Advances in Computer Science Research (ACSR), volume 61
7th International Conference on Education, Management, Computer and Society (EMCS 2017)

Application of Large Data Analysis in Ceramic Art Students

Hong Ye and Xudong Wu

(Jiangxi Ceramic Arts and Crafts Vocational and Technical College, Jingdezhen, Jiangxi 333000)

Abstract. Based on the link between the university library management system and the educational administration management system, this paper established three kinds of analysis models including cluster analysis, association rule and decision tree, to study the correlation between the information of books borrowing and achievements by ceramic art students. The experimental results show that the student's book borrowing situation has a positive correlation with student achievement.

Keywords: Association rules; Clustering analysis; Decision tree

大数据分析在陶瓷艺术专业学生中的应用

叶虹[1] 吴旭东[2]

(江西陶瓷工艺美术职业技术学院，江西 景德镇 333000)

摘要：论文通过高校图书管理系统与教务管理系统之间的联系，建立聚类分析、关联规则和决策树三种分析模型，对陶瓷艺术专业学生图书借阅信息与成绩进行相关性研究，实验结果表明，学生的图书借阅情况与学生成绩存在正相关。

关键字：关联规则；聚类分析；决策树

中图分类号：G250.7　　文摘标识码：A

引言

目前，在高校图书馆开展数据挖掘技术已经非常广泛，其中包括对读者借阅行为的关联分析、聚类；利用决策树对图书采购进行科学管理；利用神经网络对读者进行预测等。虽然这些研究已经取得了一定的成果，然而这些仅仅只停留在读者和图书馆之间，对高校学生和教师的教学方面作用不大。论文利用图书馆的借阅数据和学生成绩数据进行关联分析，对学院陶瓷艺术设计专业 400 多名学生，在校期间 20 多门主干课程的考试成绩与图书利用情况相结合，利用数据挖掘算法试图找出图书馆利用和学习成绩之间的关系，进而对学生学习、教师授课、图书馆辅助教学参考用书，提供一定意义上的指导。

1 聚类分析

1.1 聚类分析 k 均值（k-means）算法

所谓聚类问题，就是给定一个元素集合 D，其中每个元素具有 n 个可观察属性，使用某种算法将 D 划分成 k 个子集，要求每个子集内部的元素之间相异度尽可能低，而不同子集的元素相异度尽可能高。其中每个子集叫做一个簇。与分类不同，分类是示例式学习，要求分类前明确各个类别，并断言每个元素映射到一个类别，而聚类是观察式学习，在聚类前可以不知道类别甚至不给定类别数量，是无监督学习的一种。目前聚类广泛应用于统计学、生物学、数据库技术和市场营销等领域，相应的算法也非常的多。论文采用 k 均值算法。

1.2 实验过程

通过对原数据库的分析，该数据库不能直接运用 SPSS Clementine 进行挖掘。对图书馆的数据库中的数据和教务管理系统数据库中的数据进行联合操作，用 SQL Server2000 软件所提供 SQL 语言对读者信息

表、文献信息表、条码表、流通信息表、学生成绩表进行操作。通过 SQL 查询分析器查询所需要的记录，为分析图书馆中图书的借阅情况、学生成绩情况，运用数据库的视图功能建立虚拟表。如图 1-1 所示。

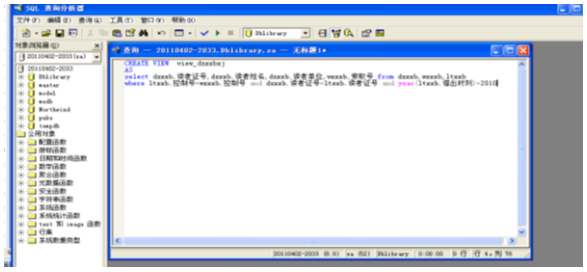


图 1-1 利用 SQL 查询分析器查询建立视图

通过分析，发现数据库视图表中还存在一些问题，如果这些数据直接用于数据挖掘将对结果产生较大的影响。记录中存在读者借阅图书重复的信息，某条记录存在字段信息不全。鉴于以上问题，需要对数据进行预处理，以保证信息的准确性，删除重复、信息不全的记录，如表 1-1 所示。

表 1-1 基于学生成绩的图书借阅信息表

图 1-2、图 1-3 为数据挖掘具体操作流程和设置，其中把学生读者分成三类。

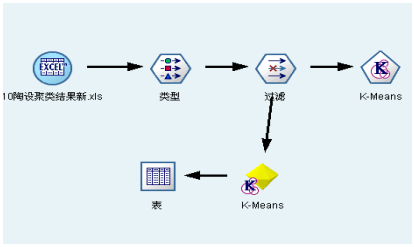


图 1-2 聚类挖掘流程图

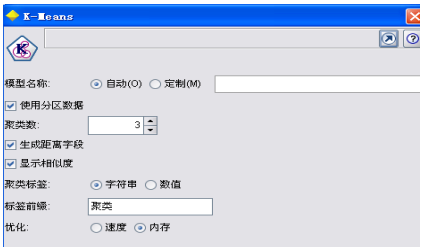


图 1-3 聚类数设置图

如图 1-4、1-5 结果所示，将学生读者分为三类，第一类学生读者平均图书借阅量为 25 册，所有课程的平均分为 80 分；第二类读者为 22 册，平均分为 79 分；第三类读者 11 册，平均分为 67 分；从各科的平均分来看，借书册书越来多，平均分就越高。借书量与学生成绩存在着一定关系，借书量在 11 册第三类读者学生学习成绩普遍偏低，体育课除外。特别是思想品德修养与法律基础课程，平均分才 20 分，非常明显区别于前两类，这说明这些学生普遍思想认识和思想水平有待提高。其中操行分，也就是平时出勤分和表现也普遍偏低，这也充分证明了平时不爱学习的，学习成绩偏差，那么来图书馆借书也特别少。借书量在 22 册的第二类读者，无论是专业成绩还是基础课成绩都明显高于第三类读者。

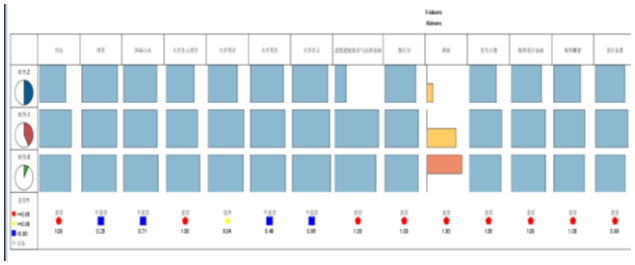


图 1-4 聚类汇总

	分类	借书量平均分		设计素描	大学英语	大学生心理学	设计色彩	装饰设计基础	摄行分	国画山水	陶瓷产品造型设计	体育	大学英语	大学语文	花鸟人物	设计美学	书法	装饰雕塑	思想道德修养与法	陶瓷彩绘装饰(1)
7	聚类-1	25	80	83.0	74.2	65.3	81.3	88.0	87.1	86.4	80.3	63.5	71.9	77.4	83.9	82.8	82.2	80.9	75.4	84.7
8	聚类-2	22	79	75.6	73.0	70.4	80.7	88.8	86.0	89.1	73.9	62.6	73.1	78.3	81.9	77.5	80.6	79.6	70.7	85.8
9	聚类-3	11	67	75.3	60.7	58.3	69.8	77.5	80.2	85.3	70.5	65.0	68.8	73.0	69.8	60.3	68.0	69.7	20.0	82.3	60.0
0	差值	14	12	8	14	7	12	11	7	1	10	-1	3	4	14	23	14	11	55	8	25

图 1-5 聚类结果

借书量在 25 册的第一类读者学生的专业成绩普遍强于二、三类读者。其中，第一类读者的专业课程平均分比第三类读者高许多。第一、三类平均分差值为 12 分、借书量差值为 14 册、其中分差较大的课程分别是思想道德修养与法律基础差值为 55 分，陶瓷彩绘装饰差值为 23 分，大学语文、花鸟人物、书法差值为 14 分，装饰设计基础、装饰雕塑差值 11 分。这说明经常来图书馆借书的读者对专业课程学习的有非常明显的帮助一般差距都在 10-20 分左右，而对基础帮助也不小，特别是大学语文和思想道德修养与法律基础。

利用聚类分析算法模型，建立了陶设专业的学生分类标准，通过学生的借阅情况来分析和预测学生的专业成绩，并得出了学生的学习成绩与借阅量情况存在正相关。

2 关联规则分析

2.1 数据预处理

把图书借阅信息数据与学生成绩数据合并，经过初步整理，为了能适应关联数据挖掘模型，对数据进行抽象、离散化等预处理。为了便于数据挖掘的进行，学生成绩采用的是百分制，需要将课程的名称以及考试成绩映射成字符。课程名称按数据表中的顺序依次映射成为英语字母 A、B、C 等；课程成绩映射的方法为:60 分以下设置为“3”,60 分至 80 分之间(包括 60 分)设置为“2”,80 分以上(包括 80 分)设置为 “1”。将借阅量用字母 T 表示，借阅量也分成 0、1、2、3、4 档，学生借阅总量中 T0 表示借书量在 0-9 册，T1 表示 10-19 册，T2 表示 20-29 册，T3 表示 30-39 册，T4 表示 40 册以上。经过处理以后如图 2-1、图 2-2 所示。

图 2-1 基于学生成绩的借阅信息表　　图 2-2 预处理后借阅信息表

2.2 实验过程与分析

实验过程如图 2-3、2-4 所示，共有 13 条规则，其中前 7 条是无效的结果，后 6 条是有效的。规则 1 中显示，大学生心理学 80 分以上和借阅量 20-29 册的学生占 6.76%，其中大学生心理学 80 分以上的同时有 60.0%的学生借阅量 20-29 册。规则 2 中显示，体育 80 分以上和借阅量 20-29 册的学生占 6.76%，其中体育 80 分以上的同时有 60.0%的学生借阅量 20-29 册。以上两条规则充分说明，大学生心理学成绩优秀、心态好、心理、身体健康的学生有 60.0%常来图书馆借书。规则 3 中显示，设计素描 80-60 分和借阅量 0-9 册的学生占 43.24%，其中设计素描 80-60 分的同时有 53.12%的学生借阅量 0-9 册。这说明设计素描课程一般的学生，利用图书馆也一般。规则 4 中显示，国画山水 80-60 分和借阅量 40 册以上的学生占 5.41%，其中国画山水 80-60 分的同时有 50.0%的学生借阅量 40 册以上。这点强有力的证明了，经常利用图书馆的学生，专业成绩也好。学院图书馆以陶瓷、艺术类收藏为重点，收藏许多画册，经常来馆的学生充分利用这些资源，所以中国画山水成绩较好。

如图 2-5 所示，以 10 陶设 1 班李硕、10 陶设 11 班金娟娟两位学生为例，借阅册数分别为 143、177 册，其中 135、99 册都是 J 类，而且以美术类书籍居多。

规则 5 显示，大学英语 60 分以下和借阅量 0-9 册的学生占 50.0%，其中大学英语 60 分以下同时有 2.7%的学生借阅量 10-19 册。这说明英语成绩差的学生，利用图书馆普遍不高。

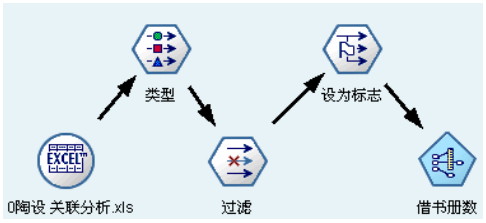
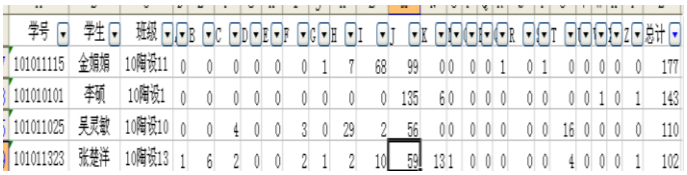


图 2-3 关联规则分析流程图

后项	前项	支持度 %	置信度
借书册数 = T0	借书册数_T0	37.84	100.0
借书册数 = T2	借书册数_T2	21.62	100.0
借书册数 = T1	借书册数_T1	17.57	100.0
借书册数 = T4	借书册数_T4	12.16	100.0
借书册数 = T3	借书册数_T3	10.81	100.0
借书册数 = T1	装饰雕塑 = P3	1.35	100.0
借书册数 = T1	书法 = O3	1.35	100.0
借书册数 = T2	大学生心理学 = E1	6.76	60.0
借书册数 = T2	体育 = C1	6.76	60.0
借书册数 = T0	设计素描 = B2	43.24	53.12
借书册数 = T4	国画山水 = I2	5.41	50.0
借书册数 = T1	大学英语2 = K3	2.7	50.0
借书册数 = T1	设计美学 = N3	2.7	50.0

图 2-4 关联结果



学号	学生	班级	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	总计
101011115	金娟娟	10陶设11	0	0	0	0	0	0	1	7	68	99	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	177
101010101	李硕	10陶设1	0	0	0	0	0	0	0	0	0	135	6	0	0	0	0	0	0	0	0	0	0	1	0	1	0	143
101011025	吴灵敏	10陶设10	0	0	4	0	0	3	0	29	2	56	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	110
101011323	张楚洋	10陶设13	1	6	2	0	0	2	1	2	10	59	13	1	0	0	0	0	0	0	4	0	0	0	1	0	0	102

图 2-5 学生借阅各类目数量详细图

规则 6 中显示，设计美学 60 分以下和借阅量 10-19 册的学生占 50.0%，其中设计美学 60 分以下同时有 2.7%的学生借阅量 10-19 册。这说明设计美学成绩差的学生，利用图书馆普遍不高。

3 决策树分析

3.1 决策树算法

论文采用的决策树算法是 C5.0 算法，包括生成规则的改进，C5.0 模型根据能够带来最大信息增益的字段拆分样本。第一次拆分确定的样本子集随后再次拆分，通常是根据另一个字段进行拆分，这一过程重复进行直到样本子集不能再拆分为止。最后，重新检验最低层次的拆分，那些对模型值没有显著贡献的样本子集被剔除或者修剪。

3.2 实验过程

利用决策树（C5.0）算法构建模型，对准备好的数据进行挖掘，将数据源设置为 EXCEL 表的形式，导入名为决策树.xls 的文件。对数据源的各个字段类型进行合适的设置。

考虑到学生专业学习成绩和借阅的专业书籍有一定的联系，那么可以假设学生借阅的专业类书籍越多，平均专业成绩越高。把陶瓷设计专业的学生的专业课程：设计素描、设计色彩、装饰设计基础、国画山水、陶瓷产品造型设计与制作 1、花鸟人物、设计美学、书法、装饰雕塑、贴花纸设计制作、陶瓷彩绘装饰(1)、B 、H 、I、 J、 K、 T 类书籍作为输入项，平均成绩作为输出成绩。如图 3-1、图 3-2 所示。

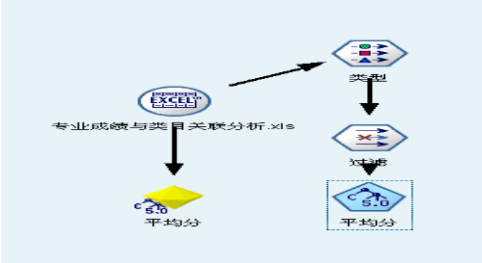


图 3-1 决策树流程图

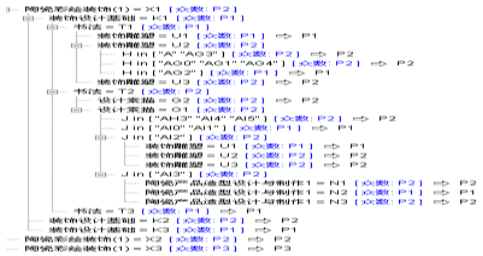


图 3-2 决策树结果

结果显示，陶瓷彩绘装饰成绩在 80 分以上且装饰设计基础在 80 分以上且书法在 80 分以上且装饰雕塑在 60-80 分且 H 类图书的借阅量在 2-9 册，那么平均分就在 80 分以上。陶瓷彩绘装饰成绩在 80 分以上且装饰设计基础在 80 分以上且书法在 60-80 分且设计素描 80 分以上且 J 类图书的借阅量在 30 册以上，那么平均分就在 60-80 分。结果表明，专业课程和图书的借阅量是成正比关系的，借阅专业书籍如 H 类和 J 类的图书的数量大，那么陶瓷设计专业学生的专业成绩就普遍偏高。

4 结论

传统的统计分析在图书借阅与学生成绩管理中已经应用了许多年，由于各种原因效果不佳。论文通过对图书借阅系统与教务管理系统相结合，运用数据挖掘中最常见的三种算法对学生的借阅情况和课程成绩相关性进行挖掘分析。实验结果表明，学生的图书借阅情况与学生成绩存在正相关。

致谢

［基金项目］江西省教育科学“十二五”规划科研项目资助（11YB472）

Acknowledgement

[Fund Project] Jiangxi Provincial Education Science "Twelfth- Five "Year Plan Research Project (11YB472)

参考文献：

[1]黄斯达. 基于图书馆借书信息的学生成绩挖掘模型研究[J]. 现代计算机, 2008（10）

[2]唐海萍. 基于数据挖掘技术在图书馆管理模式[J]. 现代情报, 2008（9）:109-110

[3]于徽. 数据挖掘技术及其在图书馆的应用[J]. 黑龙江科技信息，2008（9）:97-98

[4]臧卫华. 数据挖掘技术在高校图书馆中的应用研究[J]. 现代情报, 2008（3）:38-39

作者简介：

叶　虹（1984-），女，江西陶瓷工艺美术职业技术学院, 硕士, 江西省工艺美术师，研究方向：艺术学

吴旭东（1982-），男，江西陶瓷工艺美术职业技术学院讲师、图书馆员、网络工程师；研究方向：数据挖掘，图书馆个性化服务、群体智能

References:

[1] Huang Sida. Research on student achievement mining model based on library bookg borrowing information [J]. Modern Computer, 2008 (10)

[2] Tang Haiping. Based on data mining technology in library management mode [J]. Modern Information, 2008 (9): 109-110

[3] Yu Hui. Data mining technology and its application in library [J]. Heilongjiang Science and Technology Information, 2008 (9): 97-98

[4] Zang Weihua. Application of Data Mining Technology in University Library [J]. Modern Information, 2008 (3): 38-39