树以及挖掘频繁项集三部分组成。FP树的建立依赖项头表的建立，在第一次扫描候选数据集时，得到所有频繁一项集的计数，然后删除支持度低于阈值的项，并按支持度降序排列。第二次扫描数据，剔除原始记录中非频繁1项集，并按照支持度降序排列。有了项头表和排序后的数据集，就可以开始建立FP树，我们需要将一条条排序后的数据集插入到FP树中。排序考前的是祖先节点，靠后的是子孙节点，若有共同祖先，则对应的公共祖先节点计数加1。如果有新节点，则项头表对应的节点会通过节点链表接上新节点，直到所有数据插入到FP树中，FP树建立完成。基于项头表，FP树以及节点链表，我们就可以挖掘到项头表每一项的条件模式基。条件模式基是以我们要挖掘的节点作为叶子节点所对应的FP子树，我们将子树中每个节点的计数设置为叶子节点的计数，并删除计数低于支持度的节点，从这个条件基中，我们就可以递归挖掘得到频繁项集，其中同时满足最小支持度和最小置信度的规则称为强关联规则。

由上一章的学生行为画像分析，我们已经建立了描述学生行为特征的标签体系，即艺术水平优秀的学生，具有何种水平的思想品德、身体素质、知识掌握能力、勤劳程度和消费习惯，而学生的各类行为数据我们通过设置最小的阈值，并转化为离散化的标签集合建立行为画像，如表3.7所示。本章在此基础上，应用FP-Tree分类算法对学生的行为画像进行情景化描述，识别出不同层次学生的行为特征，及时引导和纠正教育方案。针对上述建立的学生标签体系，我们采用多频繁模式树算法挖掘其中的强关联规则。首先将最小支持度minsup和最小置信度mincon都设置为0.02，接着采用FP-Tree算法识别不同层次学生的行为特征，具体的算法步骤如下：

Step1：扫描每个学生ID对应的六个离散化标签，即六个类别符号，得到所有频繁1项的计数，删除低于支持度的频繁1项，并按照降序排序插入到项头表中。

Step2：扫描每个学生ID下的原始记录，删除包含的非频繁1项，并对每条记录按照支持度降序排序，建立学生行为标签体系项头表。

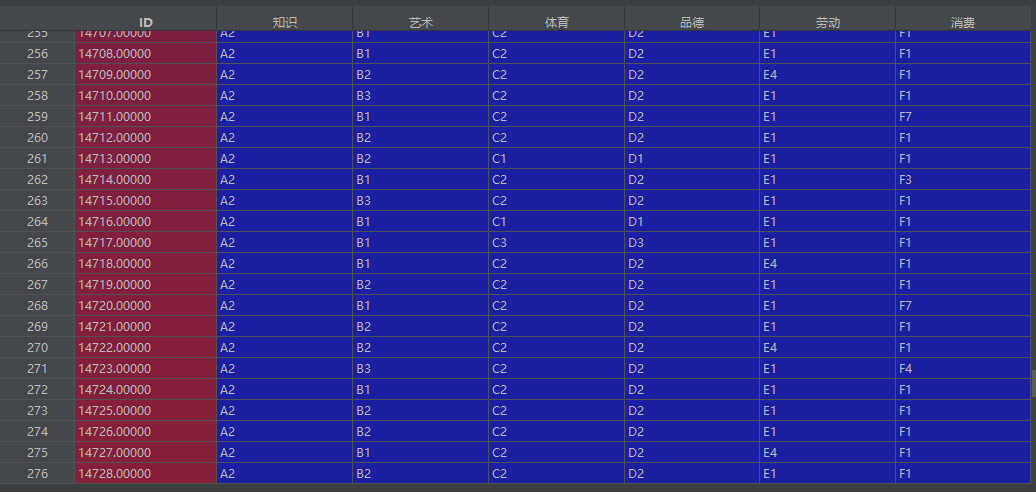
Step3：从项头表中读入每条记录，并按照如下规则建立FP树：将每条记录中排序在前的频繁1项作为祖先节点，排序在后的作为子孙节点，若有共同祖先，则对应公共祖先节点的支持度计数加1，若插入的是新节点，则通过节点链表将频繁1项的项头表链接到新的项节点，直到所有的频繁1项中插入到多频繁模式树中，建立学生行为标签体系FP树。

Step4：接着从项头表中自底向上地挖掘项头表中每一项对应的FP子树，找到子树中对应的条件模式基，并从条件模式基中不断挖掘多频繁模式项。

Step5：根据预先定义的支持度和置信度中发现满足要求的多频繁模式，算法结束。

## 实验结果与分析

本章的实验依旧基于校园日志数据，根据上一章得到的学生数据离散化标签集合，我们在本章节通过采用多频繁模式树挖掘在知识水平方面表现出的优秀学生以及普通学生的行为特征。如图4.1为真实数据下的学生行为数据的标签体系。



4.1 学生数据的离散化标签集合

本文根据多频繁模式树的分类算法挖掘学生数据中的多频繁模式，并且根据固定的行为标签挖掘所包含的FP子树，以反映其余行为特征对该行为标签的影响。如图4.2-4.5为学生数据的多频繁树结构。我们可以通过每个等级艺术水平对应的FP子树发现与其相关性最高的频繁项，从而可以分析出哪些因素影响了学生的艺术水平分布。具体来说，与节点距离越近，影响力越大，从树的节点分布我们就可以清洗明朗地发现各因素与艺术水平之间的关系。

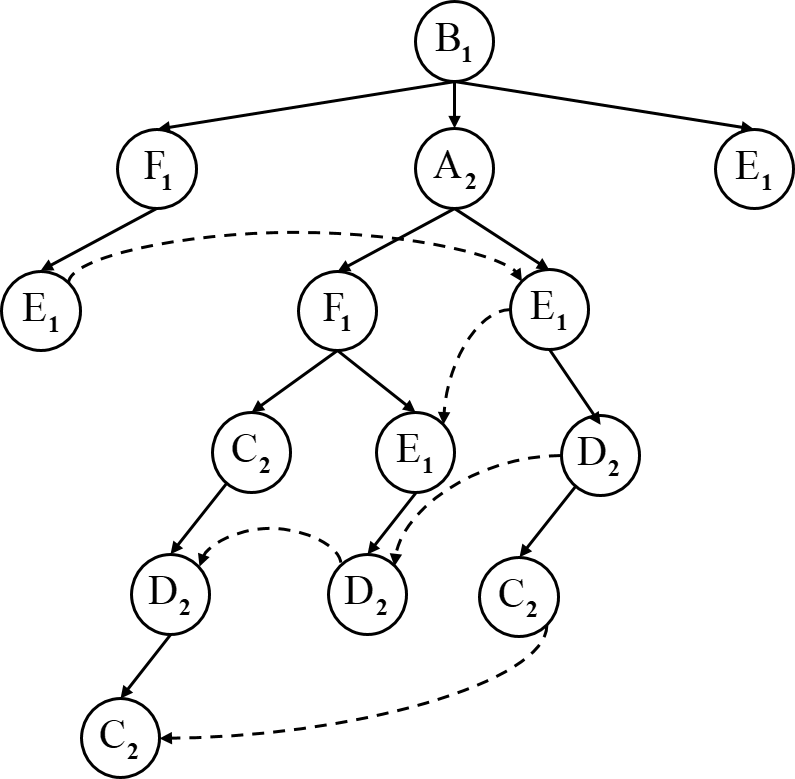


图4.2 艺术水平优秀学生的多频繁模式树

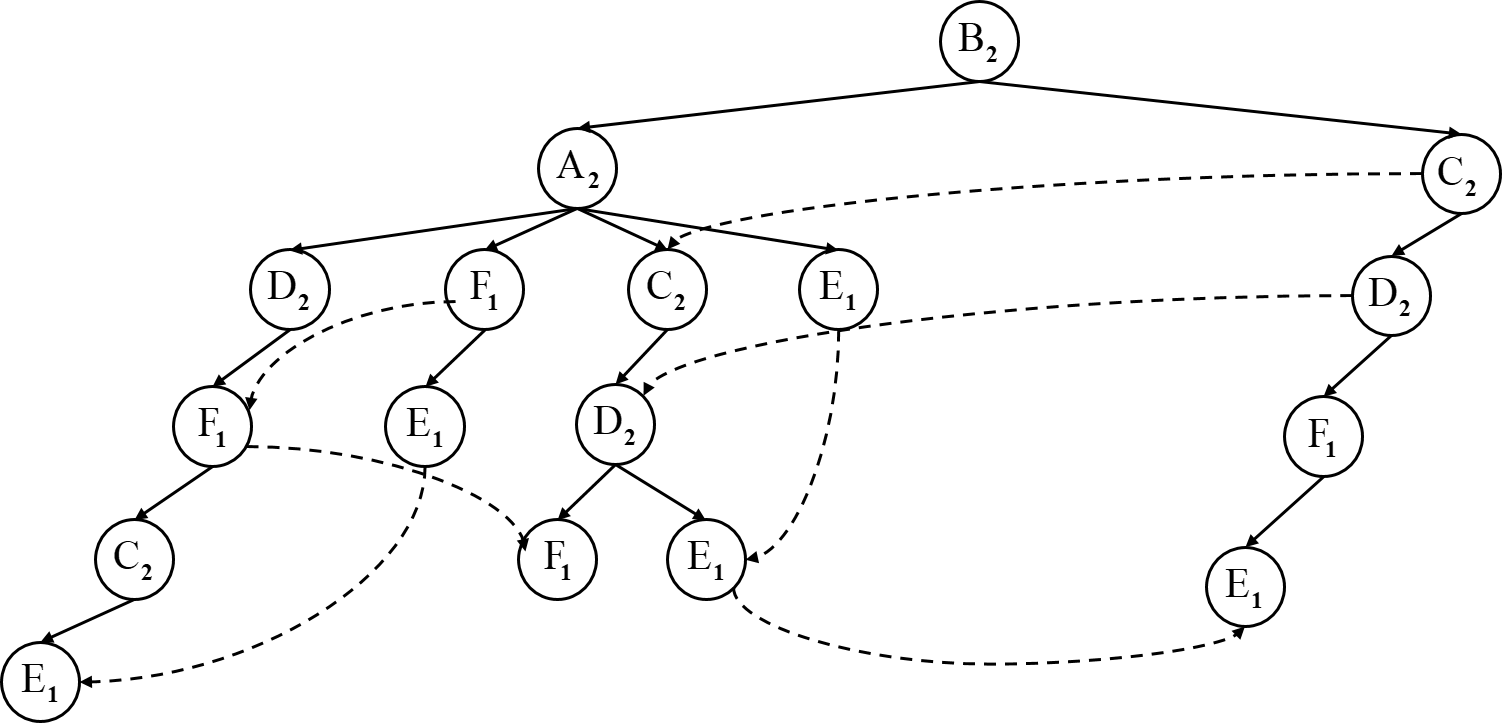


图4.3 艺术水平良好学生的多频繁模式树

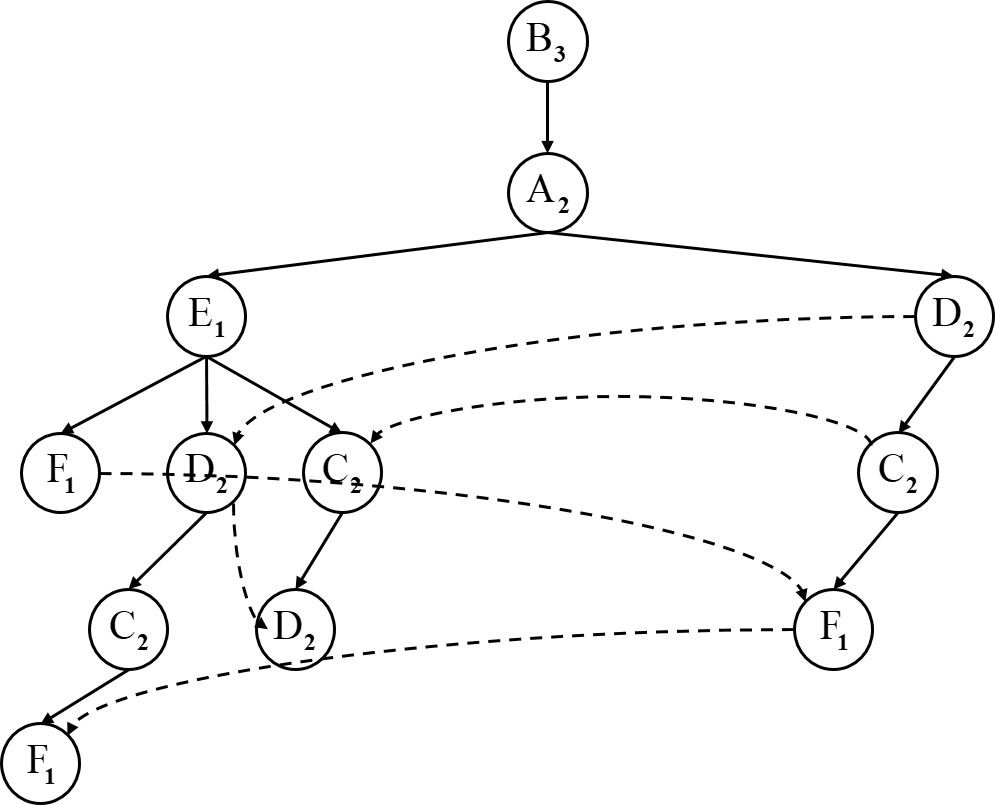


图4.4 艺术水平一般学生的多频繁模式树

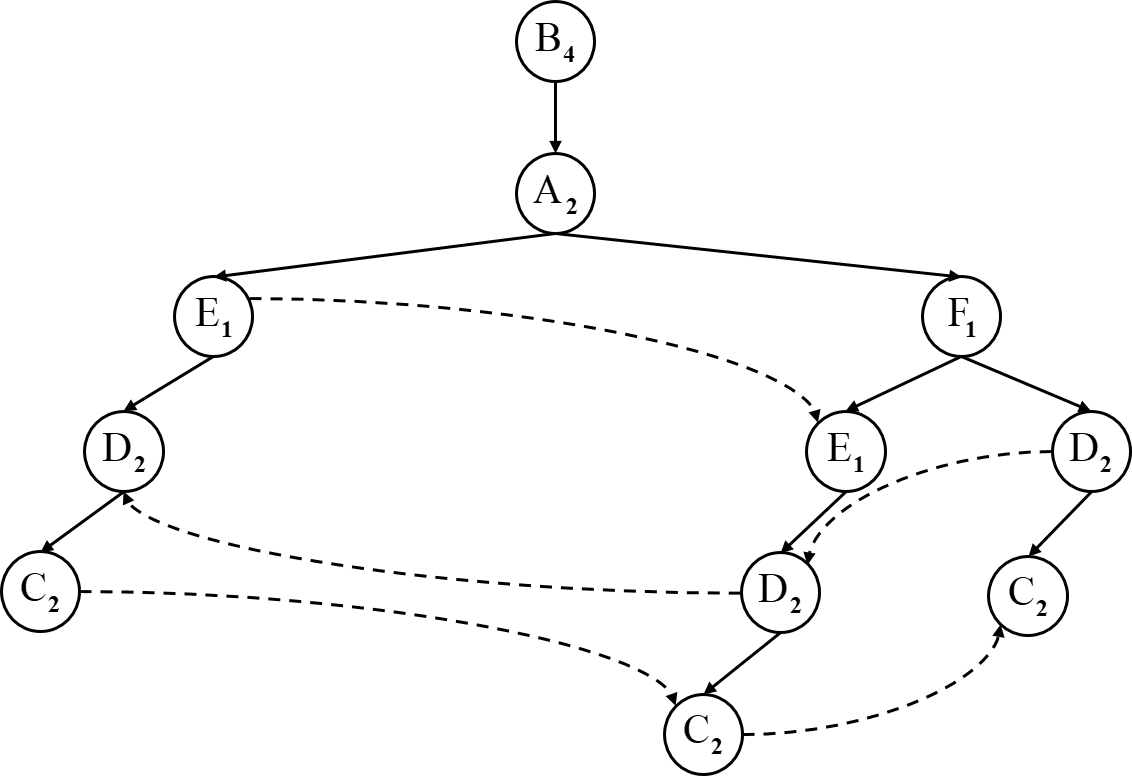


图4.6 艺术水平差学生的多频繁模式树

当确定了每个级别的艺术水平的对应关系，我们就可以发现其中潜在的强关联规则，我们根据学生数据的多频繁模式树识别出学生数据中各类行为特征的关联规则表，如表5.1-5.4所示。

由表5.1可知，艺术水平优秀的学生一般有三种行为特征，第一种是：体育水平一般，高消费，品德水平普通，知识水平良好。第二种是：高消费，知识水平良好。第三种是：知识水平良好，勤劳程度优秀。

表5.1 艺术水平优秀学生的行为特征多频繁模式表

|  |  |
| --- | --- |
| 规则 | 支持度 |
| {'C2', 'F1', 'D2', 'A2'}=>{ 'B1'} | 0.30 |
| {'F1', 'A2'}=>{ 'B1'}  {'A2', 'E1'}=>{ 'B1'} | 0.40  0.30 |

由表5.2可知，艺术水平良好的学生有三种行为特征，第一种是：高消费，知识水平良好，品德水平普通，体育水平一般。第二种是：知识水平良好，品德水平普通，体育水平一般，勤劳程度优秀。第三种是：知识水平良好，高消费。

表5.2 艺术水平良好学生的行为特征多频繁模式表

|  |  |
| --- | --- |
| 规则 | 支持度 |
| {'F1', 'A2', 'D2', 'C2'}=>{ 'B2'} | 0.30 |
| {'A2', 'D2', 'C2', 'E1'}=>{ 'B2'}  {'A2', 'F1'}=>{ 'B2'} | 0.30  0.40 |

由表5.3可知，艺术水平一般的学生有三种行为特征，第一种是：勤劳程度优秀，体育水平优秀，高消费，知识水平普通，品德优秀。第二种是：勤劳程度优秀，品德普通，体育良好，知识良好，消费极低。第三种是：勤劳程度优秀，品德普通，高消费，体育良好，勤劳程度一般。

表5.3 艺术水平一般学生的行为特征多频繁模式表

|  |  |
| --- | --- |
| 规则 | 支持度 |
| {'E1', 'C1', 'F1', 'A2', 'D1'}=>{ 'B3'} | 0.33 |
| {'E1', 'D2', 'C2', 'A2', 'F7'}=>{ 'B3'} | 0.33 |
| {'E1', 'D2', 'F1', 'C2', 'A2'}=>{ 'B3'} | 0.33 |

由表5.4可知，艺术水平差的学生有三种行为特征，第一种是：勤劳程度优秀，品德普通，知识水平普通，高消费，体育水平普通。第二种是：品德普通，体育良好，知识水平良好，勤劳程度极低，高消费。第三种是：勤劳程度优秀，品德普通，体育良好，知识水平良好，消费极低。

表5.5 艺术水平差学生的行为特征多频繁模式表

|  |  |
| --- | --- |
| 规则 | 支持度 |
| {'E1', 'D3', 'A2', 'F1', 'C3'}=>{ 'B4'} | 0.33 |
| {'D2', 'C2', 'A2', 'E4', 'F1'}=>{ 'B4'} | 0.33 |
| {'E1', 'D2', 'C2', 'A2', 'F7'}=>{ 'B4'} | 0.33 |

## 本章小结

本章通过使用多频繁模式树FP-Tree在学生行为特征画像的基础上进一步挖掘分析了影响学生艺术水平的各项因素。通过调整支持度、置信度等指标，使FP-Tree算法在挖掘特征间的关联性表现更好。进而区分出优秀艺术水平学生与普通学生之间的区别，细分影响因素，为学生之后的教学仿真提供智能化调整机制。

# 总结和展望

## 工作总结

在“智慧教育”的广泛普及和高速发展的背景下，实现学生在校行为智能管理成为未来发展不可或缺的一部分。本文在研究教育数据挖掘和用户画像技术的基础上，对采集的校园日志数据进行数据处理，挖掘背后的潜在信息，通过符号化标签体系建立了学生行为画像，并进一步对不同层次学生的行为特征进行了融合分析，为校园的管理者、教育工作者和学生个体展示了更加全面详实的学生画像。本文的主要工作和贡献如下：  
 第一，本文研究现有的数据挖掘算法的基础上，对采集到的校园日志数据进行预处理：数据清洗，利用拉格朗日插入法和箱装图方法对缺失值和异常值进行清洗；数据集成，由于采集到的数据是多源异构分布，我们按照德、智、体、美、劳和消费水平对采集到数据进行规约和整合，将学生在校各类行为所产生的不同的量化指标集成到不同的维度表内；数据融合，我们将各维度表的指标进一步细分，得到描述学生行为的二级指标，即加权平均值，月平均消费峰值，活动得分等指标，建立相应的学生画像特征库。

第二，为更好地拟合数据分布，对学生的各类行为数据进行聚类分析，本文采用基于密度的聚类算法DBSCAN，对，得到各类行为数据的聚类类别并用相应的类别符号表示，从而完成对学生个体进行分类并构建学生的行为标签集合。

第三，为进一步识别不同层次的学生行为特征，本文在构建的学生各类行为数据离散化标签集合的基础上，以学业成绩为研究对象，采用多频繁模式树算法挖掘影响学生成绩的各个因素，筛选出优秀学生与普通学生的行为特征。

## 工作展望

尽管本文通过聚类分析与关联规则挖掘算法建立了学生在校行为画像，但由于作者知识受限和经验不足，此项研究依旧存在许多改进的地方，未来可以进一步研究方向有如下三点：

第一，缺乏进一步的学生细分和生活应用研究，例如从学生的消费水平，我们可以对学生进行贫困生认定，并对贫困生的困难情况进行分级，以便针对不同等级的贫困生进行不同程度的帮扶。

第二，缺乏进一步的学生行为趋势预测与分析，例如从学生的历史成绩，我们可以采用时序预测等方法，得到学生未来成绩的走向，帮助校园工作者和学生及时发现问题并调整教育方案。

第三，缺乏进一步的多维度学生行为画像构建，在构建学生行为画像时，我们只从德、智、体、美、劳和消费六个方面对构建了学生行为画像，在后续的工作中我们应该考虑融入更多维度的学生行为数据。

参考文献

1. 张齐. 中国受高等教育人口统计分析[D].山东大学,2021.DOI:10.27272.
2. 王媛媛.英语专业高职学生信息素养现状调查及启示——基于《教育信息化2.0行动计划》的启示[J].广东开放大学学报,2020,29(04):98-102.
3. 胡慧睿.基于教育信息化2.0行动计划,科学教育与信息技术融合要点分析[J].现代职业教育,2018(24):140-141.
4. F. de Paula Santos, C. P. Lechugo and I. F. Silveira-Mackenzie, "“Speak well” or “complain” about your teacher: A contribution of education data mining in the evaluation of teaching practices," 2016 International Symposium on Computers in Education (SIIE), 2016, pp. 1-4.
5. H. Jin, T. Wu, Z. Liu and J. Yan, "Application of Visual Data Mining in Higher-Education Evaluation System," 2009 First International Workshop on Education Technology and Computer Science, 2009, pp. 101-104.
6. M. H. Saraee and B. Theodoulidis, "Knowledge discovery in temporal databases," IEE Colloquium on Knowledge Discovery in Databases (Digest No. 1995/021 (A)), 1995, pp. 1/1-1/4.
7. 唐小燕,楼桦.智慧教育下的教学数据可视化应用探索[J].江苏经贸职业技术学院学报,2022(03):56-59.DOI:10.16335.
8. 吕坛,刘锦秀,徐向明.国内大学生创新创业教育研究的热点主题和演进路径——基于CiteSpace的数据挖掘及可视化分析（2000-2021年）[J].创新创业理论研究与实践,2022,5(02):12-17+24.
9. 梁珊,王选飞.疫情防控常态化背景下混合式教学模式探索——以“数据挖掘”课程为例[J].黑龙江教育(理论与实践),2021(12):70-71.
10. Cristóbal Romero, Sebastián Ventura, and Enrique García. 2008. Data mining in course management systems: Moodle case study and tutorial. Comput. Educ. 51, 1 (August, 2008), 368–384.
11. Wenke Lee and Salvatore J. Stolfo. 1998. Data mining approaches for intrusion detection. In Proceedings of the 7th conference on USENIX Security Symposium - Volume 7 (SSYM'98). USENIX Association, USA, 6.
12. L. Xu, M. Chow and L. S. Taylor, "Power Distribution Fault Cause Identification With Imbalanced Data Using the Data Mining-Based Fuzzy Classification $E$-Algorithm," in *IEEE Transactions on Power Systems*, vol. 22, no. 1, pp. 164-171.
13. F. Shi, Q. Miao and D. Mei, "The application of data association mining technology in university curriculum management," *2012 IEEE Symposium on Robotics and Applications (ISRA)*, 2012, pp. 521-524.
14. 李宇帆,张会福,刘上力,唐兵.教育数据挖掘研究进展[J].计算机工程与应用,2019,55(14):15-23.
15. 钟瑜,辜向东,肖巍.我国外语教育数据挖掘研究综述[J].外语与翻译,2018,25(01):79-86+98.DOI:10.19502.
16. 第三期“中国基础教育大数据发展蓝皮书”项目核心成果[J].现代远程教育研究,2021,33(02):113.
17. 柴艳妹，雷陈芳.基于数据挖掘技术的在线学习行为研究综述[J].计算机应用研究，2018（5）：1287-1293.
18. 于方，刘延申“以用户为中心”的教育数据挖掘应用研究[J].电化教育研究，2018，39（11）：69-77.
19. X. Huang, Z. Qiu, J. Su, Z. Shi and L. Yan, "Knowledge Graph-based User Portrait Construction for Electricity Enterprise Suppliers," 2021 International Conference on Communications, Information System and Computer Engineering (CISCE), 2021, pp. 81-84.
20. A. Amato, V. Di Lecce and V. Piuri, "A Smart Distributed Measurement Data Management System for DSM," 2007 IEEE Instrumentation & Measurement Technology Conference IMTC 2007, 2007, pp. 1-5.
21. R. Hashemi et al., "Effects of Phenotypical Patterns on Epigenetic Markers," 2018 International Conference on Computational Science and Computational Intelligence (CSCI), 2018, pp. 1351-1356, doi: 10.1109/CSCI46756.2018.00262.
22. A. Mason, O. Korostynska, T. T. Håseth, O. Alvseike and M. S. Sødring, "Toward detecting boar taint compounds using low. power microwave sensors," 2019 13th International Conference on Sensing Technology (ICST), 2019, pp. 1-6.
23. Cynthia Waters, "Materials Technology Gaps in Metal Additive Manufacturing," in Materials Technology Gaps in Metal Additive Manufacturing , SAE, 2018, pp.i-xviii.
24. 王宪朋. 基于视频大数据的用户画像构建[J]. 电视技术,2017,41(6):20-23. DOI:10.16280/j.videoe.2017.06.005.
25. 马安华. 基于用户行为分析的精确营销系统设计与实现[D]. 江苏:南京邮电大学,2013.
26. R. M. Esteves, T. Hacker and C. Rong, "Competitive K-Means, a New Accurate and Distributed K-Means Algorithm for Large Datasets," 2013 IEEE 5th International Conference on Cloud Computing Technology and Science, 2013, pp. 17-24.
27. Mustakim, E. Rahmi, M. R. Mundzir, S. T. Rizaldi, Okfalisa and I. Maita, "Comparison of DBSCAN and PCA-DBSCAN Algorithm for Grouping Earthquake Area," *2021 International Congress of Advanced Technology and Engineering (ICOTEN)*, 2021, pp. 1-5.
28. Z. Pan, P. Liu and J. Yi, "An Improved FP-Tree Algorithm for Mining Maximal Frequent Patterns," 2018 10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), 2018, pp. 309-312.
29. P. R P. and K. Prasad Rao, "A Comparative Study On Apriori And Reverse Apriori In Generation Of Frequent Item Set," 2019 1st International Conference on Advances in Information Technology (ICAIT), 2019, pp.
30. H. Wu, Z. Lu, L. Pan, R. Xu and W. Jiang, "An Improved Apriori-based Algorithm for Association Rules Mining," 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 2009, pp.