Creative Education Studies 创新教育研究, 2017, 5(2), 143-152
Published Online May 2017 in Hans. http://www.hanspub.org/journal/ces
https://doi.org/10.12677/ces.2017.52024

Hans汉斯

Predicting At-Risk Students Based on the Campus Card and Students' Basic Information

Yong Wang, Rui Xu

Ocean University of China, Qingdao Shandong
Email: markwy@126.com

Received: Apr. 29th, 2017; accepted: May 12th, 2017; published: May 23rd, 2017

Abstract

Accurate prediction of at-risk students in freshmen is extremely important to improve the graduation rate of a university. This paper explores the relationships between the campus card records and the performance of students. In addition, combining with the basic information of students and their previous exam scores, the paper proposes a method of predicting at-risk students based on machine learning and statistics. The experiment conducts on a dataset with 4.194 million items of 3680 freshmen of grade 2013 from the Ocean University of China, and the result shows that the recall rate is 52% and the precision is 77%, with good practical performance.

Keywords

Campus Card, Student's Basic Information, Machine Learning

基于校园一卡通与学生基本信息预测落后生

王　勇，许　蕊

中国海洋大学，山东　青岛
Email: markwy@126.com

收稿日期：2017年4月29日；录用日期：2017年5月12日；发布日期：2017年5月23日

摘　要

准确预测大一新生中的落后生对提高高校毕业率有重要的影响。本文研究了校园一卡通打卡记录和学生

文章引用: 王勇, 许蕊. 基于校园一卡通与学生基本信息预测落后生[J]. 创新教育研究, 2017, 5(2): 143-152.
https://doi.org/10.12677/ces.2017.52024

成绩之间的关系，并结合学生基本信息和前期成绩记录，提出一种基于机器学习与统计学的预测落后生的方法。在中国海洋大学2013级3680名新生共计419.4万条记录上进行了实验，结果显示所提出的方法查全率达到52%，查准率达到77%，具有较好的实用性。

关键词

校园一卡通，学生基本信息，机器学习

1. 引言

高等教育在我国已变得大众化，高考录取率从 20 世纪 80 年代的 7%上升到目前的 75%左右[1]，然而大一部分新生因挂科、留级常常不能正常毕业，使学生、家庭和学校承受了很大的压力，也对国家人才培养造成了一定的损失。为了减少此类现象的发生，需要一种方法可以提前识别出学业有问题的学生，以便及时对他们进行帮扶。挂科、留级的学生在日常生活与学习中有不同于正常学生的表现，这些表现可以从用餐、去图书馆等行为上有所体现。通过分析校园一卡通数据可能提取到这些影响学业表现的特征，设计出预测落后学生的方法。现有关于学生学业的研究通常是探讨学分制下基于学科间关系的学业警示，属于事后报警[2]。在学业预警方面，已有的研究工作主要有利用父母受教育程度、家庭收入等因素预测学生的成绩表现[3]；利用学生上网使用论坛的情况预测单科成绩[4]；利用在线学习系统的行为记录预测该学科的考试成绩[5] [6] [7] [8]；根据作业完成情况、课堂出勤、小测验成绩预测学生单科成绩[9]。相比而言，本论文研究面对全校新生，不分专业，不分学科，不使用与具体课程过程相关的特殊数据。目前尚没有与本文完全类似的研究。

2. 数据源介绍、预处理与特征提取

2.1. 数据源介绍

为了找到标识学生学习生活状态的数据，本文从教务处、网络中心、图书馆信息中心采集学生基本信息、成绩表与校园一卡通记录。通过沟通发现 2013 级为唯一一届有完整的大一时期全部记录的学生群体。因此，本研究基于 2013 级学生的数据。该研究工作的实验数据列表如表 1 所示。

Table 1. Data list of 3680 freshmen of grade 2013 in 18 colleges
表 1. 2013 级 18 个学院 3680 名学生大一期间数据列表

表名	共有记录数(条)
学生基本信息表	3680
成绩表	77,421
图书馆入馆记录	220,299
借还书记录	98,169
就餐刷卡记录	1,994,709

上述数据中，借还书记录只包括学号与借还书时间 2 项，不包括借书书目。就餐刷卡记录只包括学号、刷卡 POS 机号、刷卡时间 3 项，不包括刷卡金额。

2.2. 数据预处理

做好数据清洗工作是建立模型的基础。为了保护学生隐私，首先对所有记录中的学号转码，进行加密处理。删除各表中姓名、身份证号等个人隐私信息。对研究无意义的属性，降噪处理。针对不一致的属性值，进行标准化。各表经过预处理后，包括的信息如下：

- 学生基本信息表(学号，性别，出生日期，学院，系别，专业，专业类别，入学方式，生源地，考生类别)。
- 成绩表(学年学期，学号，课程代码，课程名称，平时成绩，期中成绩，技能成绩，考试成绩)。
- 图书馆入馆记录(学号，入馆打卡时间)。
- 借书记录(学号，借书打卡时间)。
- 早餐记录(学号，打卡时间)。
- 午餐记录(学号，打卡时间)。
- 晚餐记录(学号，打卡时间)。

将以上各表存到数据库文件中，数据库中的数据结构如下图 1 所示。

2.3. 特征空间的构造

每位学生都可以由从日常学习与生活的数据中提取的特征来描述，根据学生在各个特征变量的不同表现可以对学生进行预测分类。本文的特征提取从以下几方面入手：

第一、学生基本信息。

第二、图书馆入馆行为。

第三、图书馆借书行为。

第四、用餐行为。

第五、成绩信息。

2.3.1. 学生基本信息特征提取

学生基本信息中共有 10 个字段，经过分析，提取代表学生特征的以下字段：生源地、考生类别、入学方式、专业类别。

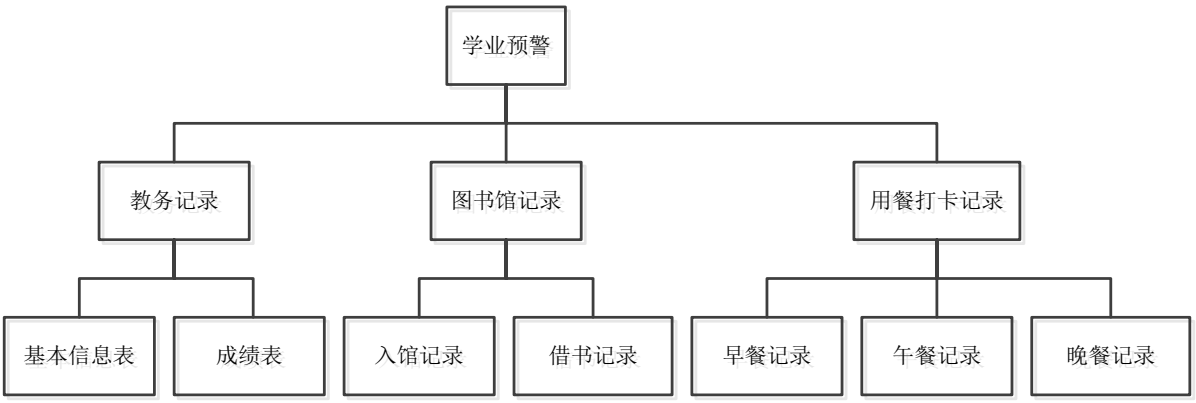

Figure 1. Data structure of the academic early warning research
图 1. 学业预警研究所用数据结构图

2.3.2. 图书馆入馆行为特征提取

从入馆记录提取。利用原始信息与生成的特征，构造以下 12 个特征变量：

1) totalCount：学生个人自 9 月到次年 5 月进入图书馆总次数。

2) intervMeanTime：学生个人入馆日期间隔天数的平均值。

3) earliestTime：学生个人自 9 月到次年 5 月入馆的最早时间。

4) earlyCount：“早”入馆时间计数，定义 8:00~9:30、11:30~13:00、17:00~18:00 这三个时间段为属于“早”的时间段。

5) meanTimeDuration：入馆次数比，表达式为：

$$\text{meanTimeDuration} = \text{入馆总次数}/\text{入馆总天数}。$$

6) stdvM：学生各月进图书馆次数的规律性，反映学习状态的稳定性。

7) meanM：月入馆次数平均值。

8) firstTerm、nextTerm：上学期开学月份与平常月份入馆次数的差值、下学期开学月份与平常月份入馆次数的差值，表达式为：

$$\text{firstTerm} = 9\text{ 月份入馆次数} - (10\text{ 月份入馆次数} + 11\text{ 月份入馆次数})/2$$

$$\text{nextTerm} = 3\text{ 月份入馆次数} - (4\text{ 月份入馆次数} + 5\text{ 月份入馆次数})/2。$$

9) fnCount：上下学期入馆次数差值，表达式为：

$$\text{fnCount} = \sum_{i=9}^{11} i\text{月份入馆次数} - \sum_{i=3}^{5} i\text{月份入馆次数}。$$

10) completeTerm：整学年开学月份与平常月份入馆次数差值，表达式为：

$$\text{completeTerm} = \text{firstTerm} + \text{nextTerm}。$$

11) sMultiMean：月份入馆标准差与平均值的乘积。

12) mtSquaDuration：平均在馆时长与入馆总次数乘积，表达式为：

$$\text{mtSquaDuration} = \text{meanTimeDuration} * \text{totalCount}。$$

2.3.3. 图书馆借书行为特征提取

从借书记录提取。将去图书馆每月累计借书册数作为特征变量，分别用 abook9-abook5 (从 2013 年 9 月到 2014 年 5 月各月累计借书册数)来表示。

2.3.4. 用餐行为特征提取

从就餐刷卡记录提取。每位学生并不是每周内的五天工作日首节都有课，在此做如下假设：大一学生在每周至少有三天的时间是首节有课的。因此将每周中早餐打卡时间最早的三天定为首节有课的情况。针对每周五天工作日的早餐时间，分为首节有课、首节无课两种情况，提取以下特征变量：

1) 首节有课－用餐时间平均值。

2) 首节有课－用餐次数缺少率。

3) 首节有课－用餐时间标准差。

4) 首节无课－用餐时间平均值。

5) 首节无课－用餐次数缺少率。

6) 首节无课－用餐时间标准差。

针对周末的早餐时间记录，提取以下特征变量：

7) 周末－用餐时间平均值。

8) 周末－用餐次数缺少率。

9) 周末－用餐时间标准差。

最后不考虑以上三种情况的差异，提取以下特征变量：

10) 总体－用餐时间平均值。

11) 总体－用餐次数缺少率。

12) 总体－用餐时间标准差。

针对午餐与晚餐记录，将各月累计用餐次数作为特征变量：alun8-alun5、adin8-adin5 (从 2013 年 8 月到 2014 年 5 月各月累计午餐、晚餐用餐次数)。

2.3.5. 成绩特征提取

从成绩记录提取。由于该预测模型的考察范围为全校学生在下学期期末考试的成绩表现，因此提取以下特征变量：上学期考试所得综合成绩、专业排名、所修学分以及是否受警示。

经过以上特征变量提取，构造 62 维的特征空间，明细如下图 2 所示。

早餐用餐记录	
PK	学号
	首节有课--用餐时间平均值 首节有课--用餐次数缺少率 首节有课--用餐时间标准差 首节无课--用餐时间平均值 首节无课--用餐次数缺少率 首节无课--用餐时间标准差 周末--用餐时间平均值 周末--用餐次数缺少率 周末--用餐时间标准差 总体--用餐时间平均值 总体--用餐次数缺少率 总体--用餐时间标准差

图书馆入馆记录	
PK	学号
	totalCount stdvM meanM sMutiMean firstTerm nextTerm completeTerm fnCount earliestTime earlyCount mtSquaDuration intervMeanTime

图书馆借书记录	
PK	学号
	abook9 abook10 abook11 abook12 abook1 abook2 abook3 abook4 abook5

成绩记录	
PK	学号
	上学期所得学分 上学期所得成绩 上学期所得专业排名 上学期是否受警示

午餐用餐记录	
PK	学号
	alun8 alun9 alun10 alun11 alun12 alun1 alun2 alun3 alun4 alun5

晚餐用餐记录	
PK	学号
	adin8 adin9 adin10 adin11 adin12 adin1 adin2 adin3 adin4 adin5

学生基本信息	
PK	学号
	生源地 考生类别 入学方式 专业类别

Figure 2. Characteristics extracted from each table
图 2. 各表提取的特征变量

3. 落后生分类方法

本节重点解决如何定义落后生的问题，即对学生进行标记，把学生总体分为正常学生与落后学生两类。

在统计留级生的过程中，注意到部分同学初次考试有多门数理化学科不及格，即使补考通过而没有被留级，但初试不及格仍能反映出该学生学习能力不够，处在学业危险的边缘。因此本文在学校留级规则基础上扩充，将符合以下条件之一的定义为落后生：

- 一学年中有 8 门次(含 8 门次)以上课程考核不合格。
- 一学年中修课取得的学分少于 16 学分。
- 数理化必修课初试与补考超过三门(含三门)以上考试不合格(比如若同一门课初试与补考都没通过，则计数为 2)。
- 下学期受到学业警示(学业警示的定义：两次及两次以下一学期内修课所得学分不足 12 学分)。

根据以上规则共筛选出落后生 386 人，给每位学生标上类标签(正常生标记为 zero，落后生标记为 one)。

4. 实验方法

在本研究中，利用多种机器学习算法建模预测。使用的算法如下[10] [11]：

- NaiveBayes：基于贝叶斯定理和属性之间相互独立假设的分类方法。
- J48：基于 C4.5 算法实现的一种决策树。
- AdaBoostM1：依据每次训练时样本是否被正确分类，以及上次分类的总精度，来确定每个样本的权重的迭代算法。
- Bagging：在原始数据集上有放回的抽样，构成 N 个新训练集来训练分类器。
- IBK：将训练集中最相似的 K 个数据中出现次数最多的分类标签作为测试集中新数据的分类类别。
- Randomtree：经随机过程建立的决策树，不包含属性选择过程。
- RandomForest：由通过随机过程建立的多个决策树构成的森林，决策树之间是相互独立的。
- REPTree：采取降低错误率剪枝的策略。
- Logistic [12]：一种广义的线性回归分析模型，由多个预测变量计算分类变量的概率。
- SimpleLogistic [13]：与 Logistic 类似，算法结构更为简化。

由于规定的落后生在整体数据中占比 10%左右，而实验的关键在于能否正确地识别落后生，单纯地使用总体精度无法体现出这一点，因此引入查准率、查全率、F-Measure、ROC 曲线以及 AUC 面积等性能评价指标。若分类器的混合矩阵如下表 2 所示。

则总体精度为：

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total number of students}}$$

Table 2. Confusion matrix statistics table
表 2. 混合矩阵统计表

	被分类器分为正常生	被分类器分为落后生
实际为正常生	TP	FN
实际为落后生	FP	TN

分类为落后生的查准率为：

$$P = \frac{TN}{FN + TN}$$

分类为落后生的查全率为：

$$R = \frac{TN}{FP + TN}$$

分类为落后生的 F-Measure 指标为：

$$F = \frac{2PR}{P + R}$$

F 值越高，说明查准率与查全率都较高，分类性能越好。

ROC 曲线指以真正类率 TPR 为 Y 轴，假正类率 FPR 为 X 轴绘制的曲线，其中：

$$TPR = \frac{TN}{FP + TN}$$

$$FPR = \frac{FN}{FN + TP}$$

该曲线与 X 轴围成的面积为 AUC 面积，面积值越大，分类性能越好。

5. 实验结果及分析

5.1. 初始结果分析

结合特征空间与学生类标签构建训练集。借助 WEKA 平台分别利用上节提到的分类器训练模型，对学生的分类做预测。为了避免过拟合的问题，每个分类模型都是采用十折交叉验证，并选用默认的参数设置。各分类器预测结果如下表 3 所示。

对总体精度而言，除 NaiveBayes 分类器有较大误差外，其它分类器都有较高的精度。但本文目的在于对落后生的分类预测，接下来比较分类为落后生的性能评价指标。查全率与查准率都较高的分类器有：

Table 3. Summaries of the predicted results of each classifier
表 3. 各分类器预测结果汇总表

分类器	总体精度	落后生		正常生		落后生 F-Measure	落后生 ROC Area
		查全率	查准率	查全率	查准率		
NaiveBayes	67.53%	74.90%	20.80%	66.70%	95.80%	0.408	0.773
J48	92.20%	47.40%	68.50%	97.40%	94.10%	0.56	0.796
meta-----AdBoostM1	90.63%	69.40%	54.10%	93.10%	96.30%	0.608	0.891
meta-----Bagging	92.26%	47.20%	69.20%	97.50%	94%	0.561	0.896
lazy----IBK	85.87%	24.10%	29.10%	93.10%	91.30%	0.263	0.586
Randomtree	85.87%	31.30%	32.20%	92.30%	92%	0.318	0.634
RandomForest	91.06%	17.40%	87%	99.70%	91.10%	0.289	0.904
SimpleLogistic	93.18%	50%	76.90%	98.20%	94.40%	0.606	0.926
Logistic	92.47%	50.30%	69.50%	97.40%	94.40%	0.583	0.904
REPTree	91.98%	52.10%	64.60%	96.70%	94.50%	0.577	0.848

AdBoostM1 分类器、SimpleLogistic 分类器、Logistic 分类器与 REPTree 分类器。F-Measure 值较大的有 AdBoostM1 与 SimpleLogistic，说明这两个分类器的查全率查准率都表现较好。这两个分类器的 ROC 曲线如下图 3、图 4 所示。

由于 ROC 曲线是以假正率 FPR 为 X 轴，真正率 TPR 为 Y 轴，一个好的分类器追求低 FPR 高 TPR，因此 ROC 曲线越靠近左上角，分类性能越好。因此 SimpleLogistic 分类器训练的模型最优，分类结果最好。

5.2. 分类器的优化调整

利用 WEKA 中 CVParameterSelection 算法对 Simplelogistic 分类器的参数进行优化。将参数 numBoostinglterations 的优化范围设置为[0 100]，参数 maxBoostinglterations 的优化范围设置为[300 800]，参数 heuristicStop 的优化范围设置为[30 40]，参数 WeightTrimBeta 的优化范围设置为[0 0.5]。优化结果为图 5。

即 numBoostinglterations = 36，maxBoostinglterations = 300，heuristicStop = 30，WeightTrimBeta = 0。据此调整 SimpleLogistic 分类器的参数，得出分类结果。优化前后 SimpleLogistic 分类器的分类结果比较如下表 4。

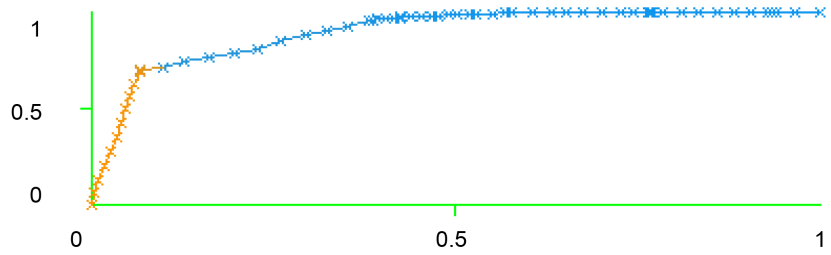

Figure 3. ROC curve of at-risk students classified by AdBoostM1
图 3. AdBoostM1 分类为落后生的 ROC 曲线

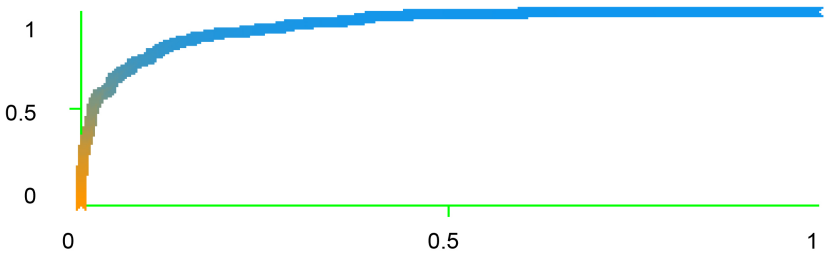

Figure 4. ROC curve of at-risk students classified by SimpleLogistic
图 4. SimpleLogistic 分类为落后生的 ROC 曲线

```
Cross-validated Parameter selection.
Classifier: weka.classifiers.functions.SimpleLogistic
Cross-validation Parameter: '-M' ranged from 300.0 to 700.0 with 9.0 steps
Cross-validation Parameter: '-H' ranged from 30.0 to 70.0 with 9.0 steps
Cross-validation Parameter: '-W' ranged from 0.0 to 0.5 with 11.0 steps
Cross-validation Parameter: '-I' ranged from 30.0 to 40.0 with 11.0 steps
Classifier Options: -M 300 –H 30 –W 0 –I 36
```

Figure 5. Parameter optimization results
图 5. 参数优化结果

Table 4. Comparison on the results of SimpleLogistic classifier before and after optimization
表 4. Simple Logistic 分类器优化前后结果比较

分类器	总体精度	落后生		正常生		落后生 F-Measure	落后生 ROC Area
		查全率	查准率	查全率	查准率		
优化前 SimpleLogistic	93.18%	50%	76.90%	98.20%	94.40%	0.606	0.926
优化后 SimpleLogistic	93.40%	52.30%	77.40%	98.20%	94.60%	0.624	0.926

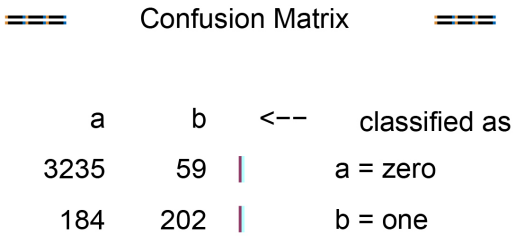

Figure 6. Confusion matrix of optimized SimpleLogistic classifier
图 6. 优化后 SimpleLogistic 分类器的混合矩阵

由表可知，F-Measure 值有很明显的提高，查全率与查准率都有所提升。优化后 SimpleLogistic 分类器的混合矩阵如图 6。

此分类器预测出 261 名学生为落后生，其中 202 名学生为真正落后生，漏掉 184 人，其中 59 个正常学生被误判为落后生。

综上，通过综合分析各分类器的分类结果，选出效果最好的 SimpleLogistic 分类器，通过参数优化调整，得到针对落后生的优化预测模型：参数为 numBoostinglterations = 36，maxBoostinglterations = 300，heuristicStop = 30，WeightTrimBeta = 0 的 SimpleLogistic 模型，预测落后生的查全率为 52.3%，查准率为 77.4%，精度较高，具有实用价值。

6. 结束语

本文提出一种针对大一新生的基于校园一卡通数据及学生基本信息的落后生预测分类模型。该模型使用的数据可方便地大规模自动采集，可对全校新生中的落后生进行预测和分类，而不局限于一门课或一个专业。实验结果显示，该模型最佳状态下查全率可达 52%，查准率达 77%，表示出较好的性能，有较高的适用性与可行性。落后学生预警模型的提出，有助于准确地识别出学业困难的学生，学校可据此有针对性地采取帮扶措施，从而减少留级的人数，对高校提高人才培养质量起到了良好的促进作用。后续研究将继续探索有助于分类的组合特征，如学生宿舍距离食堂的路程，去图书馆打卡数据则考虑距考试周的远近等因素，提高预测效率与精度。

参考文献 (References)

[1] 袁安府, 张娜, 沈海霞. 大学生学业预警评价指标体系的构建与应用研究[J]. 黑龙江高教研究, 2014(3): 79-83.

[2] Wu, H.F., Cheng, Y.S. and Hu, X.G. (2012) Model Design of Achievement Pre-Warning in High Education Based on Data Mining. Springer, Berlin Heidelberg, 501-506.

[3] Ramaswami, M. and Bhaskaran, R. (2010) A CHAID Based Performance Prediction Model in Educational Data Mining. *International Journal of Computer Science Issues*, **7**, 10-18.

[4] Romero, C., Pez, M.I., Luna, J.M., *et al*. (2013) Predicting Students' Final Performance from Participation in On-Line Discussion Forums. *Computers & Education*, **68**, 458-472.

[5] Jovanovic, M., Vukicevic, M., Milovanovic, M., *et al*. (2012) Using Data Mining on Student Behavior and Cognitive Style Data for Improving e-Learning Systems: A Case Study. *International Journal of Computational Intelligence Systems*, **5**, 597-610. https://doi.org/10.1080/18756891.2012.696923

[6] Şen, B., Uçar, E. and Delen, D. (2012) Predicting and Analyzing Secondary Education Placement-Test Scores: A Data Mining Approach. *Expert Systems with Applications*, **39**, 9468-9476.

[7] Minaeibidgoli, B., Kashy, D.A., Kortemeyer, G., *et al*. (2003) Predicting Student Performance: An Application of Data Mining Methods with an Educational Web-Based System. *Frontiers in Education*, Westminster, 5-8 November 2003, T2A-13.

[8] Rovai, A.P. (2003) In Search of Higher Persistence Rates in Distance Education Online Programs. *Internet & Higher Education*, **6**, 1-16.

[9] Marbouti, F., Diefes-Dux, H.A. and Madhavan, K. (2016) Models for Early Prediction of At-Risk Students in a Course Using Standards-Based Grading. *Computers & Education*, **103**, 1-15.

[10] Pang, N., Michaelsteinbach, V. 数据挖掘导论: 完整版[M]. 北京: 人民邮电出版社, 2011: 89-451.

[11] Kubat, M. (2015) An Introduction to Machine Learning. Springer International Publishing, Berlin, 19-274. https://doi.org/10.1007/978-3-319-20010-1_2

[12] Edition, S. (2016) Applied Logistic Regression Analysis. *Technometrics*, **38**, 184-186.

[13] Trappey, C.V. and Wu, H.Y. (2008) An Evaluation of the Time-Varying Extended Logistic, Simple Logistic, and Gompertz Models for Forecasting Short Product Lifecycles. *Advanced Engineering Informatics*, **22**, 421-430.