可拓数据挖掘在学生成绩分析中的应用研究

刘大莲, 田英杰

引用本文:
刘大莲,田英杰. 可拓数据挖掘在学生成绩分析中的应用研究[J]. 智能系统学报, 2022, 17(4): 707–713.
LIU Dalian,TIAN Yingjie. Application of extension data mining in student achievement analysis[J]. *CAAI Transactions on Intelligent Systems*, 2022, 17(4): 707–713.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202112020

DOI: 10.11992/tis.202112020
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.tp.20220419.1703.007.html

可拓数据挖掘在学生成绩分析中的应用研究

刘大莲[1,2], 田英杰[3]

（1. 北京联合大学 数理部, 北京 100101; 2. 北京联合大学 数理与交叉科学研究院, 北京 100101; 3. 中国科学院虚拟经济与数据科学研究中心, 北京 100190）

摘　要: 为了充分利用教育大数据资源, 促进教学改革良性发展, 本文利用可拓支持向量机、可拓 k-均值聚类等多种可拓数据挖掘方法及皮尔逊相关系数, 对高校学生数学课程的平时作业、期中和期末考试成绩等进行挖掘和分析, 探索试卷设计的科学性, 学生对知识点的掌握程度, 以及哪些题目是影响学生成绩的主要因素, 针对每个学生给出其该门课程日后学习的侧重点等。将不断发展的前沿科研方法应用于需要不断改革的教育教学中, 同时也对长期沉睡的庞大的学生成绩数据加以充分利用, 科研指导教学, 教学反哺科研, 起到很好的示范作用。

关键词: 可拓学; 数据挖掘; 分类; 聚类; 支持向量机; 皮尔逊相关系数; 教育大数据; 学生成绩分析
中图分类号: TP18　　**文献标志码**: A　　**文章编号**: 1673−4785(2022)04−0707−07

中文引用格式: 刘大莲, 田英杰. 可拓数据挖掘在学生成绩分析中的应用研究 [J]. 智能系统学报, 2022, 17(4): 707–713.
英文引用格式: LIU Dalian, TIAN Yingjie. Application of extension data mining in student achievement analysis[J]. CAAI transactions on intelligent systems, 2022, 17(4): 707–713.

Application of extension data mining in student achievement analysis

LIU Dalian[1,2], TIAN Yingjie[3]

(1. Institute of Mathematics and Physics, Beijing Union University, Beijing 100101, China; 2. Institute of Fundamental and Interdisciplinary Sciences, Beijing Union University, Beijing 100101, China; 3. Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: To make full use of the educational big data resources and promote the sound development of teaching reform, this paper applies several extension data mining methods, including an extenics support vector machine, an improved k-means algorithm based on extension distance, etc., and the Pearson's correlation coefficient, to analyze the usual homework, midterm, and final examination results of a college students' mathematics course, to explore the scientificity of test paper design, students' mastery of knowledge points, and which topics are the main factors affecting students' performance. Furthermore, some advance information is given to each student to tell them which point they should focus on in this course later. This paper applies the constantly developing, cutting-edge scientific research methods to the education and teaching that need constant reform and makes full use of the huge student achievement data that has been sleeping for a long time. It has played a good example of scientific research guiding teaching and teaching feeding scientific research.

Keywords: extenics; data mining; classification; clustering; support vector machine; Pearson's correlation coefficient; big data in education; analysis of student score

当代社会, 随着信息技术的突飞猛进, 高等学校的教育教学改革的深入化也受到了深刻的影响。尤其处于大数据时代, 数据挖掘的各种方法被应用到教育行业[1-6], 为寻找更好的教育教学方法提供了新思路。学生成绩在高等学校里不但是衡量学校人才培养水平的一个重要指标, 同时也是教育大数据中的一个重要内容。由于学生成绩具有数据类型相对统一, 数据量较大, 相对容易获取等特点, 因此依据恰当的数据挖掘技术, 对学生成绩进行不同角度的深入挖掘和分析, 从而得到指导教学的新方法或新理论的研究成为高等学校教学改革的一个研究热点。丁智斌等[7] 利用

收稿日期: 2021−12−11.　**网络出版日期**: 2022−04−22.
基金项目: 国家自然科学基金面上项目(72071049); 教育部人文社科规划基金项目(18YJAZH049); 北京联合大学教育教学研究与改革项目(JJ2021Y053).
通信作者: 田英杰. E-mail: tyj@ucas.ac.cn.

决策树中的 ID3 算法对学生成绩进行分析，从而得出了影响学生成绩的内部原因及一些其他相关结论。喻铁朔等[8]是基于支持向量机（support vector machine, SVM）等 4 种数据挖掘的方法对学生成绩进行预测，从不同角度对 4 种模型进行对比，得出不同模型适用于不同课程的结论，对高校学生课程成绩预测。钟文精等[9]基于 k-means 聚类算法，对学生成绩进行聚类分析，为进行深入的教学改革和设计提供数据依据。本文依据可拓数据挖掘中的几种重要算法及皮尔逊相关系数，对北京某高校经管类学生的数学课程相关成绩进行多角度深入分析，从而得到一些和教学相关的重要结论，为改进教学方法，提高教学质量给出合理化建议。

1 基础知识与算法

1.1 可拓支持向量机

可拓学是由广东工业大学蔡文研究员创立的一门原创学科。在众多专家学者的不懈努力下，历经 30 余年的潜心研究，建立了可拓论体系和可拓创新方法体系[10-18]。可拓数据挖掘[19-20]是将可拓学的理论和方法与挖掘数据的方法技术相结合的一门新技术，可拓支持向量机[21]就是其中一种经典机器学习算法与可拓理论深入结合而产生的新算法。与标准的支持向量分类机不同，可拓支持向量机是解决可拓分类问题的，其在进行标准分类问题预测的同时，更注重于找到那些通过变化分量（特征）的值而转换类别的样本，这样的样本称为可拓样本，而相应的变量称为可拓变量。

算法 可拓支持向量分类机算法（ESVM）

1）给定训练集：其中 $x_i \in \mathbf{R}^n$，$y_i \in \mathbf{Y} = \{1,-1\}$，$i = 1,2,\cdots,l$。给定可拓样本 x_k 的可拓变量 j 的可拓区间 a_j^k 和 b_j^k，$j = 1,2,\cdots,n$；选择合适的惩罚参数 $C > 0$；

2）构造并求解最优化问题：

$$\min_{\alpha} \quad \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{l}\alpha_i$$

$$\text{s.t.} \quad \sum_{i=1}^{l} a_i y_i = 0$$

$$0 \leqslant \alpha_i \leqslant C,\ i = 1,2,\cdots,l$$

得最优解 α_i^*；

3）计算 b^*：

$$b^* = y_j - \sum_{i=1}^{l}\alpha_i^* y_i (x_i \cdot x_j),\ j \in \left\{ j \middle| 0 \leqslant \alpha_j^* \leqslant C \right\};$$

4）构造决策函数：

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{l}\alpha_i^* y_i (x_i \cdot x) + b^* \right);$$

5）对于输入 x_k，首先用决策函数 $f(x_k)$ 得到其对应的预测类别 y_k，然后用其可拓变量对应的可拓区间 a_j^k 和 b_j^k 分别代替 $[x_k]_j$，这样对 $|E|$ 个可拓变量，就得到 $2^{|E|}$ 个不同的组合值。相应的，基于 x_k 得到了 $2^{|E|}$ 个新的输入，分别用决策函数来判断，若有一个被判断为 $-y_k$，则认为该输入是可变换的。

1.2 基于可拓距的 k-means 聚类算法

可拓 k-means[22]基于可拓学中点 x 与区间 $X_0 = <a,b>$ 的距离定义，提出了一种选取 k-means 算法初始聚类中心的新方法，算法描述如下：

1）计算出两两样本间距离及等效密集距离区间 $Z = [A,B] = [\min(D), \max(D)]$，其中 $\left\{ D = d \middle| d = \sqrt{\sum_{p=1}^{m}(x_i^p - x_j^p)^2} \right\}$ 为两两样本间距离集合；

2）按照两样本的距离 Z 对区间的左右测距的距离和左、右侧距[9]的定义，将距离映射为左侧距 $\rho_i^{(i,j)}\rho_i^{(i,j)}$ 及可拓右侧距 $\rho_j^{(i,j)}\rho_r^{(i,j)}$，将 $\rho_i^{(i,j)}\rho_i^{(i,j)}$ 按从小到大顺序依次排序，同 $S = \{(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)\}$ 时计算样本间可拓平均左侧距 $\overline{\rho_l}$ 及可拓平均右侧距 $\overline{\rho_r}$；

3）遍历排序好的可拓距，将其中首个大于样本间可拓平均左侧距 $\overline{\rho_l}$ 的可拓距对应中心点坐标作为第一个初始聚类中心；

4）计算排好序可拓距中下一个值对应中心点坐标并依次计算出其与已确定的初始聚类中心的可拓距，将其与样本平均可拓右测距 $\overline{\rho_r}$ 进行比较，若其均大于 $\overline{\rho_r}$，则该中心点坐标作为下一个初始聚类中心；否则重新执行步骤 4；

5）如果遍历一次后，初始聚类中心未达到 K，则按式（1）计算出缩小因子 η，动态缩小样本平均可拓右侧距 $\overline{\rho_r}$，重新回到步骤 3；

$$\eta = \begin{cases} 1 + \dfrac{c_n^2 - k'}{c_n^2}, & k' \neq K \\ 1, & k' = K \end{cases} \tag{1}$$

式中：k' 为每次遍历后所获得的初始聚类中心个数；K 为指定聚类中心数

6）若聚类中心数达到 K 时，则完成初始聚类中心的选取。

1.3 皮尔逊（Pearson）相关系数

Pearson 相关系数[23]用于分析定量数据，当数据满足正态分布时可用 Pearson 相关系数查看变量间相关性。其公式为

$$r = \frac{\sum_{i=1}^{n}(x_i - \hat{x})(y_i - \hat{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \hat{x})^2 (y_i - \hat{y})^2}}$$

式中：相关系数 r 的取值范围为 $-1 \leqslant r \leqslant 1$。$r > 0$ 为正相关，$r < 0$ 为负相关，$0 < |r| < 1$ 表示相关程度。

2　高校学生成绩特点分析

2.1　数据描述

收集了北京联合大学 2018—2019 学年包括旅游学院、管理学院和商务学院 3 个学院共计 929 名学生的数据, 包括经管类概率论与数理统计(I)课程的平时作业、期中和期末考试成绩等。根据期末试卷的 5 道客观题(记为 kg_1~kg_5)和 10 道主观题(记为 zg_6~zg15)共 15 道题目, 总结出 15 个主要的知识点。为便于分析, 我们把每个学生的知识点掌握描述成一个 15 维向量, 向量的每个分量即为该生在某个知识点上的掌握程度。而知识点的掌握程度则根据学生的平时作业成绩、期中和期末试卷上考核相应知识点的得分, 综合计算得到。最后根据每个学生期末试卷的考试总成绩的及格与否把学生分成正负两类, 及格为正类, 不及格为负类。这样把所有学生组成一个大小为 929 的两类分类问题的数据集 1, 记为 S_1。

收集了我北京联合大学 2018—2019 学年包括旅游学院、管理学院和商务学院 3 个学院共计 841 名学生的数据, 包括微积分(II)课程的平时作业、期中和期末考试成绩等。根据期末试卷的 6 道客观题(记为 kg_1~kg_6)和 12 道主观题(记为 zg_7~zg_18)共 18 道题目, 总结出 18 个主要的知识点。同上述 S_1 数据处理类似, 我们把每个学生的知识点掌握描述成一个 18 维向量, 根据每个学生期末试卷的考试总成绩的及格与否把学生分成正负两类。这样把所有学生组成一个大小为 841 的两类分类问题的数据集 2, 记为 S_2。

下面将基于 S_1 和 S_2 进行学生成绩特点的挖掘分析。

2.2　基于可拓 SVM 的试卷题目影响力分析

基于成绩数据集 1, 探索哪些知识点是影响学生及格与否的主要因素, 从而检测试卷是否满足出题意愿; 进一步, 对每个学生, 可以给出决定其及格与否的某个或某几个具体题目, 以便学生以后有所侧重学习。

首先, 对建立的训练集 $S_1=\{(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)\}\in(\mathbf{R}^n\times\mathbf{Y})^l$, 其中 $x_i\in\mathbf{R}^{15}$, $y_i\in\mathbf{Y}=\{1,-1\}$, $i=1,2,\cdots,929$, 利用 5–折交叉验证方法, 选取最优的参数 C 和径向基核函数参数, 并用最优参数对整个训练集进行训练, 得到最终的决策函数。利用此决策函数进行规则抽取[24], 可以得到基本的分类规则, 我们这里将分类规则按照决策树的形式表示如图 1 所示。

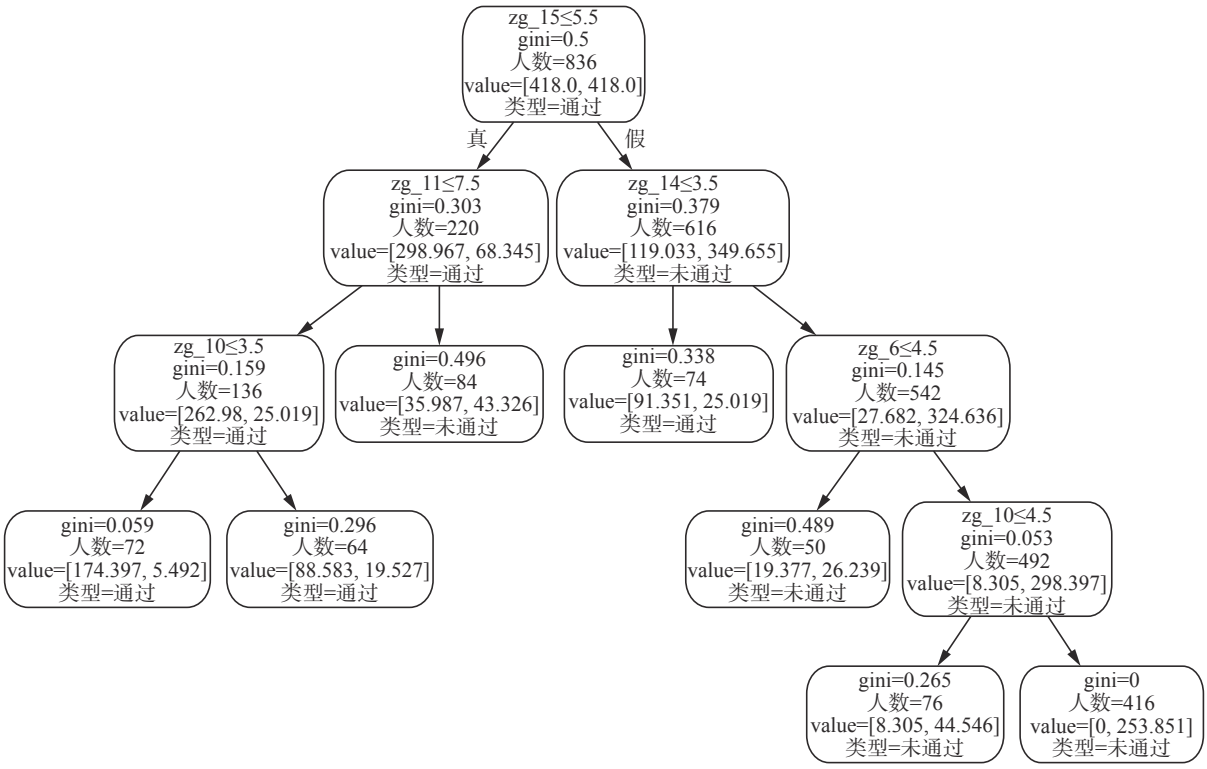


图 1　分类规则图

Fig. 1　Classification rule diagram

由图 1 可以看出, 据此规则得到的节点数为 13, 叶子节点数为 7, 树的最大深度为 5, 最基本的区分规则是选择那些对学生是否及格判断起主要作用的题型及题号。从树中可以看出在众多规

则中 zg_15, zg_14, zg_11, zg_6 都被作为分枝的因素。从根节点带有特征取值范围来看，根节点的两个分支分别代表两类学生成绩分布，一类是 zg_15 的得分大于 5.5 分，另一类是 zg_15 得分小于 5.5 分。

从根节点的左分支中关于 zg_11 得分是否大于 7.5 的分支对比观察中可以发现，即便学生对 zg_11 得分小于 8.5，学生的及格率依然很高，由此可见，对该分支的进一步挖掘，可以找出更加具备辨识度的特征以及取值范围。

从根节点的右分支出发，我们可以发现，第个二分支节点判断的特征为 zg_14 的得分是否小于等于 3.5。从选择人数上看，zg_14 的得分大于 3.5 的学生比相应得分小于 3.5 的人数高出 468 人，但是在 zg_14 得分超过 3.5 的同学不及格的概率更高。由此可见，在众多主观题中，zg_15 对学生的成绩及格影响更高，而 zg_14 对是否成绩及格的概率呈现出较低的相关性，所以导致在 14 号主观题得分高的同学在最后的及格率分析中影响度不高。

将上述分析进一步总结到规则表 1，从中可以看出，影响学生对概率统计及格率的主要因素有以下 3 点：

表 1　S_1 及格率规则

Table 1　Pass rate rules of S_1

类型	选择规则	合计人数
通过	zg_15≤5.5 → zg_11≤7.5	136
通过	zg_15>>5.5 → zg_14≤3.5	74
未通过	zg_15≤5.5 → zg_11>7.5	84
未通过	zg_15>5.5 → zg_14>3.5 → zg_6≤4.5	542

1）第 15 号主观题：从 5 条分支规则中可以发现，将第 15 号主观题得分作为根节点分支范围的合计人数最多，由此可以推断第 15 号主观题是影响学生对于概率统计课程及格率的主要因素。

2）第 14 号主观题：在所有的规则中同样也对第 14 号主观题的得分范围进行了划分，基于前面的第 15 题的分支背景，第 14 号主观题的取值范围也有了相应的调整。

3）第 11 号主观题：在规则表中，存在前馈规则一致的两条规则。第 11 号主观题的得分是否超过 7.5 分是区分他们的关键。

另外，从结果上看，对规则主要的考虑因素也集中在主观题型中，而客观题影响度较低。为了进一步探究一张试卷中各个题型之间的重要性，我们对概率统计试卷上的题型进行了影响度可视化操作，可视化结果如图 2 所示。

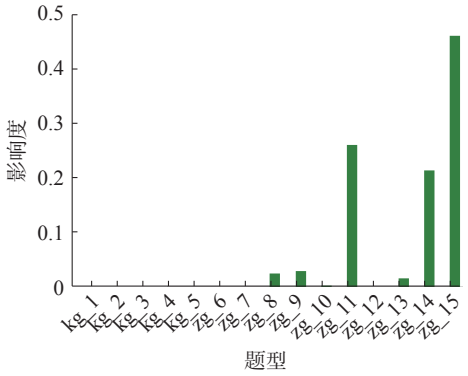


图 2　各题影响度可视化图

Fig. 2　Impact of each question

主观题第 15 题 zg_15 作为影响学生概率统计及格的重要因素，该现象在管理学院、旅游学院尤为明显。主要原因在于 zg_15 得分难度低，导致该题得高分的同学较多；对于其他的主观题，例如 zg_14 和 zg_11 也有类似的趋势。而反观客观题的影响比例，可以看到影响力几乎为 0，原因在于客观题题型分值较小，且相对得分容易获得，所以导致客观题所占的影响力整体较低。结合上述表 1 的分析研究，绝大多数及格学生的提分关键在于第 11、14 和 15 号主观题。

在上面已得到普遍规律的前提下，进一步分析影响每个学生是否及格的关键知识点：

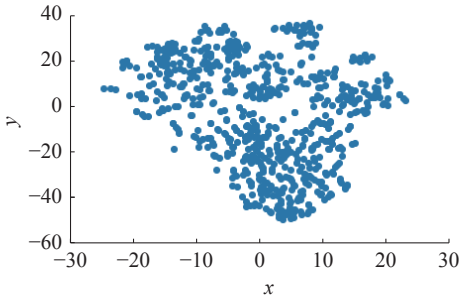
因每个题目学生得分都有不同，所以每个题目对应的变量都是可拓变量。首先定义所有题目 $j(j=1,2,\cdots,15)$的可拓区间，即$[a_j,b_j]$。这里将每个题目不得分和得最高分设为可拓区间上下界，即$a_j=0$，b_j为该题目的得分。针对每个学生x_k的每个题目对应的变量，用其可拓变量对应的可拓区间a_j^k和b_j^k分别代替$[x_k]_j$和$[y_k]_j$，这样对$|E|=15$个可拓变量，就得到2^{15}个不同的组合值。相应的，基于x_k，利用决策函数得到了2^{15}新的输入，分别用决策函数来判断，若有一个被判断为$-y_k$，则认为该输入是可变换的。

以学生 t_1 为例，我们得到 kg_4, zg_13 是影响其及格与否的 2 个关键题目，即如果学生 t_1 在 kg_4 和 zg_13 对应的知识点掌握程度从最低变为最高的情况下，其将由不及格而变成及格；而对于学生 t_2，同理可知学生对 kg_2，zg_13, zg_15 对应的知识点掌握程度是影响其及格与否的关键。

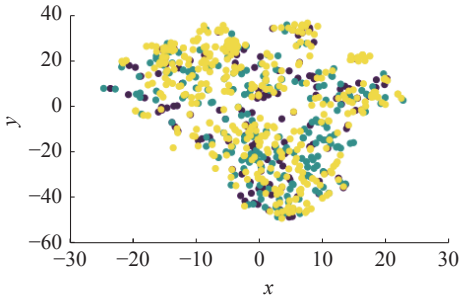
2.3　基于可拓距的 k-means 聚类算法成绩特定分析

基于成绩数据集 S_2，我们拟分析学生成绩分布的整个规律.首先建立数据集$S_2=\{(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)\}$，其中$x_i\in\mathbf{R}^{18}$，$y_i\in\mathbf{Y}=\{1,-1\}$，$i=1,$

2,⋯,841。为了对数据有整体的了解和把握, 以便于进一步从不同角度进行分析。首先, 我们对数据利用 t-SNE 方法进行降维和可视化展示, 图 3(a) 是微积分 (II) 课程的全体成绩分布图。可以发现, 图中的成绩数据分布较为紧密, 紧密的样本分布为数据聚类添加了难度。同时, 为了验证“同一学院的学生, 该门课程的总体水平较为接近”这一设想, 我们按照学院划分, 将管理学院、旅游学院和商务学院的学生成绩作为不同类别的数据, 利用 t-SNE 方法进行降维和可视化展示, 如图 3(b) 所示。很明显看出, 结果和我们预期吻合。(可视化图均为示意图, 坐标无实际意义。)



(a) 全体成绩分布图

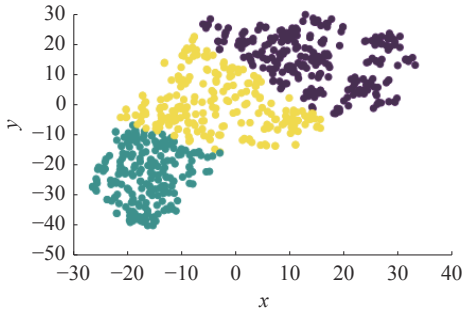


(b) 按学院划分成绩分布图

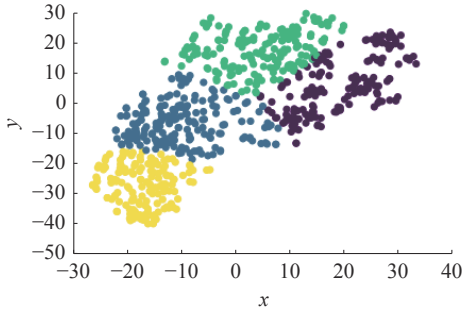
图 3　整体数据可视化图

Fig. 3　Visualization of the overall data

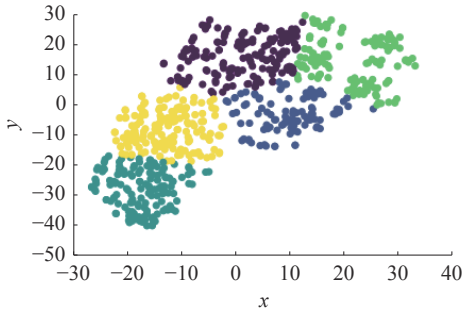
对于具有上述特征的数据, 采用上述 1.2 节中所阐述的基于可拓距的 k-means 聚类算法, 把 k 分别取为 3、4、5, 并利用 t-SNE 方法进行降维和可视化展示得到如下结果(如图 4), 可以看出 k=3 时效果比较好。



(a) k=3 可视化图



(b) k=4 可视化图



(c) k=5 可视化图

图 4　k-means 可视化图

Fig. 4　Visualization of k-means

进一步, 我们对聚类的 3 类进行分析, 对每一类中所有点的每个分量求均值, 探索每类的特点, 得到表 2。可以看出类别 2 与 1, 3 在各个题目对应的知识点掌握程度都有明显区别, 也就是类别 2 的学生, 几乎对所有知识点掌握都较差, 这些学生需要全面补习; 而类别 1 和 3 之间只在某些知识点上取值差别稍大, 比如 zg_18。

表 2　k=3 聚类分析表

Table 2　k=3 Cluster analysis table

类别	个数	kg_1	kg_2	kg_3	kg_4	kg_5	kg_6	zg_7	zg_8	zg_9	zg_10	zg_11	zg_12	zg_13	zg_14	zg_15	zg_16	zg_17	zg_18
1	314	2.21	2.68	2.70	2.60	2.76	0.39	5.88	5.92	5.87	5.7	5.98	6.3	6.68	6.08	6.04	9.27	9.79	1.41
2	200	**1.73**	**1.43**	**2.11**	**1.83**	**1.98**	**0.27**	**4.45**	**3.90**	**3.68**	**3.34**	**3.35**	**4.46**	**4.26**	**4.2**	**3.29**	**6.26**	**0.96**	**0.17**
3	225	2.32	1.69	2.45	2.31	2.72	0.25	5.26	5.07	5.15	5.25	5.81	6.19	6	5.88	5.66	9.027	7.25	0.67

2.4　基于 Pearson 相关系数的试卷题目相关性分析

基于数据集 S_1, 利用 Pearson 相关系数进行相关性分析, 结果如图 5 所示, 其中颜色越深代表着相关性越大。可以发现: 正对角线代表着当前特

征与特征自身的相关性计算值，正对角线上的值均为 1，颜色最深。其余部分代表着当前特征与其他特征的相关性计算，颜色的深浅代表着相关性的强弱。具体而言：客观题 kg_1,kg_2,kg_3,kg_4,kg_5 之间相关性热力图颜色为浅绿色，说明它们之间相关性较弱，但是总体保持着正相关的关系。据此可以推断，客观题一道题的得分情况对另外一题的得分情况影响较低，或者说题目本身考查的知识点不相关。而主观题之间的相关性则更加复杂。根据主观题之间的相关性热力图分布，它们之间存在负相关和正相关两种相关关系。相关性的数值越接近 1 或−1，说明两组数据之间正向或反向线性关联越强。例如，zg_6 与 zg_7、zg_7 与 zg_8，zg_8 与 zg_9 之间的相关性热力图颜色为黄色，说明它们之间的相关性为负相关。与之相反的情况为：zg_7 与 zg_9、zg_11 与 zg_12 之间的相关性热力图颜色为蓝色，说明具有很强的正相关性，两个特征的相关密切程度比较高。此时就要引起注意，试卷中 zg_7 与 zg_9、zg_11 与 zg_12 之间是否考察知识点重合，还是题目难易程度相近引起的高度相关。如果出现命题知识点重合，是否符合我们考核的目的，从而对考试后试卷命题合理性分析给出提示。

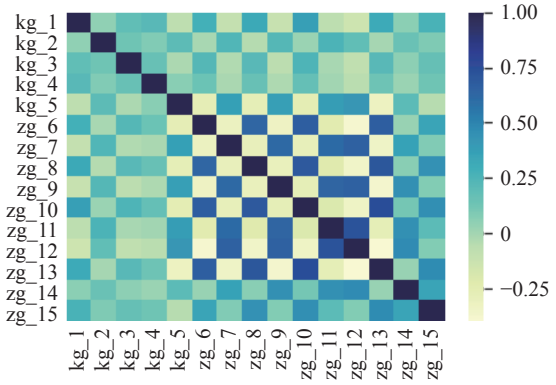


图 5　题目相关性热力图

Fig. 5　Correlation map

3　结束语

本文主要基于可拓数据挖掘的几种重要方法及皮尔逊相关系数，对高校学生成绩利用不同模型，从不同角度进行分析，从而分析影响学生成绩的主要题目，探索学生对知识点的掌握程度。进一步，对每个学生，可以给出决定其及格与否的某个或某几个具体知识点，以便学生以后有所侧重学习。试卷中各题目相关性强弱分析的结论，也对课程考核等方面给出合理化指导和建议。将不断发展的、前沿的科学技术、科研方法应用于不断深化改革的教育教学中，同时也对长期沉睡的庞大的学生成绩数据加以充分利用，教学促进科研，科研反哺教学，起到了示范作用。采用的相关算法是我们精心选取的算法，针对相关成绩数据分析有一定的优势。将来我们可以进一步深入研究，探讨如何将解决矛盾问题的可拓学和机器学习的相关算法深度融合，起到如虎添翼的作用。深究如何进一步将科研的方法应用到教育大数据中，从而对推进教学改革，进一步提高高校教学质量做出贡献。同时也希望上述分析能起到抛砖引玉的作用。

参考文献：

[1] 徐承俊，朱国宾. 数据挖掘在全国计算机等级考试(NCRE)成绩分析中的研究及应用 [J]. 计算机应用与软件, 2020, 37(8): 64–67,73.
XU Chengjun, ZHU Guobin. Research and application of data mining in national computer rank examination (NCRE) achievement analysis[J]. Computer applications and software, 2020, 37(8): 64–67,73.

[2] 郭鹏，蔡骋. 基于聚类和关联算法的学生成绩挖掘与分析 [J]. 计算机工程与应用, 2019, 55(17): 169–179.
GUO Peng, CAI Cheng. Data mining and analysis of students' score based on clustering and association algorithm[J]. Computer engineering and applications, 2019, 55(17): 169–179.

[3] 唐笑林. 数据挖掘技术的研究和应用 [J]. 华东理工大学学报(自然科学版), 2008, 34(2): 290–295.
TANG Xiaolin. Application and research of data mining[J]. Journal of east China University of Science and Technology (natural science edition), 2008, 34(2): 290–295.

[4] 张树滑. 基于 ID3 算法的大学生成绩数据挖掘与体能分析系统设计 [J]. 现代电子技术, 2019, 42(5): 104–106,110.
ZHANG Shuhua. Design of sports achievement data mining and physical fitness analysis system based on ID3 algorithm[J]. Modern electronics technique, 2019, 42(5): 104–106,110.

[5] 王小根，陈瑶瑶. 多模态数据下混合协作学习者情感投入分析 [J]. 电化教育研究, 2022, 43(2): 42–48,79.
WANG Xiaogen, CHEN Yaoyao. Analysis of blended collaborative learners' emotional engagement based on multimodal data[J]. E-education research, 2022, 43(2): 42–48,79.

[6] 沈苗，来天平，王素美，等. 北京大学课程推荐引擎的设计和实现 [J]. 智能系统学报, 2015, 10(3): 369–375.
SHEN Miao, LAI Tianping, WANG Sumei, et al. Design

and implementation of the course recommendation engine in Peking University[J]. CAAI transactions on intelligent systems, 2015, 10(3): 369–375.

[7] 丁智斌, 袁方, 董贺伟. 数据挖掘在高校学生学习成绩分析中的应用 [J]. 计算机工程与设计, 2006, 27(4): 590–592.
DING Zhibin, YUAN Fang, DONG Hewei. Application of data mining to analysis of university students' grades [J]. Computer engineering and design, 2006, 27(4): 590–592.

[8] 喻铁朔, 李霞, 甘玲. 基于学生成绩回归预测的多模型适用性对比研究 [J]. 中国教育信息化, 2020(17): 23–28.

[9] 钟文精, 焦中明, 蔡乐. 基于 K-Means 算法的学生成绩聚类分析 [J]. 教育信息技术, 2021(5): 56–58.

[10] 蔡文, 杨春燕, 何斌. 可拓逻辑初步 [M]. 北京: 科学出版社, 2003.

[11] CAI Wen, YANG Chunyan, LIN Weiehu. Extension engineering methods[M]. Beijing: Science Press, 2003

[12] 蔡文. 可拓集合和不相容问题 [J]. 科学探索学报, 1983(1): 83–97.
CAI Wen. Extenics and incompatibility[J]. Journal of scientific exploration, 1983(1): 83–97.

[13] 李文军, 杨春燕, 汤龙, 等. 可拓学中相关关系的变换方法研究 [J]. 智能系统学报, 2019, 14(4): 619–626.
LI Wenjun, YANG Chunyan, TANG Long, et al. Research on the transformation method for the correlation relation in extenics[J]. CAAI transactions on intelligent systems, 2019, 14(4): 619–626.

[14] 杨春燕, 李卫华, 汤龙, 等. 基于可拓学和 HowNet 的策略生成系统研究进展 [J]. 智能系统学报, 2015, 10(6): 823–830.
YANG Chunyan, LI Weihua, TANG Long, et al. Strategy-generating system based on extenics and HowNet[J]. CAAI transactions on intelligent systems, 2015, 10(6): 823–830.

[15] 王丽萍, 叶季平, 苏学灵, 等. 基于可拓学理论的防洪调度方案评价研究与应用 [J]. 水利学报, 2009, 40(12): 1425–1434.
WANG Liping, YE Jiping, SU Xueling, et al. Evaluation of flood control operation program based on extenics theory and its application[J]. Journal of hydraulic engineering, 2009, 40(12): 1425–1434.

[16] 杨春燕, 李兴森. 可拓创新方法及其应用研究进展 [J]. 工业工程, 2012, 15(1): 131–137.
YANG Chunyan, LI Xingsen. Research progress in extension innovation method and its applications[J]. Industrial engineering journal, 2012, 15(1): 131–137.

[17] 杨春燕, 蔡文. 可拓学与矛盾问题智能化处理 [J]. 科技导报, 2014, 32(36): 15–20.
YANG Chunyan, CAI Wen. Extenics and intelligent processing of contradictory problems[J]. Science & technology review, 2014, 32(36): 15–20.

[18] 杨春燕, 蔡文, 涂序彦. 可拓学的研究、应用与发展 [J]. 系统科学与数学, 2016, 36(9): 1507–1512.
YANG Chunyan, CAI Wen, TU Xuyan. Research, application and development on extenics[J]. Journal of systems science and mathematical sciences, 2016, 36(9): 1507–1512.

[19] 蔡文, 杨春燕, 陈文伟, 等. 可拓集与可拓数据挖掘 [M]. 北京: 科学出版社, 2008.

[20] 杨春燕, 蔡文. 可拓数据挖掘研究进展 [J]. 数学的实践与认识, 2009, 39(4): 134–141.
YANG Chunyan, CAI Wen. Recent progress in extension data mining[J]. Mathematics in practice and theory, 2009, 39(4): 134–141.

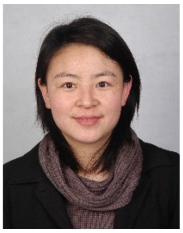
[21] 陈晓华, 刘大莲, 田英杰, 等. 可拓支持向量分类机 [J]. 智能系统学报, 2018, 13(1): 147–151.
CHEN Xiaohua, LIU Dalian, TIAN Yingjie, et al. Extension support vector classification machine[J]. CAAI transactions on intelligent systems, 2018, 13(1): 147–151.

[22] 赵燕伟, 朱芬, 桂方志, 等. 基于可拓距的改进 k-means 聚类算法 [J]. 智能系统学报, 2020, 15(2): 344–351.
ZHAO Yanwei, ZHU Fen, GUI Fangzhi, et al. Improved k-means algorithm based on extension distance[J]. CAAI transactions on intelligent systems, 2020, 15(2): 344–351.

[23] 盛骤, 谢式千. 概率论与数理统计及其应用 [M]. 2 版. 北京: 高等教育出版社, 2010.

[24] YANG Sixiao, TIAN Yingjie, ZHANG Chunhua. Rule extraction from support vector machines and its applications[C]//2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. Lyon: IEEE, 2011: 221–224.

作者简介：



刘大莲，副教授，主要研究方向为最优化理论与方法、数据挖掘。发表学术论文 18 篇。

田英杰，教授，博士生导师，中国科学院大学经济与管理学院副院长，主要研究方向为机器学习、大数据挖掘与最优化。出版中英文专/合著 5 部，近 5 年发表学术论文 50 余篇。