文章类型：论文|刊号（ISSN）：2630-5178 /（中图刊号）：380GL019

基于数据建模的高校学业预警技术关键点分析

倪维晨
天津理工大学
DOI:10.12238/mef.v5i2.4818

[摘　要] 学业预警是高校学生学业状态的重要反馈预警机制，将其与基于数据驱动的机器学习建模方法结合，可提高预警的及时性与有效性。但在应用中却存在关键的技术瓶颈，本文对建模数据的获取、数据不平衡以及模型的可解释性问题进行了阐释与分析，并提出了对应的解决措施。
[关键词] 学业预警；机器学习；数据不平衡；模型可解释性
中图分类号：G647　　文献标识码：A

Analysis of Key Points of College Academic Early Warning Technology Based on Data Modeling

NI Weichen
Tianjin University of Technology

[Abstract] Academic early warning is an important feedback early warning mechanism for college students' academic status. Combining it with data-driven machine learning modeling method can improve the timeliness and effectiveness of early warning. However, there are key technical bottlenecks in the application. This paper explains and analyzes the acquisition of modeling data, data imbalance and the interpretability of the model, and puts forward the corresponding solutions.
[Key words] academic early warning; machine learning; data imbalance; model interpretability

学业预警制度是高校学业管理制度之一，其目的是在学生的学业可能出现或者已经出现问题时引起学生、教师以及家长的注意，从而避免学生出现更严重的学业问题。但目前的学业预警机制还不够完善。学业预警是以学期为周期的，要在一个学期结束后，使用各科期末成绩计算学业完成度，当完成度达不到培养计划与学籍管理规定的要求时进行预警，属于事后警示。预警出现时，学生的学业问题已经存在一段时间了。随着大数据时代的到来，数据挖掘、云端计算、机器学习等技术与高校教育的结合，使基于数据驱动的在校生学习状态分析的研究如火如荼。

在此背景下，近年来提出的“智慧教育”就是在以学生为中心的前提下，通过对教学数据与资源的采集与汇总，进行实时的教学监测，以关注每位在校大学生的学习情况，为教师、辅导员等教学管理者提供教学管理与措施的制定、学习困难学生重点帮扶的数据支撑。学界对大数据推动智慧教育的作用问题上多持积极、正面的态度，认为大数据作为一种工具，可以进一步量化高校的教育教学质量，为学生绘制学业“画像”，促进因材施教等。提高学业预警的时效性也是当前高校教育与大数据结合的热点研究方向，但目前来看，国内教育大数据的应用，特别是应用于高校学生学业预警的应用还不成熟，还有一些关键问题、难点问题亟待解决，只有解决这些问题，才能将目前的研究转变成可落地的“灵丹妙药”。本文就一些关键问题展开分析，并尝试从技术的角度给出一些可行的解决方案。

1 建模关键技术分析

1.1建模数据的获取

大数据技术需要使用采集数据构造一个与现实教育完全镜像的世界，这样才能以此为基础，开展数据分析以及学业异常的监测等工作。所以获取到最丰富完备的数据是建立模型的基础，如果这一环节没有做好，所建模型的准确性是无法保证的。然而，教育与学习是人的精神层面的活动，特别是学生的思想状态是无法像物质层面的事物一样能够被量化与采集的。学生的学习状态是受社会环境与学生自身等多种因素影响的。比如，学生看到社会生活中短视频与网络直播的少数从业人员的高收入，就片面地认为不学习、没有专业技能一样可以赚钱，导致对所学专业不感兴趣，沉迷游戏，无节制追剧，学业基础差，容易受家庭环境影响等。这些影响因素是无法通过测量手段直接获取的。

现在的学业预警数据中，各学期的成绩是最基础的数据。为了能更全面反映学生的学习状态，有些研究使用与校园卡关联的图书借阅情况、上网行为、

消费习惯与定位信息等，以建立完善的“学业画像”。但获得这些数据存在伦理维度的问题——学生的隐私。校园卡和校内账户的使用，方便了学生在校园的生活与学习，但对些数据全方位的过度的关注必然会侵犯学生的隐私权。从另一方面来看，学生的校园活动数据是实时产生的海量数据，完整获取、保存并运用这些数据是一个庞大的系统工程，需要大量存储与运算设备以及运维人员，大部分高校是负担不起的。因此这种方法在理论研究上可能有较好的效果，但实施难度较大。

在课程的成绩预测研究中，授课教师为了及时预测学生的学习效果，提出利用出勤情况、回答问题的次数、作业完成情况等过程性评价数据。这些数据是随着课程进行而产生的，其数据量也比期末成绩更大。可以考虑将这些数据作为学业预警模型的变量。这些数据是可以通过在线的教学平台如雨课堂、智慧树等获得的。由于疫情的影响，目前基本都采取线上授课形式，教师与学生对这些教学平台都是比较熟悉的。与学生的校园卡生活数据相比，此方法获得的数据不涉及学生隐私，而且也是更直接、细化的学习数据，可以更完整、更及时地体现学生的学习状态。

从建模角度来看，模型的输入变量应具有较高的信息熵且不同变量之间应有较低的互信息，即低耦合。但在实际建模时，为了获得更准确的模型，往往使用较多的输入变量且变量之间的耦合度较高。输入变量的数量直接关系到建模方法的选取。因此，为获得更多样的数据，选择的建模方法要适用于高维输入的应用特点。

1.2数据不平衡问题

一个分类数据集内若存在类别数量比例的失衡，就被称为不平衡。在数据集中占比较多的类被称为多数类，而那些占比例较少的类则是少数类。对于学业数据而言，正常无预警的学生为多数类，有预警的学生为少数类，且少数类与总体的比远小于1%，属于极端不平衡。在使用这样的数据建模时，多数类样本的信息量会淹没少数类样本，导致少数类的严重分类误差。但在学业预警的建模研究中此问题却没有被重视起来。

学业预警与金融领域的欺诈、贷款违约监测，生物医学领域的疾病诊断、蛋白质鉴定，工业系统领域的设备故障监测等数据不平衡情况类似，仅仅依靠分类的准确性来衡量模型的性能是不可行的。比如，100位学生中仅1位学生有学业预警，若模型仅简单地将所有学生都分类为正常无预警，模型的准确率可达99%。准确度虽然高，但没有起到预警的作用。学界常使用F-measure来评价不平衡数据集的建模效果，它的定义为：

$$F_1 = \frac{2 \times precision \times recall}{precision + recall}$$

其中，precision为精确率，代表预测结果中被正确预测为预警学生的个数占预测结果中被预测为预警学生数的百分比；recall为召回率，代表被正确预测为预警样本数占样本中真实的预警样本数量的百分比。上述例子中因为没有预警样本，所以其recall=0，导致F-measure为0。precision与recall两个指标也可以用来综合评价模型，但同一个模型往往无法同时提高这两项指标，即模型的精确率高时召回率低，反之也是如此。对于学业预警问题，我们更关注召回率，即需要把所有应预警的学生都预测出来，不要漏报，即使会因此产生少量的误报也没有关系。

在数据预处理层面，应考虑重采样技术，这种方法被用来对少数类或多数类进行过采样或欠采样。对于一个不平衡的数据集，我们可以使用替换法对少数类样本进行超额多次取样。这种方法被称为过采样。同样地，我们可以随机删除多数类的样本，使其与少数类相匹配，这被方法被称为欠采样。在对数据进行重采样后，可以得到一组平衡的数据集。对少数类进行过采样，会使模型过分集中于过采的样本，从而导致模型的过拟合现象。

为了解决这一问题，Chawla在2002年提出了SMOTE算法，此算法及其改进算法已经是处理数据不平衡问题的常用方法。其基本思想是不直接使用已经存在的样本补充少数类，而是通过人为合成的方法产生新的样本。新样本的生成方法可简单总结为：将随机选中的少数类样本与其邻近的一个随机样本做差，将差值乘以（0，1）范围之内的数，然后将乘积加到选中的样本上，以构造出新样本。这样做不会改变原数据集中的样本分布，并且可以通过合成的样本使分类器建立更大的少数类决策区域。这样就可以避免过拟合现象，从而提高模型对少数类识别的准确度。

如果将SMOTE算法直接应用于学业预警，可能还存在问题。由于学业数据中包含各门课程的总成绩及每门课的过程性考核数据，数据维数过多，直接使用SMOTE算法的结果还不及欠采样方法。这就需要使用数据降维算法对原数据进行降维操作。数据降维算法可分为选择方法与融合方法，即特征选择方法与特征融合方法。常用的特征选择方法包括：方差选择、相关系数选择、互信息选择以及基于树模型的选择方法等。常用的特征融合方法包括：主成分分析法（PCA）及其改进算法，以及线性判别方法。特征融合方法由原模型的变量生成新的降维后的变量，由于新变量是由原变量融合生成的，所以已经失去了物理意义。而由特征选择方法获得的降维变量仍然保留其物理意义，这对于模型的理解是有益的。所以，建议优先使用特征选择方法进行降维，然后再使用SMOTE算法构建模型的训练样本集。

在算法层面，应考虑使用代价敏感学习与单类学习等方法。代价敏感学习的特点在于对不同类数据的错分有不同的惩罚强度，在数据不平衡问题中，加强了少数类错分的惩罚力度，并以总体错分代价最低为学习目标。单类学习只对多数类样本进行学习，预测时通过相似性度量方法判断样本是否属于异常点，相对多类别的异常点即为少数类。

1.3模型预警结果的可解释性

数据驱动的机器学习模型方法属于黑箱建模方法，其建模过程如图1中“标准机器学习过程”所示。作为最终的用

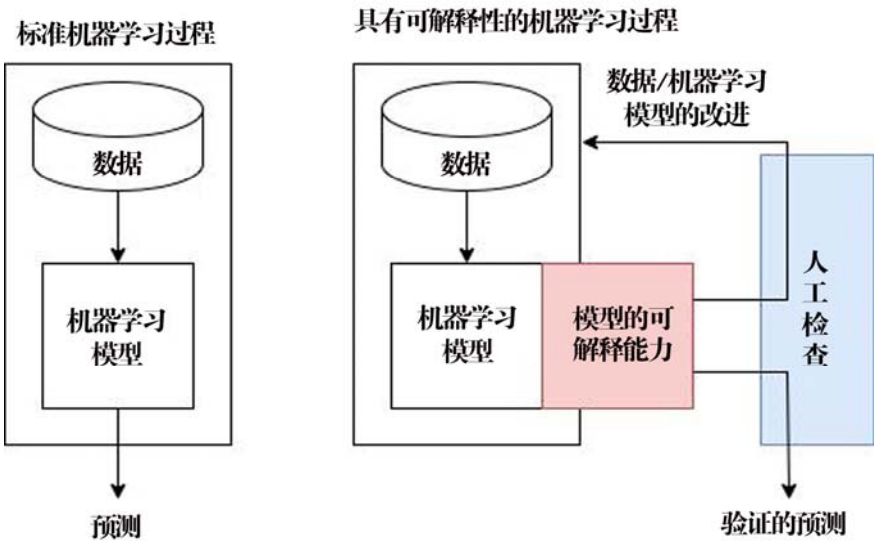

图 1 标准机器学习过程与具有可解释性的机器学习过程

户，高校的管理者只能获得预测结果，却无法了解做出具体预测的原因，这会引起人们的怀疑。只有当用户能够理解他们为什么要做出具体的决定时，才会信任这个模型，并产生使用的意愿。具有可解释的机器学习方法将内部运行机制或者预测的原因呈现给用户，这样教育管理者不仅可以得到更准确的验证后的预测结果，还可以了解预测背后的原因。同时，对于用户来说，若建模数据或方法中存在显而易见的错误，可立即发现并改正，如图1中“具有可解释性的机器学习过程”所示。

一个较为简单的增强可解释性的做法是将预警学生的数据与平均值比较。例如，平均出勤率为90% ，而预警生的出勤率只有50%;多个课程作业的平均分为80%，而预警学生为40%。当然也可以对综合指标进行这样的比较：将所有技术类课程（如单片机、C语言、现代电气控制）的出勤率平均值、课程作业平均值与预警学生的对应数值相比较。上述方法是较简便的，为了更好从根本上理解与厘清学业预警的运作机理，实际上应从教育学原理、课程与教学论、教育心理学、教育技术等教育学与心理学的基本理论出发，探索与学业预警相关的基本规律与原理，在此基础上，指导学业预警模型的特征选择、模型构建，并与模型的预警结果结合，从根本上增强模型的可解释性，同时提高模型的预报性能。

2 结语

大数据在教育领域的应用，特别是与学业预警的结合，是带有技术复杂性的。这提醒我们在进行大数据应用研究时，一定要考虑实际的应用需求以及应用的特点。现实中，大数据在教育领域的应用还远不如其他领域，其中的原因值得深思。本文仅从技术角度进行了分析，当然还有其他限制性因素。如何看待并解决这些问题，直接关乎大数据与结合教育之路当如何前行，以及如何走得更远。

[参考文献]

[1]吴文峻.面向智慧教育的学习大数据分析技术[J].电化教育研究,2017(6):88-94.

[2]唐汉卫,张姜坤.大数据教育应用的限度[J].华东师范大学学报(教育科学版),2020(10):60-68.

[3]韩凤霞.大数据时代高校学籍预警机制的探索与研究[J].中国教育信息化(高教职教),2015(10):46-49.

[4]Cao Y. Portrait-Based Academic Performance Evaluation of College Students from the Perspective of Big Data [J].International Journal of Emerging Technologies in Learning (iJET), 2021, 16(04):95.

[5]李艳霞,柴毅,胡友强,等.不平衡数据分类方法综述[J].控制与决策,2019(04):673-688.

[6]万建武,杨明.代价敏感学习方法综述[J].软件学报,2020(1):113-136.

[7]周国华,申燕萍,殷新春.基于凸壳的在线单类学习机[J].西南大学学报(自然科学版),2018(12):163-172.

作者简介：

倪维晨（1985-），女，汉族，天津市人，助理研究员，硕士，研究方向：教育管理。