Advances in Social Sciences 社会科学前沿, 2018, 7(8), 1149-1156
Published Online August 2018 in Hans. http://www.hanspub.org/journal/ass
https://doi.org/10.12677/ass.2018.78169

Hans汉斯

Establishment of University Students' Academic Early Warning Model Based on Discriminant Analysis

Jinhua Ye, Xiaoqiu Yu, Xiaojing Zhou, Debao Gao, Di Lang, Hui Qin

College of Science, Heilongjiang Bayi Agricultural University, Daqing Heilongjiang
Email: zhouxiaojing7924@126.com

Received: Jul. 18th, 2018; accepted: Aug. 1st, 2018; published: Aug. 8th, 2018

Abstract

Students tend to slack off after entering university; therefore, it is very important to study the early warning model for college students to graduate successfully and find a job. In this article, through combing the literature and our school students' actual situation, academic early warning index system is established on five elements: the daily study time each week; student course selection number; number of late, leaving early and truancy in one term; make-up exam course credits; semester grades. Based on distance discriminant method, Fisher discriminant method and Bayes discriminant method of discriminant analysis method, three academic warning models are established, and the test results show that distance discriminant method and Bayes discriminant method have an accuracy of 100%. Finally, the three methods are compared and analyzed.

Keywords

Academic Warning, Discriminant Analysis, SPSS

基于判别分析法的大学生学业预警模型建立

野金花，于晓秋，周晓晶，高德宝，朗　迪，秦　辉

黑龙江八一农垦大学理学院，黑龙江　大庆
Email: zhouxiaojing7924@126.com

收稿日期：2018年7月18日；录用日期：2018年8月1日；发布日期：2018年8月8日

摘　要

学生进入大学以后，很容易出现懈怠学业的现象，因此研究预警模型对大学生能够顺利毕业以及求职十

文章引用: 野金花, 于晓秋, 周晓晶, 高德宝, 朗迪, 秦辉. 基于判别分析法的大学生学业预警模型建立[J]. 社会科学前沿, 2018, 7(8): 1149-1156. DOI: 10.12677/ass.2018.78169

分重要。本文通过梳理文献及本校学生实际情况，确定以每周的日常学习时间，学生选课数目，学期迟到、早退、逃课次数，学期补考课程学分，学期成绩总和五个要素建立学业预警指标体系，再基于判别分析法中的距离判别法、Fisher判别法以及贝叶斯判别法建立3个学业预警模型，并对结果进行检验，其中距离判别法和贝叶斯判别法的准确率达100%。最后进行了三个方法的比较分析。

关键词

学业预警，判别分析，SPSS

1. 引言

学业预警是高校在教育管理实施过程中，对学生在校期间已经发生或者即将发生的学习方面的问题和学业上的困难进行危机干预，并有针对性地采取防范和补救措施，通过学校、学生及其家长三者之间的沟通与协作，帮助学生克服学业困难，使学生重返正常的学习、工作与生活常态并顺利完成学业的一种信息沟通和危机干预制度[1]。

普渡大学的 Course Signals (课程信号)系统是数据量化和监测学生学习状态的系统。对课程 GPA 成绩、等级考试成绩、排名、学术经历及学生登录课程网站频率等进行了分析。普渡大学的研究者发现，那些曾经被亮黄灯，即处在中度学业失败危险的学生，收到预警邮件后会在课堂上表现得更好。而那些直接被亮红灯，即处于高危群体的学生，即便收到了预警邮件，他们在课堂表现上也不会有太大改观。由此也可以看出，早期预警对成绩不佳的学生顺利完成学业十分重要。国外有预测系统性银行危机的模型，采用了信号方法和 Logit 模型的方法进行研究[2]。为了防止钢铁企业事故发生，利用组合加权和灰色预测模型对钢铁企业事故预警系统进行了探讨[3]。BP 神经网络模型应用于寿险偿付能力预警体系，利用改进的 BP 网络预测寿险公司的偿付能力[4]。

2007 年广西师范学院的陈钦华按照以学生为本、过程管理和目标管理的统一、全面性、便利性原则，设计了学业预警机制[5]。深圳大学管理学院华金秋介绍了我国台湾高校学业预警制度背景、规定和特点，然后就建立大陆高校学习预警机制进行了探讨[6]。2009 年苏州科技大学的丁福兴、蔡熹芸、翁丽祥针对苏州科技大学的现实情况，以该大学的学习困难群体为研究对象，选取 2007 年 9 月和 2008 年 9 月的学籍处理数据进行比较，得出的研究结论认为：实施学生学业预警措施在降低学籍处理率方面的绩效是显著的[7]。2010 年张海舰等提出了以学生的平均绩点来作为界限标准的建议，并就学业预警的工作程序和运行进行了论述，但是如何公平评价学生的问题没有得到解决[8]。2011 年范冰通过构建高职院校学业预警机制，提出应高度重视、加强防范初期的准备不足，应利用教务管理系统建立学生预警数据库，应加强辅导员的监管工作力度，应加强落实后期等措施[9]。关于学业预警模型的建立，2015 年郑俊玲基于核主成分分析建立了学业预警模型[10]。

综上所述，国外各高校已经具备了相对完善、严格、科学的学业辅导制度，而我国的学业预警机制还没有深入实施，大部分停留在定性分析的研究上，因此，本文针对数据难以人工处理、复杂多样的特点，选取判别分析方法建立学业预警模型，对大量数据进行量化分析处理，对需要预警的指标进行分析和计算，利用现有数据总结得出判别分析方法的出错率低，为预警的确定和预警等级的划分提供直观的

依据，有利于采用定量的方式为学业预警工作提供高效率高质量的保障，相关部门可依据分析完成后的数据商讨如何指导学业预警工作的展开，进行接下来的学业帮扶工作。

2. 大学生学业预警主要指标的选取

学业预警的前提是建立一套科学的预警评价指标体系，它决定着预警对象的准确性和预警程度的精确性，还直接影响着后续帮扶与激励的效果。指标的选取要能够反映评价对象的各方面状况，而且针对指标采集的数据应该容易收集、计算或估计的。

根据指标体系建立的原则以及文献资料，再结合本校的实际预警情况，本文选取了日常学习时间/周，学生选课数目，学期迟到、早退、逃课次数，学期补考课程学分，学期成绩总和共五个指标构成学业预警模型的指标体系。

数据采用黑龙江八一农垦大学的学业预警评价体系：根据学生的实际情况及所产生影响的程度，发出不同等级的预警信息，分为口头预警、蓝色预警、黄色预警、橙色预警、红色预警五个级别。此次选取了黑龙江八一农垦大学某学期获得学业预警 40 位同学的相关数据来进行模型的建立，采集的指标有日常学习时间，选课数目，学期迟到、早退、旷课次数，补考课程学分，学期六门成绩总和这五个指标。其中，las、xtg、erd、ers 四位同学作为待判样本。借助的数学分析软件为 SPSS。

3. 判别分析模型建立

判别分析是多元统计中用来判别样本类型的统计分析方法，在已知的主题下用某种方法分成几类情况，判断新样本属于什么类型的多变量统计分析方法。通常使用判别分析方法来测量新样本与每个已知样本的接近度，即判别函数，同时规定一种判别准则来判定新样本的归属。判别准则是衡量新样本与已知群体接近度的理论基础和方法标准。常用的有距离判别准则、Fisher 判别准则、贝叶斯判别准则等。

3.1. 距离判别模型的建立

基本思想[11]：首先根据已知分类的数据，计算每个类别的重心，即分组(类别)的平均值，距离判别准则是对任何新样本的观测值，如果它与第 i 类的重心距离最近则认为来自第 i 类。

1) 检验变量在各分类上均值是否有显著差异

SPSS 软件中给出了表 1 单变量方差分析的结果，分别检验了日常学习时间，选课数目，学期迟到、早退、旷课次数，补考课程学分，学期六门成绩总和这五个变量在口头预警、蓝色预警、黄色预警、橙色预警、红色预警五类中的均值是否相等。原假设是各类均值相等，备择假设是不等，如果接受原假设，说明利用这五个变量进行判别分析没有意义。从该表中可以看出计算出的显著性除“选课数目”无法计算以外，其余都小于 0.01，故可以参与判别分析，变量“选课数目”是不适合的。

Table 1. Tests of equality of group means
表 1. 群组平均值的等式检定

	Wilks' Lambda (λ)	F	$df1$	$df2$	显著性
周学习时间(小时)	0.059	122.631	4	31	0.000
选课数目	.[a]				
迟到、早退、逃课次数	0.014	543.870	4	31	0.000
补考课程学分	0.039	189.474	4	31	0.000
学期六门成绩总和	0.244	24.043	4	31	0.000

[a]无法计算，因为在每一个群组中此变量都是常数。

2) 建立判别函数式

SPSS 软件中的分类函数系数表给出了判别函数的系数和常数项见表 2，从该表中得到距离判别中的判别函数为：

$$y_1 = 5.374x_1 + 1.311x_3 - 2.044x_4 + 0.620x_5 - 151.871$$

$$y_2 = 0.858x_1 + 8.851x_3 + 1.854x_4 + 0.546x_5 - 133.833$$

$$y_3 = -2.836x_1 + 17.412x_3 + 4.578x_4 + 0.546x_5 - 219.447$$

$$y_4 = -5.297x_1 + 23.532x_3 + 7.168x_4 + 0.493x_5 - 321.921$$

$$y_5 = -9.156x_1 + 31.245x_3 + 9.484x_4 + 0.500x_5 - 507.950$$

利用 SPSS 的 Compute 功能将所有样本的变量值带入到判别函数式，可以得到所有样本的判别函数式 y_1，y_2，y_3，y_4，y_5，其中 las 的判别函数值分别为 11.57，285.73，470.62，562.72，618.75；xtg 的判别函数值分别为 63.68，191.05，236.05，220.99，149.94；erd 的判别函数值分别为 104.85，143.92，107.41，28.52，−116.39；ers 的判别函数值分别为 115.13，148.93，106.59，23.00，−127.15。

因此将 las 归为第五类；将 xtg 归为第三类；erd 归为第二类；将 ers 归为第二类。这正是表 2 给出的用此判别函数对待样品判别的结果。

3) 分类结果

从表 3 可以看出用此判别函数对所有样品判别的结果和误判概率以及用交叉验证法判别结果和误判概率。可见有 100%的正确率。

3.2. Fisher 判别法模型的建立

基本思想[12]：将 m 组 n 维的数据投影到某一个方向，n 代表所选取的指标个数，投影就是将 n 维空间的样本投影到 m 空间中。在 n 维空间中找到组间距离最大的线性组合作为第一 Fisher 判别函数，以此类推得到第二、第三 Fisher 判别函数。

1) 建立判别函数

从 SPSS 中的典型区别函数系数表中可以得到未标准化典型判别函数的系数向量。根据此系数向量可以写出 4 个未标准化的典型判别函数：

$$y_6 = -0.455x_1 + 0.931x_3 + 0.372x_4 - 0.004x_5 - 2.302$$

$$y_7 = 0.156x_1 + 0.698x_3 - 0.387x_4 + 0.027x_5 - 11.374$$

$$y_8 = 0.096x_1 + 0.208x_3 + 0.170x_4 - 0.025x_5 - 1.790$$

$$y_9 = -0.022x_1 - 0.220x_3 + 0.329x_4 - 0.028x_5 - 11.869$$

Table 2. Classification function coefficients
表 2. 分类函数系数

	口头预警	蓝色预警	黄色预警	橙色预警	红色预警
周学习时间(小时)	5.374	0.858	−2.836	−5.297	−9.156
迟到、早退、逃课次数	1.311	8.851	17.412	23.532	31.245
补考课程学分	−2.044	1.854	4.578	7.168	9.484
学期六门成绩总和	0.620	0.546	0.546	0.493	0.500
(常数)	−151.871	−133.883	−219.447	−321.921	−507.950

Table 3. Classification results[a,c]
表 3. 分类结果 [a,c]

			口头预警	蓝色预警	黄色预警	橙色预警	红色预警	
原始	计数	口头预警	23	0	0	0	0	23
		蓝色预警	0	4	0	0	0	4
		黄色预警	0	0	3	0	0	3
		橙色预警	0	0	0	3	0	3
		红色预警	0	0	0	0	3	3
		未分组的观察值	0	2	1	0	1	4
	%	口头预警	100.0	0.0	0.0	0.0	0.0	100.0
		蓝色预警	0.0	100.0	0.0	0.0	0.0	100.0
		黄色预警	0.0	0.0	100.0	0.0	0.0	100.0
		橙色预警	0.0	0.0	0.0	100.0	0.0	100.0
		红色预警	0.0	0.0	0.0	0.0	100.0	100.0
		未分组的观察值	0.0	50.0	25.0	0.0	25.0	100.0
交叉验证[b]	计数	口头预警	23	0	0	0	0	23
		蓝色预警	0	4	0	0	0	4
		黄色预警	0	0	3	0	0	3
		橙色预警	0	0	0	3	0	3
		红色预警	0	0	0	0	3	3
	%	口头预警	100.0	0.0	0.0	0.0	0.0	100.0
		蓝色预警	0.0	100.0	0.0	0.0	0.0	100.0
		黄色预警	0.0	0.0	100.0	0.0	0.0	100.0
		橙色预警	0.0	0.0	0.0	100.0	0.0	100.0
		红色预警	0.0	0.0	0.0	0.0	100.0	100.0

[a]100.0%个原始分组观察值已正确地分类；[b] 仅会针对分析中的那些观察值进行交叉验证。在交叉验证中，每一个观察值都会依据从该观察值之外的所有观察值衍生的函数进行分类；[c]100.0%个交叉验证已分组观察值已正确地分类。

利用 SPSS 的 Compute 功能将所有样本的变量带入到以上 4 个判别函数中，可以得到所有样本的判别函数值(表 4)。待判样品的函数值分别为(27.94, 2.44, −0.02, 16.56)，(11.58, 0.25, −4.30, −18.41)，(1.90, 0.05, −7.20, −20.30)，(1.26, −0.44, −7.79, −20.36)。

2) 函数结果显著性的 Wilks' Lambda 检验

表 5 给出了典型判别函数的有效性检验。Wilks' Lambda 范围在 0 至 1 之间，值接近 0 表示组均值不同，值接近 1 表示组均值在统计意义上没有不同。由于 Wilks' Lambda 统计量的分布表不易找到，一般化为卡方统计量。*df* 自由度是用于计算显著性水平的自由度。sig 显著性水平是零假设被拒绝的概率，即拒绝零假设时犯第一类错误的概率。从表中可以看出判别函数显著性检验的 sig 值分别为 0.000，0.518，0.930，0.796。除了第一个以外都大于 0.05，故只有第一个判别函数显著，能够用来判别样品的归属。

3.3. 贝叶斯判别法模型的建立

贝叶斯的统计思想[13]是假设对研究的对象已有一定程度的认识，常用先验概率分布来描述这种认

识，然后抽取一个样本，用样本来修正已有的认识(先验概率分布)，得到后验概率分布。各种统计推断都通过后验概率分布来进行，将贝叶斯思想用于判别分析就得到贝叶斯判别法。

1) 以口头预警、蓝色预警、黄色预警、橙色预警、红色预警所含学生的频率作为各类的先验概率，即

$$p_1=\frac{23}{36}=0.639,\quad p_2=\frac{4}{36}=0.111,\quad p_3=\frac{3}{36}=0.083,\quad p_4=\frac{3}{36}=0.083,\quad p_5=\frac{3}{36}=0.083$$

同表 6 结果。

2) 建立贝叶斯判别函数

表 7 给出了贝叶斯判别中式的判别函数的系数和常数项。

从该表可得贝叶斯判别中的 5 个判别函数为

$$y_{10}=5.374x_1+1.311x_3-2.044x_4+0.620x_5-150.709$$

$$y_{11}=0.858x_1+8.851x_3+1.854x_4+0.546x_5-134.471$$

$$y_{12}=-2.836x_1+17.412x_3+4.578x_4+0.546x_5-220.323$$

$$y_{13}=-5.297x_1+23.532x_3+7.168x_4+0.493x_5-322.796$$

$$y_{14}=-9.156x_1+31.245x_3+9.484x_4+0.500x_5-508.826$$

Table 4. Canonical discriminant function coefficients
表 4. 典型区别函数系数

	1	2	3	4
周学习时间(小时)	−0.455	0.156	0.906	−0.022
迟到、早退、逃课次数	0.931	0.698	0.208	−0.220
补考课程学分	0.372	−0.387	0.170	0.329
学期六门成绩总和	−0.004	0.027	−0.025	0.028
(常数)	−2.302	−11.374	−1.790	−11.869

非标准化系数。

Table 5. Wilks' Lambda (λ)
表 5. Wilks' Lambda (λ)

函数的检定	Wilks' Lambda (λ)	卡方	*df*	显著性
1 至 4	0.006	158.060	16	0.000
2 至 4	0.765	8.165	9	0.518
3 至 4	0.972	0.865	4	0.930
4	0.998	0.067	1	0.796

Table 6. Prior probabilities for groups
表 6. 群组的事前机率

预警级别	在前	分析中使用的观察值未加权	分析中使用的观察值加权
口头预警	0.639	23	23.000
蓝色预警	0.111	4	4.000
黄色预警	0.083	3	3.000
橙色预警	0.083	3	3.000
红色预警	0.083	3	3.000
总计	1.000	36	36.000

3) 判别函数准确性结果检验

利用回代法和交互验证法对判别结果的准确性进行验证。从表 8 可知，准确率为 100%。

3.4. 学业预警模型比较

距离判别模型中的指标体系“选课数目”对分析无意义。但是预测方法得到的结果与实际情况一致，未出现误判现象，能够为学业预警提供技术支持。

Table 7. Classification function coefficients
表 7. 分类函数系数

	口头预警	蓝色预警	黄色预警	橙色预警	红色预警
周学习时间(小时)	5.374	0.858	−2.836	−5.297	−9.156
迟到、早退、逃课次数	1.311	8.851	17.412	23.532	31.245
补考课程学分	−2.044	1.854	4.578	7.168	9.484
学期六门成绩总和	0.620	0.546	0.546	0.493	0.500
(常数)	−150.709	−134.471	−220.323	−322.796	−508.826

Table 8. Classification results[a,c]
表 8. 分类结果 [a,c]

		预警级别	口头预警	蓝色预警	黄色预警	橙色预警	红色预警	总计
原始[a]	计数%	口头预警	23	0	0	0	0	23
		蓝色预警	0	4	0	0	0	4
		黄色预警	0	0	3	0	0	3
		橙色预警	0	0	0	3	0	3
		红色预警	0	0	0	0	3	3
		未分组的观察值	0	2	1	0	1	4
		口头预警	100.0	0.0	0.0	0.0	0.0	100.0
		蓝色预警	0.0	100.0	0.0	0.0	0.0	100.0
		黄色预警	0.0	0.0	100.0	0.0	0.0	100.0
		橙色预警	0.0	0.0	0.0	100.0	0.0	100.0
		红色预警	0.0	0.0	0.0	0.0	100.0	100.0
		未分组的观察值	0.0	50.0	25.0	0.0	25.0	100.0
交叉验证[b]	计数%	口头预警	23	0	0	0	0	23
		蓝色预警	0	4	0	0	0	4
		黄色预警	0	0	3	0	0	3
		红色预警	0	0	0	0	3	3
		口头预警	100.0	0.0	0.0	0.0	0.0	100.0
		蓝色预警	0.0	100.0	0.0	0.0	0.0	100.0
		黄色预警	0.0	0.0	100.0	0.0	0.0	100.0
		橙色预警	0.0	0.0	0.0	100.0	0.0	100.0
		红色预警	0.0	0.0	0.0	0.0	100.0	100.0

[a]100.0%个原始分组观察值已正确地分类；[b] 仅会针对分析中的那些观察值进行交叉验证。在交叉验证中，每一个观察值都会依据从该观察值之外的所有观察值衍生的函数进行分类；[c]100.0%个交叉验证已分组观察值已正确地分类。

Fisher 判别模型中只有 1 个判别函数显著，能用来判别样品的归属，而且结果与距离判别结果不一致，得到的结果不准确。

贝叶斯判别方法中，在各类预警所含学生的频率作为各类的先验概率的情况下，所得判别分析的结果误判率为 0。

4. 结论

本文就学业预警工作中的情况加以判别分析方法的应用，使学业预警工作由定性转变为定性定量相结合；确定了建立学业预警模型的指标，建立了距离判别分析模型、Fisher 判别分析模型、贝叶斯判别分析模型对学业预警情况进行了预测，其中距离判别法、贝叶斯判别法的模型，这两种模型的误判率为 0，预测精度高，可靠性强，同时为学业预警等级评价中提供可供参考的评价指标；该模型由于数据、条件的限制，所以还存在很多不足之处，如果能够尝试更多条件，相信会对学业预警机制提供指导意义。

基金项目

2017 年黑龙江省高等教育教学改革一般项目(编号：SJGY20170441)；2018 年黑龙江省高等教育科学规划重点课题(GBB1318087)；2017 年大庆市哲学社会科学规划研究项目(编号：DSGB2018110)；2017 年黑龙江省大学生创新创业重点项目(编号：2017LXY04)。

参考文献

[1] 邹慧. 独立学院“学困生”学业预警机制探析[D]: [硕士学位论文]. 南京: 南京师范大学, 2011.

[2] Asanović, Ž. (2017) Predicting Systemic Banking Crises Using Early Warning Models: The Case of Montenegro. *Journal of Central Banking Theory and Practice*, **6**, 157-182. https://doi.org/10.1515/jcbtp-2017-0025

[3] Li, C.P., Qin, J.X., Li, J.J. and Hou, Q. (2016) The Accident Early Warning System for Iron and Steel Enterprises Based on Combination Weighting and Grey Prediction Model GM(1, 1). *Safety Science*, **89**, 19-27. https://doi.org/10.1016/j.ssci.2016.05.015

[4] Wu, Y. (2011) Computing and Intelligent Systems. Springer Berlin Heidelberg.

[5] 陈钦华. 构建学分制下高校学生学业预警机制的探索[J]. 广西师范学院学报(哲学社会科学版), 2007(S2): 60-65.

[6] 华金秋. 台湾高校学习预警制度及其借鉴[J]. 江苏高教, 2007(5): 136-138.

[7] 路鹏, 王基生, 殷明均. 数据挖掘技术在高等院校学业预警中的应用[J]. 中国教育技术装备, 2009(30): 120-122.

[8] 张海舰, 吕海航, 王芹芹. 关于大学生学业预警制度若干问题的探讨[J]. 职业时空, 2010, 6(11): 73-74.

[9] 范冰. 高职院校学业预警机制的实践探索[J]. 常州信息职业技术学院院报, 2011, 10(4): 73-75.

[10] 郑俊玲. 基于 KPCA 的大学生学业预警模型及其应用[D]: [硕士学位论文]. 唐山: 华北理工大学, 2015.

[11] 刘庆军, 陈坤, 刘晓光. 煤与瓦斯突出预测 PCA-距离判别法研究[J]. 中国煤炭, 2016, 42(10): 97-101.

[12] 杨超, 史秀志. 基于 Fisher 判别法模型的矿山安全标准化等级评价[J]. 黄金科学技术, 2018, 26(2). http://kns.cnki.net/kcms/detail/62.1112.TF.20171121.0907.004.html

[13] 崔光磊, 熊伟. 贝叶斯判别法在煤与瓦斯突出预测中的应用[J]. 煤炭工程, 2013, 45(3): 96-98.

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2169-2556，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：ass@hanspub.org