**基于因果推断和多头自注意力机制的学生成绩预测**

张文奇 ，王海瑞 ，朱贵富

（1. 昆明理工大学 信息工程与自动化学院，云南 昆明 650000 ；

2. 昆明理工大学 信息化建设管理中心，云南昆明 650000）

**Abstract**

摘 要：学生成绩预测旨在为学校管理者提供决策支持，帮助教师改进教学手段，对学生进行学业指导，最终提高学生成绩。传统的学生成绩预测方法大多利用相关性分析选取重要因素，忽视多变量之间的间接联系和联系的方向性，而且在进行预测时没有区分输入特征的重要程度，因此提出一种基于因果推断和多头自注意力机制的学生成绩预测方法。该方法不仅使用因果推断选取与标签具有方向性联系的直接特征和间接特征，而且用多头自注意力机制区分不同特征对期末成绩的影响程度。在公开数据集上进行大量的实验，结果显示所提出方法的预测准确率高达 ，高于其他传统成绩预测方法。 关键词：学生成绩预测；教育数据挖掘；因果推断；多头自注意力机制；Transformer；相关性分析 中图分类号：TN911.1-34 文献标识码：A 文章编号：1004-373X(2023)17-0111-06

Students' performance prediction based on causal inference and multi-head self-attention mechanism

ZHANG Wenqi , WANG Hairui , ZHU Guifu

(1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650000, China; 2. Information Construction Management Center of Kunming University of Science and Technology, Kunming 650000, China)

**Abstract**

The purpose of students' performance prediction is to provide decision support for school administrators, help teachers improve teaching methods, provide academic guidance to students, so as to ultimately improve the students' performance. In the traditional students' performance prediction methods, correlation analysis is mostly used to select important factors, and the indirect connection and directionality of the connection between multiple variables are ignored. Moreover, the degree of importance of input features is not distinguished when making predictions. Therefore, a students' performance prediction method based on causal inference and multi-head self-attention mechanism is proposed. In this method, causal inference is used to select the direct features and indirect features that have directional connection with labels, the multi-head self-attention mechanism is used to distinguish the influence degree of different features on the final results. A large number of experiments were carried out on an open data set. The experimental results show that the prediction accuracy rate of the proposed method can reach , which is higher than the other traditional performance prediction methods.

Keywords: students' performance prediction; educational data mining; causal inference; multi - head self - attention mechanism; transformer; correlation analysis

**0 引 言**

随着时代的进步、社会的发展，教育的受重视程度越来越高。互联网技术在教育领域的普及，获取了大量的教育数据，通过对教育数据进行分析和挖掘，帮助学生了解自己的学习效果和学习进展 。数据采集技术迅速发展，出现了大量冗余变量，削弱了相关变量的作用，

因果推断是统计学和数据科学的核心问题之一，其广泛应用在医学、社会科学和经济学等多个领域。本文将因果推断应用于人工智能领域，分析出属性特征，以及前两阶段成绩和期末成绩数据之间的因果关系。针对多变量数据，经典的预测方法常使用相关性分析找出相关变量，然而相关性分析忽略了变量之间的间接关系，难以满足模型预测的要求 。因果性更强调的是具有单向性的原因和结果的联系 。

近些年来，通过学生成绩预测帮助管理者完善教学  
方法以及为学生进行个性化指导已经引起了国内外学者的广泛研究。文献[4]将累计绩点、必修课程成绩、期中考试成绩作为输人特征，对学生的期末考试成绩进行预测，模型取得了较好的预测性能。文献[5]在 18 个属性特征中根据信息增益率对属性进行排序，筛选出 8 个相关属性作为决策树预测模型的输入进行预测。文献[6]通过对中外学生数据挖掘发现学生成绩与某些属性特征有极大的相关性。

虽然上述工作在成绩预测上都取得了较好的表现，但是仍存在如下两方面的挑战：  
1）上述相关工作选取强相关性的特征作为模型输入，用相关性作为特征提取的指标，只考虑了两个变量的联系，却忽略多变量之间的间接联系和联系的方向性。  
2）上述相关工作中学生数据内部之间的关联性没有体现，忽略了各种因素对成绩的重要程度是不同的问题。

为解决上述相关工作的挑战，本文提出了一种新颖的因果推断结合多头自注意力机制的预测方法。通过因果推断选取与标签具有方向性联系的直接特征和间接特征，解决了相关性分析忽略多变量之间方向性和间接关系的问题。多头自注意力汇聚数据内部的信息加强了数据之间的交互，解决了数据内部之间没有关联性的问题。残差连接和标准化层使得数据更加标准，有效提高了模型的学习能力。

**1 相关工作**

本节主要介绍因果推断和学生成绩预测两个方面的相关工作。

**1.1 因果推断**

机器学习方法在数据分析和数据挖掘领域得到了广泛的应用，但是机器学习更多的是依赖统计方法，而忽略了因果关系。文献[7]的研究明确在因果推断时随机化的好处，首次提出了完整的潜在结果模型框架。潜在结果模型框架通过比较研究对象在不同的影响下所得结果的不同，得到干预变量和结果之间的因果关联 。文献[9]的研究为因果推断开发了一个有原则的、非参数的框架，即结构化因果模型。文献[10]针对因和果分布的不对称性提出了一种推断多维变量间线性因果关系的方法，该方法有较强的适用性，即适用于随机因果关系，也适用于确定的因果关系。为了克服对高维数据进行因果推断时的模型失效问题，文献[11]采用高维数据网络拆分策略，提出了一种基于互信息的非线性因果推断算法模型，该算法在保证准确度的同时提高了执行效率。

**1.2 学生成绩预测**

利用学生的过往成绩预测最终成绩一直是学生成绩预测的重要途径，例如文献[12]基于学期开始前提供的数据，在学期开始前准确预测表现不佳的学生，同时设计了一些可解释特征来量化引进糟糕表现的因素。文献[13]根据学生以前完成的课程学习情况来研究新学期哪些学生在课程的最初学习阶段有困难，将研究任务转换成一个多实例、多标签问题，充分利用了课程之间的关联关系。文献[14]基于学生日志数据，使用循环神经网络来预测学生最终成绩。文献[15]针对不同课程的重要性不同以及课程不是独立的问题，提出了一种基于图的分层注意神经网络模型。

考虑到学生行为习惯与他们的学习情况有着密切的联系，基于学生行为特征的成绩预测研究也一直受到国内外研究人员的广泛关注。文献[16]收集了一个真实的上网行为数据集，构建了一个基于多头自注意力机制的双层网络结构，同时引入多任务学习策略。文献[17]以高校的图书借阅历史和累计成绩作为数据样本，对所提出的有监督的感知矩阵分解方法进行了评价。研究证明学生的属性特征，例如家庭情况、学习环境等，同样对学生成绩有很大的影响。文献[18]提出一种将高考考生的短期特征与长期特征相结合的神经网络模型，短期特征包括前三次模拟考试的各科成绩，长期特征包括考生类别、考生学校，在真实数据集上进行大量实验，证明该模型预测精度优于其他经典算法。文献[19]根据 30 个属性特征在两阶段历史成绩的注意力得分，为每个属性特征分配相应的权重，区分了这些属性特征对成绩的影响程度。

**2 模型结构**

本文旨在通过分析挖掘学生的成绩数据和属性数据，为学校管理者提供决策支持，帮助教师改进教学手段，对学生进行学业指导，提出了一种基于因果推断和多头自注意力机制的学生成绩预测方法。通过因果推断描述成绩和各种因素的因果关系，选出影响学生成绩的重要因素，利用多头自注意力机制区分输入特征对期末成绩的影响程度。

预测方法框架如图1所示。

**2.1 因果推断层**

本文使用因果关系从数据中学习结构模型，使用的结构学习算法是 NOTEARS算法，该算法的主要贡献是找到一种数学建模方法：如式（1）所示，设计了一种连续的针对无环的约束方案 ，替代传统的组合约  
束的方式 ，这种方式可以通过优化反复迭代后，得到无环的结果。

式中: 表示问题的规模， 表示问题节点数； 是得分函数；DAGs是有向无环图。

![](data:application/octet-stream;base64,)

图1 本文所提出方法的架构图  
式（1）中， 的表达式如式（2）所示：

式中 tr 是矩阵的迹，也就是主对角元素之和。  
把 DAG 学习问题转化成可以用拉格朗日乘数法求最优解的问题。通过对 DAG 使用特定阈值来删除因果性较弱的边，然后通过调用结构模型函数检索到可视化图中的最大子图，在该子图中挑选出指向标签G3，或者间接指向标签 G3 的特征，作为 Transformer 层的输入特征。为了保证数据信息的完整性，因果推断层的输入在数据预处理中不进行数据清洗和数据分组，只进行数据数值化。

**2.2 Transformer 层**

Transformer层由 个相同的块堆叠构成，每一个块都有两个子层：第一个子层是一个多头自注意力层；第二个子层是一个全连接前馈网络层。在两个子层后都添加了一个残差连接，然后进行层归一化操作。

**2.2.1 多头自注意力层**

每个输入都投影到矩阵上，多头自注意力机制使用多组权重矩阵 线性变化生成多组键向量矩阵 ，查询向量矩阵 和值向量，矩阵 ，其中 表示输入到 Transformer 块的嵌入数量， 表示键向量的维度， 表示值向量的维度。多头自注意力机制的每个头都采用缩放点积注意力机制，其计算公式如下：

式中 表示键向量的维度。  
多头自注意力机制的计算公式如下：

式中： 表示第 个自注意力头的映射矩阵，其中矩阵维度表示为 ， 表示多头自注意力的头数。多头自注意力机制允许模型在 个不同的子空间学习到相关的信息，最后通过信息矩阵拼接将信息汇聚起来。

**2.2.2 前馈层**

在多头自注意力层后包含一个前馈全连接层。考虑注意力机制可能对复杂过程的拟合程度不够，第 1 个线性变换层将嵌入扩展为其大小的 4 倍，第 2 个线性变换层将其投影回其原始大小，其计算方式如下：

式中： 分别表示 2 个线性变换层的权重； 分别表示 2 个线性变换层的偏置。

**2.2.3 残差连接&层规范化**

引入残差连接的作用是为了将信息无损耗地传递得更深，来增强模型的拟合能力，残差连接需要输入和输出有同样的维度，所以把每一个层的输出维度设置成 。

层规范化是所有深层网络模型都需要的标准网络层，规范化层既可以提高模型训练的稳定性，又可以加快训练的速度。

**3 实 验**

**3.1 数据集**

本文使用葡萄牙两所公立学校在2005学年、 2006学年使用学校报告和问卷收集的数学成绩数据集来展开实验。数学成绩数据集包含 30 维属性特征以及第一阶段成绩 G1、第二阶段成绩 G2、期末成绩 G3，共有 395 条数据。本文的目的是用 30 维属性特征以及第一阶段成绩 G1、第二阶段成绩 G2，经过因果推断挑选合适的特征来准确地预测期末成绩 G3 的类别。数据集的特征描述如表1所示。  
表 1 数据集的特征描述

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 特征符号 | 特征名称 | 特征描述 |
| 1 | school | 学生学校 | 二进制:"GP"或"MS" |
| 2 | sex | 学生性别 | 二进制："F"一性或"M"-男性 |
| 3 | age | 学生年纪 | 数字: 15～22 |
| 4 | address | 学生家庭类型 | 二进制："U"-城市或"R"-农村 |
|  |  |  |  |
| 29 | health | 学生当前健康状况 | 数字:1 5 |
| 30 | absences | 学生缺课次数 | 数字:0 93 |
| 31 | G1 | 学生第一阶段成绩 | 数字:0 20 |
| 32 | G2 | 学生第二阶段成绩 | 数字:0 20 |
| 33 | G3 | 学生期末成绩 | 数字: , 输出目标 |

**3.2 数据预处理**

**3.2.1 数据清洗**

对两个数据集的期末成绩 G3 进行分析，数学成绩数据集期末成绩为 0 的有 38 人，进一步分析这些学生第一阶段的成绩和第二阶段的成绩，认定这 38 人期末缺考，将这 38 条数据删除，最终数学成绩数据集的有效数据为 357 条。

**3.2.2 数据分组**

第一阶段历史成绩 G1、第二阶段历史成绩 G2 和期末成绩 G3 使用的是 20 分评分准则， 0 是最低分， 20 是满分。通过对 3 个数据集 G1、G2、G3 进行数据分析，可以得知某些分数的样本数据量过少，结合我国的"赋分制"，将学生成绩划分为优良、中等、及格、不及格 4 个等级。每个等级对应的分数范围为：优良：16 20 分；中等：13 15 分；及格：10 12 分；不及格：<10分。

**3.3 评价指标**

针对多分类问题，本文使用准确率（Accuracy）、精确率（Precision）、召回率（Recall）、 -Measure 来评价预测方法的分类性能，计算公式如下：

式中：TP 表示分类预测正确的正样本个数；FP 表示分类预测正确的负样本个数；FN 表示分类预测错误的正样本个数；TN表示分类预测错误的负样本个数。

**3.4 实验参数设置**

在本文的实验中，因果推断设定 DAG 边的阈值 0.8 。在两个数据集中分别随机挑选出 的学生用于训练，挑选 20% 的学生用来测试。本文的模型基于深度学习框架 Pytorch 进行训练和测试。在训练时，优化器为随机梯度下降优化器 SGD，设置批处理大小 batchsize ，初始学习率 ，动量 momentum= 0.9 ，模型迭代次数 epoch=150。  
3.5 实验结果与分析  
3.5.1 因果推断结果

为了保证数据信息的完整性，因果推断层的输入在数据预处理中不进行数据清洗和成绩分组，只进行数据数值化，在数学成绩数据集中检索出的与期末成绩 G3 具有关联的特征如图2所示。

![](data:application/octet-stream;base64,)

图2 数学成绩数据集中与 G3 有关联的因果图  
从图2中可以看出，数学成绩数据集中对期末成绩 G3有直接因果关系的特征包括：第一阶段成绩 G1，有间接因果关系的特征包括第二阶段成绩 G2、接受高等 （higher）教育的意愿、学生的性别（sex）、失败（failures）次数、额外的教学支持（schoolup）。将因果推断层选取的特征作为后面模型的输入，来达到精准预测的目的。

**3.5.2 预测结果对比实验**

为了验证本文模型在成绩预测方面的表现，将本文提出的成绩预测模型（Ours）与常规方法：K近邻（KNearest Neighbor, KNN）、逻辑回归（Logistic Regression， LR）、多项式朴素贝叶斯（Multinomial Naive Bayes, MNB），以及集成学习方法：随机森林（Random Forest， RF）、极限梯度提升（Extreme Gradient Boosting, XGB）模型在数学成绩数据集上进行对比实验。为了对比实验的公正性，以因果推断的结果作为对比实验的输入特征，不同方法之间的预测结果如表2所示，预测结果可视化如图3所示。

从表2、图3的实验结果可以看出，相比其他 5 种方法，本文提出的预测方法在 4 个评价指标中都取得了最好的效果。本文提出的方法在数学成绩数据集上的预测  
准确率高达 ，精确率达到了 ，召回率达到了 -Measure 达到了 ，均明显高于其他方法。

表2 数学成绩数据集上的预测结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 方法 | Accuracy | Precision | Recall | -Measure |
| Ours | 93.06 | 93.90 | 93.06 | 93.01 |
| LR | 91.67 | 92.38 | 91.67 | 91.65 |
| RF | 91.67 | 92.17 | 91.67 | 91.63 |
| XGB | 90.28 | 91.10 | 90.28 | 90.17 |
| KNN | 84.72 | 89.34 | 84.72 | 85.03 |
| MNB | 52.78 | 58.60 | 52.78 | 49.11 |

![](data:application/octet-stream;base64,)

图3 数学成绩数据集上的预测结果  
相比本文提出的方法，逻辑回归、随机森林、极限梯度提升、 K 近邻的预测效果一般，而多项式朴素贝叶斯的效果更差。传统预测方法效果较差的原因主要是忽略了数据内部之间的联系，没有考虑不同的输入特征对学生的影响程度是不同的，将特征平等作为预测模型的输入。本文提出的方法运用多头自注意力机制，加强了数据内部之间的联系，提高了模型的学习能力，从而提升了模型的预测性能。

**3.5.3 消融实验**

为了进一步验证本文所提出的因果推断和多头自注意力机制的有效性，在数学成绩数据集上进行针对因果推断和多头自注意力机制的消融实验。NoTransformer（No-T）指的是去掉 Transformer 层，将因果推断的结果输入到线性连接层来预测期末成绩；NoCausal（No-C）指的是本文提出的方法不使用因果推断选取重要特征，把所有特征作为 Transformer块的输入来预测期末成绩；No-Causal-Transformer（No-C-T）指的是去掉因果推断层和 Transformer 层，使用线性连接层来预测 G3。消融实验的结果如表3所示，消融实验结果可视化如图4所示。  
表 3 消融实验结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 方法 | Accuracy | Precision | Recall | -Measure |
| Ours | 93.06 | 93.90 | 93.06 | 93.01 |
| No-T | 91.67 | 91.92 | 91.67 | 91.71 |
| No-C | 87.50 | 87.87 | 87.50 | 87.57 |
| No-C-T | 84.72 | 85.38 | 84.72 | 84.57 |

![](data:application/octet-stream;base64,)

图4 消融实验结果  
从表3和图4中可以看出：基于因果推断和 Transformer 的成绩预测方法的预测效果是最好的，NoTransformer 相较于 No-Causal-Transformer 准确率提高了大约 7%，其他指标也有大幅提升，这说明了因果推断的重要性，选取一定因果关系的特征，增强预测模型解释性的同时，提高了预测性能；No-Causal 方法相较于NoCausal-Transformer 方法，准确率提高了大约 ，分析原因是 Transformer层的多头自注意力机制在不同的空间内进行特征学习，不同空间的信息汇聚，加强了数学内部之间的联系，提升了模型的学习能力。基于因果推断和 Transformer 的成绩预测方法对比 No-Causal 和 NoTransformer，预测准确率分别提升了约 和 ，再次证明了因果推断和多头自注意力机制的优越性。

**4 结 语**

学生成绩预测作为计算机和教育学交叉领域的一个热门研究方向，旨在为学校管理者提供决策支持，帮助教师改进教学手段，对学生进行学业指导，最终达到提高学生成绩的目的。未来的研究工作中，随着高校信息化程度越来越高，可以获取更多学生成绩的影响因素，选取更有效的因素来提高预测性能，同时需要克服多头自注意力机制计算量大、实际应用成本高的问题。  
注：本文通讯作者为朱贵富。

**参 考 文 献**

[1] 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.  
[2] 任伟杰, 韩敏. 多元时间序列因果关系分析研究综述[J]. 自动化学报,2021,47(1):64-78.  
[3] 王东明,陈都釡. 因果推断: 起源和发展[J]. 控制工程,2021,29 (3) .  
[4] HUANG S, FANG N. Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models [J]. Computers & education, 2013, 61: 133-145.  
[5] PANDEY M, SHARMA V K. A decision tree algorithm pertaining to the student performance analysis and prediction [J]. International journal of computer applications, 2013, 61(13): 1-5.  
[6] 谢娟英, 张宜, 陈恩红. 学生成绩关键因素挖掘与成绩预测[J].南京信息工程大学学报(自然科学版),2019,11(3):316-325.  
[7] RUBIN D B. Estimating causal effects of treatments in randomized and nonrandomized studies [J]. Journal of educational psychology, 1974, 66(5): 688-701.  
[8] IMBENS G W, RUBIN D B. Causal inference in statistics, social, and biomedical sciences [M]. University of Cambridge: Cambridge University Press, 2015.  
[9] PEARL J. Causal diagrams for empirical research [J]. Biometrika, 1995, 82(4): 669-688.  
[10] JANZING D, HOYER P O, SCHÖLKOPF B. Telling cause from effect based on high-dimensional observations [EB/OL]. [2011-09-20]. <https://www.docin.com/p-261081261.html>.  
[11] 张浩, 郝志峰,蔡瑞初,等.一种适用于高维网络的方向推断算法[J].小型微型计算机系统, 2015,36(6):1358-1362.  
[12] POLYZOU A, KARYPIS G. Feature extraction for next-term prediction of poor student performance [J]. IEEE transactions on learning technologies, 2019, 12(2): 237-248.  
[13] MA Y, CUI C, NIE X, et al. Pre-course student performance prediction with multi-instance multi-label learning [J]. Science China information sciences, 2019, 62(2): 1-3.  
[14] OKUBO F, YAMASHITA T, SHIMADA A, et al. A neural network approach for students' performance prediction [C]// Proceedings of the Seventh International Learning Analytics & Knowledge Conference. New York: ACM, 2017: 598-599.  
[15] OUYANG Y, ZENG Y, GAO R, et al. Elective future: The influence factor mining of students' graduation development based on hierarchical attention neural network model with graph [J]. Applied intelligence, 2020, 50(10): 3023-3039.  
[16] 姚丽, 崔超然, 马乐乐, 等. 基于校园上网行为感知的学生成绩预测方法[J].计算机研究与发展，2022，59（8）：1770-1781.  
[17] LIAN D, YE Y, ZHU W, et al. Mutual reinforcement of academic performance prediction and library book recommendation [C]// 2016 IEEE 16th International Conference on Data Mining (ICDM). New York: IEEE, 2016: 1023-1028.  
[18] 田钰. 基于多元特征感知网络的高考成绩预测 [J]. 计算机集成制造系统，2021,27(9):2741-2748.  
[19] 李梦莹, 王晓东,院书岚, 等.基于双路注意力机制的学生成绩预测模型[J].计算机研究与发展,2020,57(8):1729-1740.  
[20] CORTEZ P, SILVA A M G. Using data mining to predict secondary school student performance [C]// Processing of the 15th Conference on European Concurrent Engineering. Ostend, Belgium: EUROSIS, 2008: 5-12.