中 文 摘 要

随着教育信息化的日益普及，教育数据挖掘的相关研究越来越丰富，在校学生的相关数据也在不断积累。对这些数据进行挖掘和分析，可以帮助学校更好地了解每个学生的各方面，为学生提供更好的学习体验和校园服务。因此，本文通过构建学生画像以及成绩预测模型，来提高学生信息化和个性化服务的水平，本文的主要研究内容如下：

（1）基于改进模糊聚类的学生行为分析。本文使用多种方法改进模糊 C 均值聚类算法，首先利用高斯密度函数确定初始聚类中心，然后使用密度敏感距离替换欧式距离，通过轮廓系数和肘部法相结合的办法来确定最佳聚类数，最后对改进聚类算法的有效性进行验证。利用改进后的聚类算法对学生三个维度进行分析，包含消费行为聚类，学习行为聚类，生活行为聚类，归纳概括不同类别学生的个性化标签结果。

（2）基于多注意力机制的成绩预测方法。利用神经网络模型，结合学生近一年的行为特征和历史成绩，先用一组平行的注意力机制计算最近两个学期成绩上各行为特征的权重；再用另一个注意力机制提取两个学期之间成绩与行为的关系，最后基于融合后的特征对下一学期成绩进行更精准的预测。实验结果表明，本模型提高了成绩预测的准确率。

（3）学生综合画像系统的设计与实现。利用聚类结果和成绩预测模型构建学生画像的标签，设计并实现学生行为管理、学生综合画像展示、成绩预测等功能，最终采用 Echarts 来实现学生画像大屏可视化展示。

关键词：模糊聚类；学生行为；成绩预测；注意力机制；学生画像

ABSTRACT

With the popularization of information technology in education, more and more attention has been paid to Educational data mining education. Data on students at school is also accumulating. Mining and analyzing these data can help schools better understand every aspect of every student. To provide students with better learning experience and campus services. Therefore, this article constructs the student profile and the performance prediction model, enhances the student informationization and the personalized service level. The main research contents of this paper are as follows:

(1) Student behavior analysis based on improved fuzzy clustering. In this paper, several methods are used to improve the fuzzy c-means clustering algorithm. First, the initial clustering center is determined by Gauss density function. Then replace the Euclidean distance with the density-sensitive distance. The contour coefficient and elbow method are used to determine the optimal cluster number. Finally, the effectiveness of the improved clustering algorithm is verified. Using the improved clustering algorithm to analyze the three dimensions of students, including consumer behavior clustering, learning behavior clustering, life behavior clustering. The behavioral characteristics of different types of students were summarized.

(2) Performance prediction method based on multi-attention mechanism. The neural network model is used to combine the students' behavior characteristics and historical performance in the past year. A parallel set of attention mechanisms was used to calculate the weights of behavioral traits on the last two semesters' grades. Another attention mechanism was used to extract the relationship between performance and behavior over two semesters. Finally, based on the characteristics of the integration of the next semester grades for more accurate prediction. The experimental results show that the model improves the accuracy of performance prediction.

(3) The design and implementation of the student integrated portrait system. Using clustering results and grade prediction model to construct the label of student portrait. Design and realize the functions of student behavior management, student comprehensive profile and grade prediction. Finally, Echarts is used to realize the

large-screen visualization of student profile.

Key words： Fuzzy clustering; Student behavior; Performance prediction; Attention mechanism; Student profile

目 录

第一章 绪论

1.1 研究背景和意义

随着教育信息化的不断进步，智慧校园建设以及教育大数据的迅速增长，教育数据挖掘（Educational Data Mining, EDM）应运而生。它是一种应用于教育领域的数据挖掘技术，通过从大量的教育数据中发现隐藏的模式、关联和趋势，来提取有关学习者、教学和评估等方面有价值的信息[1]。EDM 可以追溯到 20 世纪 90 年代，当时计算机技术和互联网的快速发展为教育领域提供了大量的数字化学习和教学数据，从而催生了对这些数据进行挖掘和分析的需求。

教育数据挖掘和教育信息化是紧密相关的概念，两者之间存在密切的关系。使用 EDM 处理实际教育问题是教育信息化的新要求，对教育数据进行深入分析和挖掘，这些教育数据蕴含着丰富的信息和潜在的价值，可以用于深入了解学生的学习行为、评估教学质量、优化学科资源配置等，从而对教育决策和教学改进提供科学依据。在此背景下，各地高校已相继着手开展数字化校园建设工作，高校信息化建设也在不断推进。

随着大数据的兴起，“数据驱动”这个概念备受关注。它紧密结合于数字化转型，即通过新技术如大数据、人工智能、云计算和移动互联网等推动业务增长。在教育研究中，数据驱动的方法被广泛应用，以实现更多的应用。然而，教育产生的数据数量庞大，价值较低，因此需要使用数据挖掘技术从中提取所需信息并进行高效利用。随着机器学习和人工智能等研究领域的不断发展，传统的分析手段和数据处理方法已经无法满足现有需求[3]，对于教育信息化而言也不例外。

高校学生在学习和生活中产生了大量的数据，但由于数据规模庞大，这些数据无法充分发挥其应有的价值，构建学生画像可以使这些数据充满意义[4]，通过“数据驱动”对学生进行画像构建实现贫困生精确资助，提高学生学习生活质量，保障学生安全等功能。大部分高校拥有自己的学生画像系统，但是都是使用简单的统计并设定一个固定值来生成标签，并没有对学生标签进行细分。本文希望通过动态生成学生标签的方法，反映最精准的学生情况。

学生在学校的学习成绩是父母、老师和学校最关注的内容，学校有必要监控学生的学业表现，并采取相应的改善措施[5]。及时预测学生的表现有很多好处，包括提前预警一些有辍学风险的学生，发现影响学生成绩的行为属性。然而，选择合适的方法来准确估计学生的表现仍然是一项复杂的工作，学生的学业成绩通常

受到许多因素的影响，包括学术和非学术属性[6]，这些因素的可变性要求建立一个复杂的预测模型。

因此学生综合画像和成绩预测都是教育数据挖掘的重要应用，通过构建学生综合画像把高校学生产生的数据进行整理分析，是智慧校园的必然要求，预测成绩是教育数据挖掘中最重要的一个环节，而使用传统的方法无法准确预测高校学生未来的成绩，需要在方法上创新的同时实际应用在高校系统中。

1.2 国内外研究现状

1.2.1 教育数据挖掘现状

教育数据挖掘研究在国外起步较早并且现在研究较为成熟，早在 1995 年便已有相关论文发表，至今已有一系列综述发表。Baker 等[7]对早期 EDM 的工作进行了概述，并认为 EDM 在未来的教育领域将有更大的影响力，但由于研究方法比较简单，并受当时的技术水平的限制，研究成果很少。

近年来，随着各类数据源的增加，教育数据挖掘方法和技术也得到了快速发展，趋向于更加多样化和复杂化。例如，Zaffar[8]使用教育数据挖掘方法来识别学生的情况和远程学习的参与模式。Toivonen[9]提出了一种新颖的增强智能方法和神经 N-树，将教育数据挖掘过程演变为具有参数和预测模型调整和可视化的开放过程。Robert[10]利用动态时间扭曲核在学习者行为的时间序列之间创建两两相似性，并将其与无监督的多尺度图聚类算法相结合，以识别具有相似时间行为的学习者组。Salloum[11]回顾了经典的数据挖掘技术、关联、分类、回归、聚类和预测，发现最新的研究特别是在在线学习环境中进行评估时，大多会使用深度学习技术，但是对于传统的教育领域比如高校，使用经典算法或者多种经典算法结合的研究比较多。随后 Xu[12]等确定了四个主要的 EDM 研究课题，包括成绩预测、对教师和学生的决策支持、行为检测和学生建模。

国内的 EDM 研究起步较晚，与国外相比在研究广度和深度上均有较大的差距。《中国基础教育大数据发展蓝皮书》指出，2015 年是“中国教育大数据元年”，EDM 领域的文献数量开始爆发式增长，同时也面临诸多机遇和挑战[13]。不同教育环境所使用的数据源不同，随着网络教育的兴起，EDM 的数据更多自于开放和智能的在线学习系统，但是在传统教育领域的发展依旧迅速。

由于大部分国内高校都在使用校园一卡通来方便学生的学习生活，所以国内学者对教育数据挖掘的工作立足于研究学生的校园刷卡消费行为，比如韩泽峰等[14]

提出了能够对消费数据时序性和关联性进行深度挖掘的模型，并在此模型的基础上构建消费异常检测模型，但是因为敏感信息问题，大多数实验数据并不公开，导致在传统教育领域教育数据挖掘并没有得到好的发展。姜绍萍等[15]利用学生行为相关性分析和神经网络对高校学生进行学业预警。李宇帆等[16]提出了四个应用方向，包括个性化学习服务、学生学习效果研究、学生辍学研究和学习行为分析等。

1.2.2 成绩预测模型现状

在成绩预测方面的研究，国外成绩预测技术相对国内比较成熟，研究者通常使用数据挖掘和机器学习方法来研究学生成绩预测，比如 Abuteir[17]对毕业生成绩数据进行分析研究，试图通过分类、聚类、异常检测等方法找到部分毕业生成绩低的原因；Ashenafi[18]基于学生的表现在几个任务分配整个课程期间，使用一个半自动的同行评估系统的数据，提出了预测模型，最后使用根均方差来评估它们的性能；Kaur[19]利用 WEKA 开源工具对学生成绩数据集进行了多种分类算法的测试和应用，并提出了基于分类算法的预测数据挖掘模型。

随着深度学习和人工智能应用的兴起，近年来许多国外学者都尝试使用深度学习算法来实现成绩预测，因为深度学习不仅可以代替人工提取特征，而且预测的精度也相对较高。例如 FokW[20]利用神经网络对学生学习成绩进行分类，建立高预测精度的 Tensorflow 深度学习模型，并确定了影响预测模型准确性的因素。AljohaniNR[21]通过使用可自由访问的开放大学学习分析数据集部署深度长短期记忆模型，解决了学生成绩预测的时间序列分类问题，对有风险学生的早期预测。CorriganO[22]提出了一种提高学习分析系统准确性的方法，通过对一所大学的所有学生应用循环神经网络，发现学生在与虚拟学习环境下的课堂表现。Mengash[23]根据大学入学以前的成绩来预测申请人的大学早期学习成绩，证明使用人工神经网络的方法优于其他传统的分类技术。Alshanqiti[24]提出一个混合回归模型，优化学生学习成绩的预测精度，以未来不同课程的成绩来衡量，预测与获得的学生成绩相关的各种因素产生的影响的定性值。虽然国外研究使用的技术多种多样，但是数据集的来源依然是过去问卷调查之类的手段，无法保证数据的真实性。

国内在成绩预测方面的研究晚于国外，但是也取得了不少优秀的成果，例如陈子健等[25]从教育数据中挖掘影响在线学习者学业成绩的因素，提出采用集成学习和嵌套集成学习的方法构建集成式学业成绩分类预测模型。吴蓓等[26]为了对风险学生提前预警，针对一门计算机课程，使用改进 C4.5 算法对学生进行成绩预测。

任占广等[27]分析学生在线学习行为，提取影响成绩的相关特征，利用神经网络建立在线课程成绩预测模型。

随着国内“智慧校园”需求的不断扩大，研究学生各方面数据的学者越来越多，但是在研究方法上的创新数量不多，例如李梦莹[28]提出基于双路注意力机制的学生成绩预测模型用前两学期成绩精准预测下一学期成绩。张阳等[29]提出基于图自编码器模型（Graph-AE）的学生成绩预测方案，该模型可以不经人工干预自动提取特征，且不需要大量的先验知识，能更好地刻画学生与课程之间的相关性和差异性。陈曦[30]等使用基于知识图谱表示学习的方法计算课程在知识层面的相似度，并将课程的知识相似度集成到传统的成绩预测框架协同过滤中来对成绩进行预测。上述方法不能十分恰当地预测学生的表现，对数据的挖掘深度不够，最近的研究主要集中在新方法上。

1.2.3 学生画像现状

学生画像是用户画像的一个应用方向，用户画像是真实用户的虚拟形象，是建立在一系列属性数据之上的用户模型。作为描述用户特征的工具，通过其获取用户偏好，已经取得了一定成果。但是绝大部分用在推荐系统之中，比如 Despenic[31]基于用户控制行为构建用户画像，获取用户对灯光的照明偏好。用户因素、项目特征对用户偏好存在非线性影响，Purkaystha[32]使用深度前馈网络，构建用户画像和项目特征，预测用户对项目的满意度。

由于学生身处复杂的学习环境并具有独特的学习行为，使其比一般用户更具特殊性，因此目前对学生画像的相关研究较少。Omheni[33]基于用户阅读过程中产生的注释痕迹构建学习者画像。Wu 等[34]基于模糊树构建学习者模型与学习活动模型，向学习者推荐学习活动。

国内早期的用户画像除了应用在推荐系统领域，更多的是描述图书馆用户的喜好，后来因为教育数据挖掘技术而迅速发展，肖君等[35]从知识特征、行为特征和态度特征 3 个维度设计在线学习者画像模型，王晓东等[36]利用学生特征构建学生模型。现在大部分高校都拥有“智慧校园”服务，并且拥有学生画像的功能，但是这些学生画像往往是通过对数据的简单统计来实现的，这样的学生画像往往不够准确和具体。

针对上述问题，杨长春等[37]针对构建智慧校园学生画像的数据缺失与高维特征问题，引入外部数据弥补缺失的数据，辅助用户建模，提出一种基于随机森林的双向特征选择算法解决高维特征问题。凌玉龙等[38]从学生消费画像和精准资助

两个角度对校园消费数据进行挖掘研究，从数据集本身的特点和 K-means 算法的缺陷两个角度出发，更好的构建了学生画像。但是以上方法对画像的构建过于单一，而画像是有层次的标签，能够用来解决实际问题。

1.3 主要研究内容

本研究旨在探究太原科技大学本科生的多项行为数据和成绩数据之间的关联性，运用聚类算法及学生画像技术分析学生的行为特征，并用神经网络算法依据学生的行为数据对学生的成绩进行预测。研究内容的整体框架，如图 1.1 所示。本文主要研究内容如下：

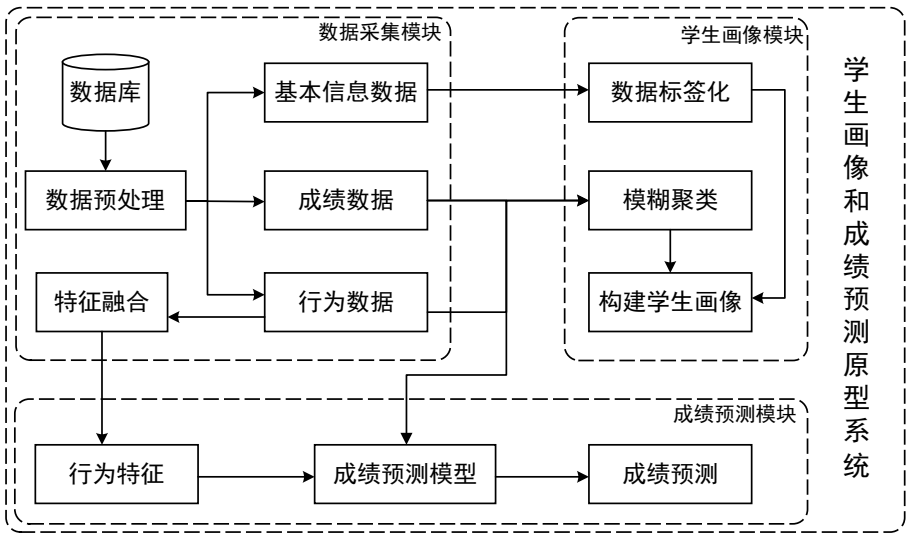

图 1.1 整体框架图

Figure 1.1 Overall frame diagram

（1）基于改进模糊聚类的学生行为分析

使用多种方法改进模糊 C 均值聚类算法。改进方法为利用高斯密度函数确定初始聚类中心，然后使用密度敏感距离替换欧式距离，对于最佳聚类数的确定，使用多种方法结合，最后对改进聚类算法的有效性进行验证。使用改进后的聚类算法对学生三个维度进行分析，包含消费行为聚类，学习行为聚类，生活行为聚类。

（2）基于多注意力机制的成绩预测方法

利用神经网络模型，结合学生近一年的行为特征和历史成绩，先用一组平行的注意力机制计算最近两个学期成绩上各行为特征的权重，再用另一个注意力机制提取两个学期之间的关系，最后基于融合后的特征对下一学期成绩进行更精准

的预测。

（3）学生综合画像系统的构建

建立画像标签是学生画像的关键工作，如图 1.2 所示是学生画像需求分析及分级标签体系。本文将通过聚类和深度网络模型来确定标签的选择，开发学生行为管理系统与画像大屏展示页面来实现智慧校园的各个需求。

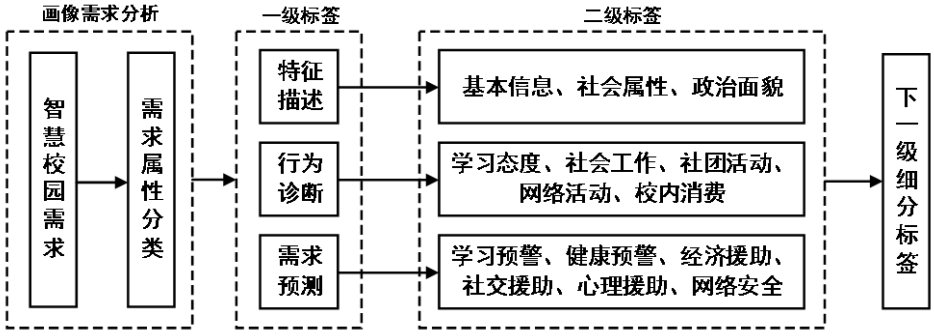

图 1.2 学生画像需求分析及分级标签体系

Figure 1.2 Student portrait needs analysis and grading label system

1.4 组织结构

根据研究内容和路线，本文共分为六章，各章节具体内容为：

第一章 绪论。主要介绍本研究的背景和意义，并对教育数据挖掘、成绩预测和学生画像的国内外研究现状进行介绍。随后对本文的研究思路和方法进行了简单介绍，最后对本文各章节的组织结构进行安排。

第二章 相关技术。主要介绍了本文中所用到的相关技术和方法，包括数据挖掘技术、模糊聚类算法、成绩预测技术以及学生画像技术的概念和流程。

第三章 基于改进模糊聚类的学生行为分析。首先分析了模糊 C 均值聚类算法不足的方面，由此确定三个方面的改进方法，并在公共数据集上说明优化算法的有效性，最后使用改进的聚类算法从三个维度对学生进行聚类分析。

第四章 基于多注意力机制的学生成绩预测模型构建。通过对学生行为数据进行特征提取和分析，并采用多种注意力机制对不同行为特征的重要性进行建模，从而有效地提高了成绩预测的准确性和可靠性，最后使用真实数据集对模型进行验证。

第五章 学生综合画像系统设计与实现。通过第三章节的聚类结果以及第四章节成绩预测结果确定合适的画像指标体系，通过可视化技术实现学生画像的展示，同时开发原型系统。