doi: 10.3969/j.issn.0490-6756.2019.02.014

基于数据挖掘技术的学生成绩预警应用研究

刘博鹏[1], 樊铁成[2], 杨 红[1]
(1. 大连海事大学信息科学技术学院, 大连 116000; 2. 大连海事大学智慧校园研究中心, 大连 116000)

摘 要: 学习成绩是评价一个学生学习情况的最重要最基础的指标,对学习成绩的分析有利于老师掌握学生的学习情况,进行针对性地进行教学辅导,而对学生而言,能提前知道自己未来课程在学习过程中出现的情况也有利于学生发现自身存在的问题并提前加以防范. 现有的研究工作大多是基于对课程、历史成绩或行为数据的分析来对学生的总成绩进行预测,很少有研究将学生行为与学生课程成绩等方面结合起来综合全面的预测学生未来所有的课程的学习情况,对此,本文从一个新的角度出发,利用学生的行为、个人属性和历史成绩等三个方面数据,根据学生未来不同课程动态的进行影响因素的选择,并利用支持向量机对学生成绩进行预警,为数据挖掘技术在教育领域的应用做了一些探索性工作.

关键词: 教育; 所有课程; 动态特征选择; 成绩预警; 支持向量机
中图分类号: G642 **文献标识码**: A **文章编号**: 0490-6756(2019)02-0267-06

Research on application of early warning of students' achievement based on data mining

LIU Bo-Peng[1], *FAN Tie-Cheng*[2], *YANG Hong*[1]
(1. Institute of information science and technology, Dalian Maritime University, Dalian 116000, China;
2. Wisdom research center, Dalian Maritime University, Dalian 116000, China)

Abstract: Academic achievement is the most important and most basic indicator for evaluating a student's learning situation. The analysis of academic performance is beneficial to teachers to master a student's learning situation and conduct the targeted teaching and counseling, while for the students, knowing in advance what happens to their future learning is also beneficial for students to discover their own problems which could be prevented in advance. Most of the existing research work is based on the analysis of the curriculum, historical performance or behavioral data to predict the student's total score, while few studies focus on combining a student behavior with his or her grades to comprehensively predict a student's learning in all future courses. This paper, from a new perspective, uses the three aspects data of student behavior, personal attributes and historical achievements which are identified by the influencing factors based on students' different curriculum dynamics in the future, the early warning of students' achievement is predicted with support vector machine, the experiment results show the exploratory work of data mining applied in education has certain significant meaning for the teachers as well as for students.

Keywords: Education; All courses; Dynamic feature selection; Performance warning; Support vector machine

收稿日期: 2018-07-20
基金项目: 中国高等教育学会重点课题 (2016XXZD03)
作者简介: 刘博鹏(1994-), 男, 甘肃白银人, 硕士研究生, 主要研究领域为教育数据挖掘.
通讯作者: 樊铁成. E-mail: ftc@dlmu.edu.cn

1 引 言

在高校的教学活动中,学生成绩是评估学生对所学知识掌握程度的重要依据,也是评估老师教学质量好坏的关键因素[1].现如今,随着管理模式的不断改进创新以及在线教育的迅猛发展,大学生可以根据自身的实际情况来制定学习计划,自由度很高,这也导致了影响学生学习成绩的因素不再单一.大学校园本身就是一个小型的社会系统,其内部的服务体系可以满足学生大部分日常生活的需求,随着各个高校招生规模的扩大和信息技术的发展,各个高校都建立自己完备的信息系统以提高自身的管理效率的同时也积累了海量的数据,如:学生的课业成绩数据以及就餐、沐浴、上网等日常生活行为数据.然而,大家并不知道这些数据与学生未来学习成绩之间的关系,如果对这些数据加以分析,发现其背后隐藏着的信息和规律,从而通过学生的行为数据以及历史课业成绩数据对学生未来的课业成绩加以预测不仅有利于教师在授课之前制定针对性的计划便于管理,也有利于学生发现自身存在的问题并提前加以防范,提高学习效率.因此,使用数据挖掘技术对学生成绩进行分析与预测,在实际应用中具有极为重要的意义[2].

目前国内外关于学生在校成绩的影响因素与预测方面的研究有很多,比如,胡祖辉[3]等人,以学生在校的上网数据和学生成绩数据为数据基础,采用决策树,逻辑回归等数据挖掘算法,对学生上网行为的相关属性与学生学习质量之间的关系进行研究,并根据最终挖掘结果,划定了合理的上网时长,为网络部门的管理提供了合理的依据.张静[4]等人通过FP-growth关联规则算法,发现了学校各个课程之间在成绩方面的关联关系,并为学校提供了合理的课程安排修改建议.刘譞等人提取了学生在校的行为数据和学生的学习成绩数据,通过C4.5,朴素贝叶斯等分类算法对学生未来的学习成绩进行分类[5].吴暾华等人则在刘譞等人的基础上,添加了学生过往的学习情况维度,来预测学生未来的整体学业表现[2].通过阅读文献我们可以发现,虽然目前国内外关于学生在校成绩的影响因素与预测方面的研究有很多,但大多都是预测学生未来的总体学习情况,并不能很好地具体预警未来某门学科的学习情况,因此本文在综合目前国内外关于学生未来学业预警问题的研究基础之上,通过分析学生在校课程与课程之间,课程与行为之间的关联关系,动态组合处理相应的特征维度,从而达到预测学生未来各门课程成绩的目的.

2 动态特征选择

2.1 学生行为及个人属性特征的动态选取

本文所做工作的目的是找出学生在校的日常生活行为,个人属性及过往学习成绩对未来的学习成绩的影响.在学生日常行为及个人属性特征提取的问题上,本文结合文献[1-5],选取了学生性别、生源地等个人属性特征,学生的消费、门禁、上网等日常行为特征共36维特征作为备选特征.但由于影响学生课程成绩的因素有很多,且不同课程的影响因素也有所不同.比如,对于学生的英语课程成绩而言,女生的成绩[6]在总体上是略高于男生的;而对于早上一二节开始的课程而言,是否吃早餐这个行为特征很大程度的反映了学生该门课程的学习情况[6].如果选取所有的这些特征来对学生未来课程成绩进行预警,不仅影响算法的执行效率,而且会带来很大的噪声从而影响算法的准确性.

互信息(MI)能够定量地反映变量间的相关程度,并且在描述线性和非线性变量时都具有较好的性能.文献[7]提出用条件期望剔除变量间的相关关系再计算其互信息,称为偏互信息(PMI)法,有效地提高了变量选择的精确性,已被应用于火电厂选择性催化还原法(SLR)脱硝系统建模的输入变量选择,结果表明PMI算法有良好的学习和泛化能力.因此,本文采用偏互信息(PMI)法来动态的选取适合于不同课程的不同行为特征.

2.2 学生课程关联网络构建

学生过往的学习成绩对于学生未来的学生成绩也起着不和忽视的作用,如果学生在大学期间如高数,线代等基础课程学习状态不佳对于未来专业课程的理解可能有阻碍作用[3].基于此,本文利用改进的apriori关联规则算法[8],对往届毕业生所有课程成绩进行关联性分析.由于全体学生的数据量过于庞大,在现有条件下无法处理,同时不同专业所学课程以及上课时间的差异会使影响因素变得更加复杂,为了使数据挖掘达到一个较好的效率和准确的结果,本文只采集了本校2009级——2013级海事管理专业毕业生的所有数据.经反复实验,最终确定最小支持度设为0.38,最小置信度设为0.6,Kulc值在0.3到0.7之间,IR值小于1,将不符合学校教学计划的关联规则删除课程名和对应标识符示以及最终得到的课程之间的有向关

系网络如表 1 和图 1 所示.

表 1　课程名和对应标识符

Tab. 1　Course name and corresponding identifier

标识符	1	2	3	4	5	6	7	8	9	10
课程名	体育(1)	大学英语(2)	高等数学(1)	大学英语听说(1−1)	船舶结构与设备	海洋法公约	体育(2)	高等数学(2)	大学物理	大学物理实验
标识符	11	12	13	14	15	16	17	18	19	20
课程名	中国近现代史纲要	计算机程序设计基础(C)	大学英语听说(1−2)	轮机管理	行政法与行政诉讼法	体育(3)	船舶防污染法	海上安全公约	马克思主义基本原理	大学英语听说(1−3)
标识符	21	22	23	24	25	26	27	28	29	30
课程名	体育(4)	船舶污染监测技术	航海学(1)	海事信息管理	航海气象与海洋学	大学英语听说(1−4)	毛泽东思想和中国特色社会主义理论体系概论	航海学(2)	船舶原理与货运	GMDSS 通信业务与设备
标识符	31	32	33	34	35	36	37			
课程名	精通救生艇筏和救助艇	高级消防	精通急救	公务员考前培训	船舶与船员管理	海事调查与分析	海上搜寻与救助			

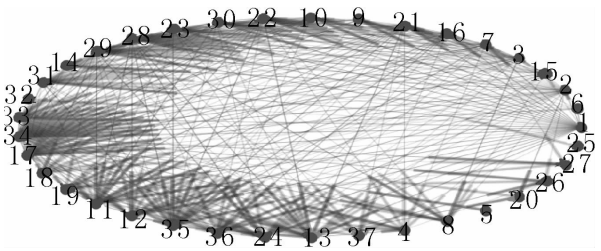

图 1　课程间的有向关系网络

Fig. 1　The directed relation network between courses

3　基于动态特征选择与支持向量机的学生成绩预警方法

支持向量机(SVM)是一种二分类模型，它的基本模型是定义在特征空间上的间隔最大的线性分类器[9]. 基于 SVM 的学习成绩预警方法的步骤如下.

第 1 步，结合文献构建学生在校影响学习成绩的特征向量，本文主要提取了学生在校的日常生活行为，个人属性及过往学习成绩等数据来作为本次研究的样本集合，并将数据进行标准化处理，以消除量纲不同而带来的影响.

第 2 步，根据课程的不同动态的选取不同的特征向量，具体的动态特征选择方法如上文所述. 例如，该专业学生的航海学这门课，在日常行为与个人属性特征上性别，早餐次数，图书馆门禁次数与该门课程互信息较大，而在过往的学习成绩中，高等数学，轮机管理，船舶污染法等课程与该门课程呈现出较强的关联关系，因此选取这些特征来作为预测特征，将预测结果分为三类，分类规则参考文献[2]，具体分类方法如表 2 所示.

表 2　分类规则

Tab. 2　Classification rule

学业状态	好	中	差
量化方法	绩点 3.5 以上	绩点 2.8−3.5	绩点 2.8 以下
预警结果	给予表扬	辅导员或班主任督促改进	协同领导，辅导员，家长等多方共同督促改进

第 3 步，选用高斯核函数并采用交叉验证的方式来选取合适的参数，惩罚因子 C 值，并进行分类. 由于本实验属于三分类问题，而传统的 SVM 解决的是二分类问题，因此，本文采用文献[10−11]所提出的”一对多”SVM 分类器的方法，即训练时依次把某个类别的样本归为一类，其他剩余的样本归为另一类，这样 3 个类别的样本就构造出了 3 个 SVM 分类器. 分类时将未知样本分类为具有最大分类函数值的那一类.

第 4 步，利用测试数据集来验证模型构建的准确性，若正确率达到目标阈值，则可用来对该专业未来任意学生该门课的学习情况进行预测，反之则重构样本集，调整参数再次学习.

基于动态特征选择与支持向量机的学生成绩预警方法流程图如图 2 所示.

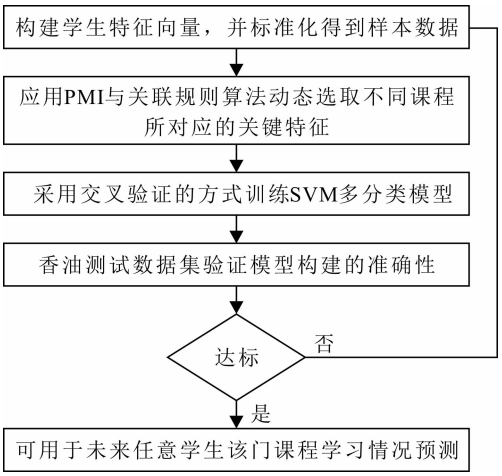

图 2 学生成绩预警方法流程图

Fig. 2 Flow chart of student achievement early warning method

4 实验与结果分析

4.1 实验方法及结果

sklearn 是机器学习中一个常用的 python 第三方模块,里面对一些常用的机器学习方法进行了封装. 我们用 sklearn 中的实现的"一对多"SVM 分类器和 PMI 特征选选择则器来进行本此实验.

本文以该专业学生大二下学期的所有课程为例进行实验验证. 该专业学生大二下学期进行的课程有海事信息管理,航海学,大学英语,大学英语听说,船舶污染物监测技术,毛泽东思想和中国特色社会主义理论体系概论(以下简称毛概)六门课. 首先,对我们提取出的学生行为及个人属性维度进行归一化处理,以消除量纲不同所带来的影响并对这六门课程以上文的分类方法分成三类. 然后,应用 PMI 变量选择的方法动态选择这六门课程行为及个人属性特征选择并应用 apriori 关联规则算法选取与之有关的学生过往课程学习成绩,动态特征选择的结果如表 3 与表 4 所示(0 代表不入选,1 代表入选).

表 3 行为特征动态选择结果

Tab. 3 Dynamic selection results of behavior characteristics

特征 课程	性别	地域	平均上网时长	……	早餐次数	一卡通消费频率
航海学	0	1	1	……	0	0
大学英语	1	1	0	……	1	0
大学英语听说	1	1	0	……	1	0
海事信息管理	0	0	0	……	1	0
船舶污染物监测技术	0	0	1	……	1	1
毛概	1	0	0	……	0	0

表 4 过往成绩特征动态选择结果

Tab. 4 Dynamic selection results of past performance characteristics

课程名	课程名	支持度	置信度	Kulc 值	IR 值
高等数学(1)	船舶污染物监测技术	0.57	0.95	0.7	0.98
高等数学(1)	海事信息管理	0.47	0.86	0.7	0.62
高等数学(1)	航海学	0.45	0.77	0.7	0.16
大学物理	航海学	0.42	0.82	0.6	0.24
大学物理	船舶污染物监测技术	0.41	0.80	0.6	0.73
船舶污染法	船舶污染物监测技术	0.41	0.72	0.7	0.2
……	……	……	……	……	……
中国近现代史纲要	毛泽东思想和中国特色社会主义理论体系概率	0.39	0.70	0.7	0.19

为了避免不同专业的差异以及数据量不足从而导致难以反映影响因素的情况发生,本研究采用了本校海事管理专业 2009～2013 级五届,共 273 位毕业生的数据及成绩数据,其中 2009～2012 级毕业生数据训练集,而 2013 级的毕业生数据则作为测试集.

本研究采用交叉验证的训练方式来确定惩罚因子 C 值,及高斯核函数 γ 值. 从训练集中划出部分样本用作 v-flod 交叉验证,称其为交叉验证集. 将交叉验证集平均分成 v 份. 对于参数 C 与 γ 的不同组和做如下操作:按顺序保留 1 个子集作为测试集,其他 $v-1$ 个子集作为训练集并训练其得到 SVM 分类器,利用该分类器对测试集进行测试并记录准确率,直到所有子集都被测试过,再取这 v 次测试的平均准确率作为该次交叉验证的准确率. 这个过程相当于对 C 和 y 进行遍历,最终选择交叉验证准确率最高时的 C 和 γ. 经过交叉验证得到的最佳参数及其预测精度如表 5 所示,在训练集上的混淆矩阵如图 3 所示.

表 5 最佳参数及预测精度表

Tab. 5 Optimum parameters and their prediction accuracy

课程名	C 值	γ 值	测试精度	训练精度
航海学	1.2	0.1	81%	72%
大学英语	8.85	0.1	87%	73%
大学英语听说	0.78	0.2	79%	73%
海事信息管理	0.48	0.1	62%	61%
船舶污染物监测技术	0.18	0.1	83%	75%
毛概	0.29	0.2	84%	54%

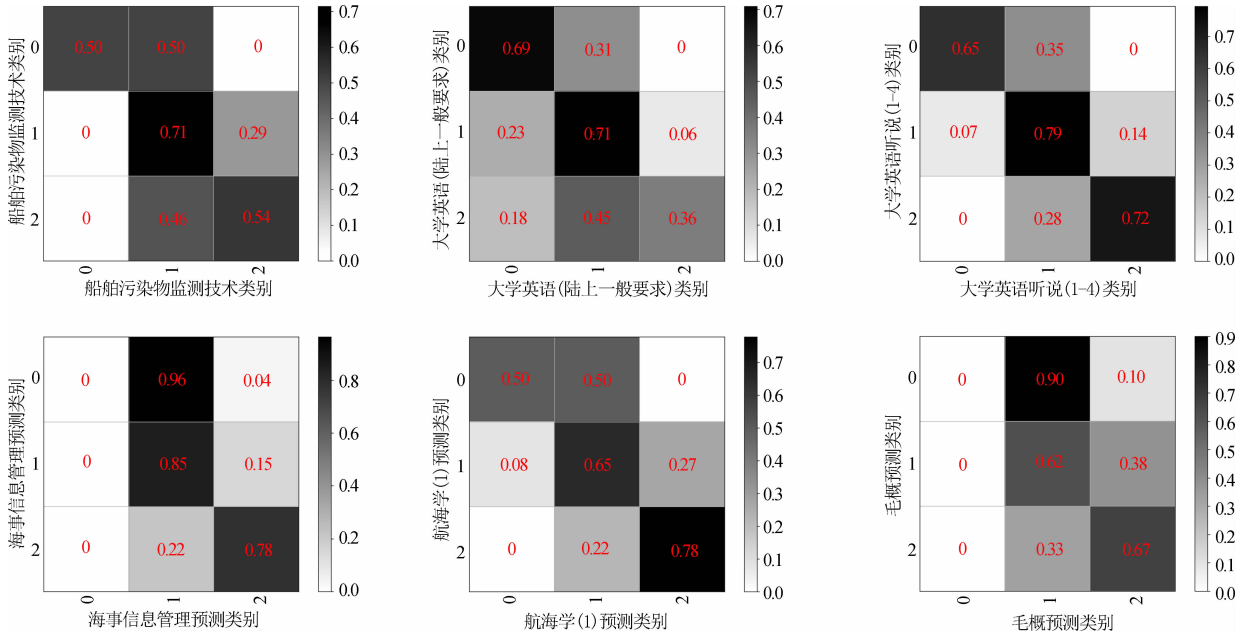

图 3　混淆矩阵

Fig. 3　Obfuscation matrix

4.2　实验结果分析

本实验结果的准确率主要受分类器的性能，特征维度与课程之间的相关性，以及类别划分三个方面的影响. 从分类器自身的性能角度出发，由于在交叉验证时实际上按“一对多”的方式训练二分类并测试准确率，而在实际训练时我们采用“一对一”方式训练，相当于拿“一对多”时分类器最优的参数去训练“一对一”分类器，存在一定的偏差，性能往往有所下降；而在特征维度的选取方面，所构建的总的行为特征集与过往学习成绩特征集中的特征并不一定与目标课程成绩存在很强的相关性，这就会导致无论是在训练集还是在验证集都不一定会取得很好的效果如表 5 与图 3 所示的海事信息管理这门课；而在分类类别的划分上，由于我们将课程根据成绩划分为三个类别，而在毛概这门课程中，绩点低于 2.8 的仅有 12 人，大部分人的绩点集中在 2.8～3.5 之间，这就导致样本分布极不均衡，进而导致分类预警效果不佳，所以在类别的划分上，不同的课程应采取更有针对性的划分方法.

5　结　论

针对学生未来课程成绩预警的问题上，构建学生行为，个人属性及过往学习成绩的特征向量并根据不同课程的特点进行动态选取，结合支持向量机来学习得到不同课程的预警模型，可自动发现学习状态不佳的同学并给予警示. 实验结果表明在大部分场景，本方法可有效提高学业监督效率，对于提高学生的个性化管理水平具有推动作用；同时，还可发现影响该课程成绩的行为及过往成绩因素，无论对于对教育工作者还是学生本身都有一定的参考价值.

参考文献：

[1]　周玉敏. 基于 Rough 集的数据挖掘在教学评价中的应用 [J]. 重庆邮电大学学报：自然科学版，2008，20：627.

[2]　吴暾华，王萍，刘婷. 基于支持向量机的大学生学业动态预警研究 [J]. 中国教育信息化，2017，23：65.

[3]　胡祖辉，施佺. 高校学生上网行为分析与数据挖掘研究 [J]. 中国远程教育，2017，37：26.

[4]　张静. 大数据技术在学生业绩分析中的研究与应用 [D]. 长春：吉林大学，2016.

[5]　刘譞. 基于学生行为的成绩预测模型的研究与应用 [D]. 成都：电子科技大学，2017.

[6]　袁洁婷，高志强. 性别因素对大学生英语成绩影响的定量分析 [J]. 海外英语，2015，18：68.

[7]　刘吉臻，秦天牧，杨婷婷，等. 基于偏互信息的变量选择方法及其在火电厂 SCR 系统建模中的应用 [J]. 中国电机工程学报，2016，36：2438.

[8]　曹莹，苗志刚. 基于向量矩阵优化频繁项的改进 Apriori 算法 [J]. 吉林大学学报：理学版，2016，54：349.

[9]　胡世前，姜倩雯，凌冰，等. 基于改进支持向量机

的空气质量监测预警模型 [J]. 江苏大学学报：自然科学版，2016，37：491.

[10] Silva C，Ribeiro B. Multiclass ensemble of one-against-all SVM classifiers [C]// International Symposium on Neural Networks. Berlin，Germany：Springer International Publishing，2016.

[11] López J，Maldonado S，Carrasco M. A robust formulation for twin multiclass support vector machine [J]. Appl Intell，2017，47：1.

引用本文格式：

中　文：刘博鹏，樊铁成，杨红. 基于数据挖掘技术的学生成绩预警应用研究 [J]. 四川大学学报：自然科学版，2019，56：267.

英　文：Liu B P，Fan T C，Yang H. Research on application of early warning of students' achievement based on data mining technology [J]. J Sichuan Univ：Nat Sci Ed，2019，56：267.