基于特征选择的学位预警方法研究

王　娜[1],李劲松[1],潘子尧[2],姚明海[1*]
(1. 渤海大学 信息科学与技术学院,辽宁 锦州 121013;
2. 渤海大学 数学科学学院,辽宁 锦州 121013)

摘　要:高校学生能够顺利获得学位,不仅对其个人就业发展至关重要,也是衡量高校教学质量的重要指标之一。学位预警是教育数据挖掘的重要应用之一,通过学位预警可以尽早地对学生的学业情况进行警示,学生能够及时调整学习状态和方法,同时准确的学位预警也可以为改进教学指导策略提供参考依据。现有的预警模型构建多是基于全部成绩数据,忽略了课程间的冗余性,使得构建的模型精度不足。因此,提出基于 Fisher 特征选择方法构建学位预警模型。利用 Fisher 得分对特征进行初步筛选;然后,利用筛选后的特征构建学位预警模型;最后,通过预警模型对获得学位情况进行预测。为检验方法的有效性,在某高校汉语言文学、化学、数学与应用数学等专业真实数据上进行了大量实验。实验结果表明,基于特征选择的学位预警方法具有良好的准确度和实用性,可以为高校学生的学位预警工作提供数据支持。

关键词:教育数据挖掘;特征选择;学位预警;支持向量机;成绩预测
中图分类号:TP301　　**文献标识码**:A　　**文章编号**:1673-629X(2023)09-0024-06
doi:10.3969/j.issn.1673-629X.2023.09.004

Research on Degree Early Warning Method Based on Feature Selection

WANG Na[1],LI Jin-song[1],PAN Zi-yao[2],YAO Ming-hai[1*]
(1. School of Information Science and Technology,Bohai University,Jinzhou 121013,China;
2. School of Mathematical Science,Bohai University,Jinzhou 121013,China)

Abstract:The successful acquisition of a degree by college students is not only crucial to their personal employment development,but also one of the important indicators to measure the quality of college teaching. Degree early warning is one of the important applications of educational data mining. The degree warning can warn students of their degree information as early as possible,student can adjust their learning state and methods in time. At the same time,accurate degree warning can provide reference for improving teaching guidance strategies. The existing early warning models are mostly built based on all performance data, which makes the accuracy of the constructed model is insufficient. Therefore,the degree early warning model based on Fisher feature selection method is proposed. Firstly,Fisher's score is used to preliminarily screen the features. Then,the degree warning model is built with the selected features. Finally,the degree obtaining situation is predicted through the early warning model. In order to test the effectiveness of the proposed method,a large number of experiments were carried out on the real data of seven majors of Chinese language and literature major, chemistry major,and mathematics and applied mathematics in university. The experimental results show that the proposed degree early warning method based on feature selection has excellent accuracy and practicality,and can provide data support for the degree early warning of college students.

Key words:education data mining;feature selection;degree early warning;support vector machines;performance prediction

0　引　言

2022 年 10 月 16 日,习近平总书记在中国共产党第二十次全国代表大会上的报告中明确提出“高质量发展是全面建设社会主义现代化国家的首要任务”[1]。推进“高质量发展”离不开高质量的人才队伍,所以提升教学质量具有重大的现实意义[2]。早在 2019 年中共中央、国务院印发的《中国教育现代化 2035》中就明确指出,要充分“利用现代技术加快推动人才培养模式改革,实现规模化教育与个性化培养的有机结合”着力提高教育质量。基于数据挖掘相关理

收稿日期:2022-11-11　　**修回日期**:2023-03-16
基金项目:辽宁省社会科学规划基金项目(L22BTJ002)
作者简介:王　娜(1981-),女,硕士,讲师,从事教育理论与实践研究;通讯作者:姚明海(1980-),男(锡伯),博士,副教授,从事模式识别与智能计算方向研究。

论的教育数据挖掘(Educational Data Mining,EDM)可以从各种教育数据中挖掘数据背后的教育规律,并可以为提高教育质量提供数据支撑,已经成为当前教育工作者们关注的焦点[3-4]。成绩预测或成绩预警作为EDM的热门研究领域之一,通过构建有效的预测或预警模型预测其学习成绩,发现成绩可能不理想甚至有辍学风险的学生,为教师提供精确的教学指导,为学生改变学习方式提供重要的参考依据,对提高教学质量具有重要的应用价值和现实意义[5-6]。

近年来,国内外学者在成绩预测和成绩预警方面开展了相关研究工作。刘晓雲等人利用多元线性回归方法构建预测高校学生毕业成绩的模型[7]。Chen 等人基于梯度提升决策树算法、人工神经网络算法和 K-means 算法,构建了一个基于学生行为特征的分析预测平台[8]。虽然国内外学者已经对成绩预测展开了相关研究,但是随着大数据时代的到来,与学生成绩相关如学生行为记录、学生消费习惯等教育数据变得越来越庞大。因为课程相关性,数据存在冗余信息等原因会影响基于这些数据构建的成绩预测或预警模型的性能。因此,有些学者开始尝试利用特征选择的方法对数据进行筛选。Gitinabard 等人采用特征选择和逻辑回归的方法来预测学生是否退课[9]。Thaher 等人利用改进的鲸鱼优化算法从学生成绩中选择出有助于构建精准预测模型的特征[10]。虽然国内外学者已经开展了相关的研究工作,但如何构建更为精准的成绩预测或预警模型仍是关注重点。

众所周知,学位能否顺利获得深刻地影响着学生的未来发展[11]。如果能在大学初期就可以向获得学位存在风险的同学发出预警,就可以督促学生及时改进学习方式,保证其顺利毕业。因此,提出基于特征选择方法构建更为精准的学位预警模型。

1　相关理论

特征选择是为了构建更精准的学习模型而从原始特征中选择出一个特征子集的理论方法。在特征选择的过程中可以有效地去除噪声、冗余等干扰信息,高效地进行维数约简,进而提高学习性能,增加对学习结果的理解[12]。

1.1　Fisher 特征选择

基于 Fisher 得分的特征选择方法是依据 Fisher 得分来寻找一组具有最好判别能力的特征子集的有监督特征选择方法[13]。其定义如公式(1)所示:

$$F_j=\frac{n_{y=+1}(\mu^j_{y=+1}-\mu^j)^2+n_{y=-1}(\mu^j_{y=-1}-\mu^j)^2}{n_{y=+1}(\sigma^j_{y=+1})^2+n_{y=-1}(\sigma^j_{y=-1})^2} \tag{1}$$

其中,$n_{y=+1}$ 和 $n_{y=-1}$ 分别为正负样本的数量;$\mu^j_{y=+1}$ 和 $\mu^j_{y=-1}$ 分别是正负样本第 j 个特征的均值;$\sigma^j_{y=+1}$ 和 $\sigma^j_{y=-1}$ 分别是正负样本第 j 个特征的标准差。F_j 值表明第 j 个特征的判别能力越强,F_j 值越大,说明该特征越重要。

基于 Fisher 的特征选择过程描述如下:

输入:训练样本集。其中,X 表示 n 个具有 d 维特征的样本;Q 是全体特征集合。

(1)利用 Fisher 算法,计算 d 维特征的 Fisher 得分,并按照得分将 d 维特征由大到小排列,得到新的特征集合 $\tilde{Q}$;

(2)构建特征集 D 为空集,特征维数 $\tilde{d}=d$,w 为阈值;

(3)从特征集 $\tilde{Q}$ 中依次选出一个特征构建新的特征子集 $\tilde{D}=[D,\tilde{Q}_i]$,并计算其特征评价函数 $f(X_{\tilde{D}})$;

(4)若 $f(X_{\tilde{D}})-f(X_D)>w$,则更新特征集 $D=[D,\tilde{Q}_i]$,特征维数 $\tilde{d}(\tilde{d}=\tilde{d}-1)$,特征集 $\tilde{Q}=[\tilde{Q}_{i+1},\cdots,\tilde{Q}_{\tilde{d}}]$,转步骤(3);

(5)若 $f(X_{\tilde{D}})-f(X_D)<w$,即新选入特征无法使评价函数性能进一步提升,则算法结束。

输出:选择出的特征集 D。

Fisher 特征选择方法通过计算原始特征的重要性得分来进行特征选择,方法简单、直观。因此,为了消除原始课程数据间的冗余信息,采用 Fisher 特征选择方法来为构建的学位预警模型筛选特征信息。

1.2　支持向量机

基于结构风险最小化理论的支持向量机(Support Vector Machine,SVM)算法[14]是一个高效的有监督分类算法。其核心思想是在特征空间中建构最优分割超平面,使得分类器能够得到全局最优解。对于线性不可分的问题,SVM 使用各种核函数将低维空间线性不可分的数据映射到高维特征空间,达到线性可分的结果。二维数据的 SVM 分类效果如图 1 所示,图中最优超平面的方程为:

$$f(x)=w\cdot x+b=0 \tag{2}$$

其中,样本(x_i,y_i),$x_i\in R^d$,$y_i\in\{+1,-1\}$,$i=1,2,\cdots,N$,w 是权重向量,b 为尺度因子,权重向量和尺度因子决定了超平面的位置。

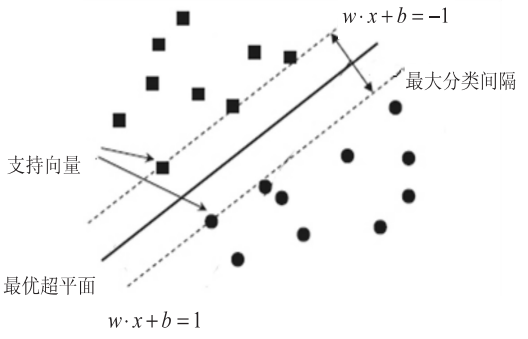

图 1　二维数据的 SVM 分类示意图

因 SVM 采用结构风险最小准则来训练分类器,能较好地处理样本特征非线性、维数高等问题,使其具

有精准的分类能力[15]。因此,提出基于 SVM 构建学位预警模型。

2 基于 Fisher 特征选择的学位预警模型

提出的基于 Fisher 特征选择的学位预警模型主要包括数据预处理、模型构建和学分预警三个部分,其算法流程如图 2 所示。考虑到学生成绩样本的特殊性,在数据预处理阶段要确保样本数据的规范化。要对学生成绩进行筛选,例如,删除选择人数较少的课程数据,删除选课较少的学生(如退学、休学等)成绩数据。此外,还要根据公式(3)对数据进行归一化处理。

$$\tilde{S} = \frac{S^j}{S_T^j} \tag{3}$$

其中,S^j、S_T^j 和 $\tilde{S}$ 分别表示第 j 门课成绩、第 j 门课的总分(一般是 100)和归一化后成绩。

在模型构建阶段首先利用 Fisher 算法进行特征选择;然后利用选择后的特征构建学位预警模型。在学位预警阶段,首先将测试样本依据特征选择规则得到新的测试样本;然后根据构建好的预警模型判断是否对学生进行学位预警。

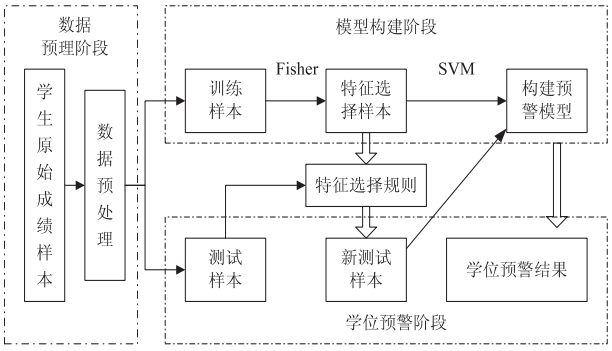

图 2 基于 Fisher 特征选择的学位预警模型流程

3 实 验

该文利用某高校 2018 级软件工程专业、化学专业、会计学专业、汉语言文学专业学生的真实成绩构建学位预警模型,通过统计大量的随机实验结果来验证用特征选择的方法构建预警模型的可行性和有效性。

3.1 实验数据

实验数据为某高校开设的包括工学类、理学类、管理学类、文学类在内的软件工程专业、化学专业、会计学专业、汉语言文学专业学生在 1、2、3 学期所获得的非学位课课程成绩和最终的平均学位绩点成绩,并分别对各专业学生成绩进行如下处理:

(1)删除选课人数过少(专业人数的 10%)的课程;

(2)将格式为“优秀”“良好”“中等”“及格”和“不及格”的等级成绩替换为“95”“85”“75”“65”和“55”百分制成绩;

(3)为了避免不同课程成绩间数量级对实验结果的影响,将学生分数成绩归一化到[0,1]区间内。

最终,利用处理后的 1、2、3 学期非学位课课程成绩和学位绩点成绩构建各专业的学位预警模型,数据情况如表 1 所示。

表 1 数据情况汇总

	软件工程	化学	会计	汉语言
学生人数	349	159	134	266
课程总数量	60	73	73	66
学位课总数量	15	16	14	16
实验用课程数量	21	25	32	27
需要给出预警的学生数量	18	8	7	13

3.2 评价指标

该文选用了有效、直观的错误率(ER)作为评价指标,其计算公式如公式(4)所示。此外,由于需要给出预警的样本仅占总样本的 5%,使得正负样本间存在严重的不平衡问题。因此,该文还用查全率(Recall)、查准率(Precision)、错误拒绝率(FRR)和错误接受率(FAR)共同作为评价指标。其中,错误率值越低说明预警模型的预警准确率越高;召回率又被称为查全率,表示需要给出预警的样本被正确给出预警的概率;查准率又被称为精准率,表示被预警模型给出预警的样本中真正需要做出预警的概率。错误拒绝率预警模型判断无需做出预警的样本中实际应该给出预警的概率;错误接受率表示无需给出预警的而被错误做出预警的概率。它们的计算公式分别为:

$$\mathrm{ER} = 1 - \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}} \tag{4}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{5}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{6}$$

$$\mathrm{FRR} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TN}} \tag{7}$$

$$\mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{8}$$

其中,TP 和 FN 分别表示预警模型对应该给出学位预警的样本正确做出预警(正确预测)的样本数量和没有做出预警(错误预测)的样本数量;FP 和 TN 分别表示预警模型对无需给出学位预警的样本错误给出预警(错误预测)的样本数量和没有做出预警(正确预测)的样本数量;TP + FN 即正样本的数量,TN + FP 即负样本的数量。

3.3 实验结果与分析

为了确保实验结果的稳定性和证明算法的有效

性,分别对每个专业进行 6 组实验。实验 1 到实验 6 分别利用不同的训练样本数量来构建预警模型,6 组实验中分别随机选择总样本的 40%、50%、60%、70%、80% 和 90% 作为训练集,其余样本数据作为测试集。每组实验都重复 100 次随机选样本,并将多次实验结果的平均值作为最终的统计结果。实验结果如表 2 至表 6 所示。

表 2　各专业学位预警错误率结果统计

训练样本占总样本比例/%	方法	会计学专业	化学专业	软件工程专业	汉语言文学专业
40	未特征选择	0.200 6	0.177 2	0.311 7	0.184 1
	特征选择	0.134 7	0.110 3	0.216 5	0.115 6
50	未特征选择	0.191 4	0.164 1	0.305 3	0.166 7
	特征选择	0.122 5	0.096 6	0.200 1	0.103 7
60	未特征选择	0.191 3	0.154 2	0.312 7	0.152 0
	特征选择	0.125 0	0.074 6	0.205 4	0.092 3
70	未特征选择	0.172 5	0.152 0	0.293 6	0.150 6
	特征选择	0.091 9	0.081 0	0.191 6	0.085 6
80	未特征选择	0.176 7	0.160 0	0.285 0	0.155 4
	特征选择	0.078 3	0.061 7	0.174 3	0.083 8
90	未特征选择	0.150 0	0.141 3	0.286 3	0.160 0
	特征选择	0.057 5	0.050 0	0.170 6	0.084 2

从表 2 中可以看出,随着训练样本的增加,各专业构建的学位预警模型的错误率普遍呈现下降趋势。其中,基于特征选择的预警模型明显具有更低的预警误差和更高的稳定性。

表 3　各专业学位预警查全率结果统计

训练样本占总样本比例/%	方法	会计学专业	化学专业	软件工程专业	汉语言文学专业
40	未特征选择	0.736 3	0.815 6	0.581 7	0.887 2
	特征选择	0.845 6	0.922 8	0.721 0	0.942 8
50	未特征选择	0.757 9	0.840 6	0.605 9	0.911 2
	特征选择	0.856 4	0.936 9	0.745 6	0.957 7
60	未特征选择	0.794 2	0.850 0	0.596 1	0.933 2
	特征选择	0.869 2	0.952 5	0.744 6	0.975 5
70	未特征选择	0.827 5	0.862 0	0.625 0	0.937 5
	特征选择	0.916 3	0.947 0	0.766 4	0.973 1
80	未特征选择	0.801 7	0.855 0	0.630 0	0.943 3
	特征选择	0.911 7	0.978 3	0.782 9	0.980 0
90	未特征选择	0.860 0	0.867 5	0.641 3	0.928 3
	特征选择	0.932 5	0.965 0	0.795 0	0.981 7

表 4　各专业学位预警查准率结果统计

训练样本占总样本比例/%	方法	会计学专业	化学专业	软件工程专业	汉语言文学专业
40	未特征选择	0.870 5	0.850 2	0.755 1	0.792 2
	特征选择	0.897 8	0.877 5	0.834 4	0.853 2
50	未特征选择	0.862 9	0.855 0	0.755 1	0.801 7
	特征选择	0.905 1	0.889 8	0.846 6	0.861 3
60	未特征选择	0.852 0	0.864 3	0.743 5	0.810 2
	特征选择	0.901 6	0.915 7	0.839 0	0.867 2

续表 4

训练样本占总样本比例/%	方法	会计学专业	化学专业	软件工程专业	汉语言文学专业
70	未特征选择	0.856 5	0.860 8	0.757 7	0.810 1
	特征选择	0.922 6	0.911 6	0.848 6	0.883 2
80	未特征选择	0.858 1	0.852 6	0.785 5	0.805 6
	特征选择	0.946 1	0.927 3	0.870 4	0.882 9
90	未特征选择	0.867 5	0.865 8	0.782 8	0.819 8
	特征选择	0.963 3	0.948 3	0.885 9	0.889 9

表 5 各专业学位预警错误拒绝率结果统计

训练样本占总样本比例/%	方法	会计学专业	化学专业	软件工程专业	汉语言文学专业
40	未特征选择	0.199 2	0.160 4	0.335 4	0.101 4
	特征选择	0.125 5	0.072 4	0.239 8	0.057 4
50	未特征选择	0.186 7	0.141 9	0.326 7	0.085 1
	特征选择	0.115 8	0.057 8	0.223 1	0.041 4
60	未特征选择	0.162 4	0.133 1	0.331 9	0.066 3
	特征选择	0.108 0	0.042 2	0.225 9	0.024 8
70	未特征选择	0.136 7	0.120 6	0.312 4	0.062 1
	特征选择	0.066 4	0.046 4	0.205 1	0.025 4
80	未特征选择	0.157 9	0.117 9	0.304 5	0.053 7
	特征选择	0.065 7	0.016 3	0.185 3	0.018 6
90	未特征选择	0.100 8	0.091 7	0.299 0	0.068 2
	特征选择	0.044 2	0.022 5	0.163 9	0.015 3

表 6 各专业学位预警错误接收率结果统计

训练样本占总样本比例/%	方法	会计学专业	化学专业	软件工程专业	汉语言文学专业
40	未特征选择	0.129 5	0.149 8	0.244 9	0.207 8
	特征选择	0.102 2	0.122 5	0.165 6	0.146 8
50	未特征选择	0.132 1	0.145 0	0.244 9	0.198 3
	特征选择	0.089 9	0.110 2	0.153 4	0.138 7
60	未特征选择	0.148 0	0.135 7	0.256 5	0.189 8
	特征选择	0.098 4	0.084 3	0.161 0	0.132 8
70	未特征选择	0.138 5	0.139 2	0.242 3	0.189 9
	特征选择	0.077 4	0.088 4	0.151 4	0.116 8
80	未特征选择	0.136 9	0.132 4	0.214 5	0.194 4
	特征选择	0.053 9	0.072 8	0.129 6	0.117 1
90	未特征选择	0.117 5	0.109 2	0.217 2	0.180 2
	特征选择	0.031 7	0.046 7	0.114 1	0.110 1

从表 3 到表 6 中也同样可以发现,各专业的成绩数据经过特征选择后构建的学位预警模型其查全率和查准率都高于没有进行特征选择的模型;而基于特征选择构建的学位预警模型的错误接受率和错误拒绝率则明显低于没有进行特征选择的模型。

综上所述,从表 2 至表 6 中的各项指标的统计结果显示,基于 Fisher 特征选择的学位预警模型具有更低的错误率和更高的稳定性。表明基于特征选择的方法可以选择出更有效的课程来构建更为精准的学位预警模型,其构建模型给出的预警结果更为可信,更有助于学生和教师及时地调整教学方式。

4 结束语

高校扩招政策的连年实施在为提升国内人口素质

的同时,也对现有高校教学管理模型带来了更高的挑战。探索学生学习的一般规律,挖掘和分析学生特征和成绩的关系,构建更为精准的学位预警模型可以更好地提高教学质量,对完善高校学位预警机制有重要的应用价值和现实意义,因此提出基于 Fisher 特征选择的学位预警模型方法。实验结果表明,构建的学位预警模型能更好地从现有成绩数据中挖掘有效信息,使预警模型具有更低的预警误差和更高的稳定性,能够更好地完成学位预警工作。但成绩预测或预警工作不仅会受到前期成绩的单一影响,还可能受到学习背景、行为习惯等因素的影响。因此,成绩预测、预警等工作仍是一个较为复杂的课题,在下一步的研究中将会利用特征选择方法充分挖掘学习背景、学习环境、行为习惯等更多因素,以构建更加精准有效的成绩预测或预警模型。

参考文献:

[1] 习近平. 高举中国特色社会主义伟大旗帜 为全面建设社会主义现代化国家而团结奋斗[N]. 人民日报,2022-10-26(001).

[2] 李晓璐. 提高教学质量适应继续教育未来发展变化[J]. 渤海大学学报:哲学社会科学版,2017,39(6):152-153.

[3] 唐亚伟,秦玉平. 基于数据挖掘的分类算法综述[J]. 渤海大学学报:自然科学版,2011,32(4):372-375.

[4] ZHANG Yupei, YUN Yue, AN Rui, et al. Educational data mining techniques for student performance prediction: method review and comparison analysis[J]. Frontiers in Psychology,2021,12:698490.

[5] 姚明海,李劲松,王　娜. 基于 BP 神经网络的高校学生成绩预测[J]. 吉林大学学报:信息科学版,2021,39(4):451-455.

[6] TOMASEVIC N, GVOZDENOVIC N, VRANES S. An overview and comparison of supervised data mining techniques for student exam performance prediction[J]. Computers & Education,2020,143(1):103676.1-103676.18.

[7] 刘晓雲,刘鸿雁,李劲松,等. 基于多元线性回归的学生成绩预测研究[J]. 计算机技术与发展,2022,32(3):203-208.

[8] CHEN Liyan, WANG Lihua, ZHOU Yuxin. Research on data mining combination model analysis and performance prediction based on students' behavior characteristics[J]. Mathematical Problems in Engineering,2022,2022(1):1-10.

[9] GITINABARD N, KHOSHNEVISAN F, LYNCH C F, et al. Your actions or your associates? predicting certification and dropout in MOOCs with behavioral and social features[C]// The 11th international conference on educational data mining (EDM 2018). New York: JEDM,2018:404-410.

[10] THAER T, ATEF Z, SANA A A, et al. An enhanced evolutionary student performance prediction model using whale optimization algorithm boosted with sine-cosine mechanism [J]. Applied Sciences,2021,11(21):10237-10245.

[11] 任　鸽,吴　猛,汗古丽·力提甫,等. 基于改进 Apriori 算法的高校课程预警规则库构建[J]. 计算机系统应用,2021,30(7):290-295.

[12] GAMIE E A, SEOUD S A E, SALAMA M, et al. Multi-dimensional analysis to predict students´ grades in higher education[J]. International Journal of Emerging Technologies in Learning,2019,14(2):4-15.

[13] 王　娜,李劲松,姚明海. 基于特征子集与特征区分度的生物认证方法[J]. 计算机技术与发展,2020,30(12):51-55.

[14] VAPNIK V. The nature of statistical learning theory[M]. New York: Springer,1995.

[15] ADMASU Y E, TEKLAY H A. Student performance prediction with optimum multilabel ensemble model[J]. Journal of Intelligent Systems,2021,30(1):511-523.