一种试卷分析及数据挖掘系统的开发

耿玉清[1]，张春生[2]
（1.科尔沁艺术职业学院 计算机系，内蒙古 通辽 028000；
2.内蒙古民族大学 计算机科学与技术学院，内蒙古 通辽 028043）

摘 要：应用统计学和数据挖掘技术，根据学生的试卷成绩进行分析，得出面向试题本身的一般质量统计指标，进行正态分布判断并输出成绩分布图，同时可以根据分析结果给出试卷评语。还可以面向教师出题情况进行分析，判断教师是否有漏题，违纪提分的可能性，并可以对一个班级不同课程，不同班级同一课程进行比较。同时利用数据挖掘技术，对学生成绩进行修补，异常成绩发现，分类，聚类，关联规则挖掘等。文中详细论述了试卷分析中各项技术的意义及实现方法。

关键词：试卷分析；指标；数据挖掘；统计学
中图分类号：G434 **文献标识码**：A **文章编号**：1673-629X(2010)10-0241-05

Development of Test Paper Analysis and Data Mining System

GENG Yu-qing[1], ZHANG Chun-sheng[2]
(1.Department of Computer, Kerqin Art Vocational College of Inner Mongolia, Tongliao 028000, China;
2.College of Computer Science and Technology, Inner Mongolia University for Nationalities, Tongliao 028043, China)

Abstract: Applying statistics and data mining, analyze students' paper marks and get the general statistical quality data oriented to test itself to judge normal distribution and draw out the distribution diagram of marks. Also, make the paper reviews on the basis of the results of analysis, and analyze test papers by teachers and judge the possibility that teachers disclose questions or raising marks against discipline, and make comparisons among different courses within one class and on one course in different classes. Meanwhile, use data mining to repair results of students' marks, discover, classify and cluster the abnormal results and mine association rules, etc. This paper demonstrates in detail the significance and the implementing method of technologies in the analysis of test paper.

Key words: test paper analysis; target; data mining; statistics

0 引 言

考试是检验教学效果和学习能力最有效的手段和措施。而试卷的出题质量至关重要，关系到对学生进行学习情况检验的成败，同时，也为以后合理地调整教学计划和教学方法提供理论依据[1,2]。

目前有关试卷分析方面的文献较多，但都不同程度地存在片面性，大部分试卷分析软件都仅仅围绕着学生总分进行一般性的统计指标分析，分析比较肤浅，不能挖掘出试卷中潜在的深层信息[3~11]。

针对以上问题，笔者从实际应用角度出发，针对每道小题，给出统计分析指标（各分数段的人数和百分比、最高分、最低分、平均分、易度、区分度等），进行正态分布判断和峰值分析，给出自动评语，提供了分析结果打印功能。同时提供教师分析，对教师的提分和漏题情况进行判断。提供多维显示功能，进行不同课程、不同班级的成绩比较。提供数据挖掘功能，进行异常成绩判断、关联规则分析、分类、聚类等等。

1 程序的总体结构

程序由8部分组成，主要包括数据处理、报表打印、信息维护、教师分析、多维显示、数据挖掘、系统维护、帮助（如图1所示）。

2 程序的主要功能

2.1 一般统计学分析

应用统计方法对试卷进行分析是试卷分析的基本功能，通过统计分析，给出一些统计分析指标，便于用户对试卷的通体情况的掌握。

（1）分段统计。

收稿日期：2010-01-14；**修回日期**：2010-04-18
基金项目：内蒙古人才基金资助项目（第8批）；内蒙古教育科研项目（NJZY07140）
作者简介：耿玉清（1961-），女，内蒙古通辽人，副教授，研究方向为数据库技术、计算机应用、计算机教学。

分段统计可以简单直观地展现出学生成绩的分布状况，是基本的分析结果和指标。

(2)最高分、最低分、平均分。

通过这三项指标可以掌握学生成绩总体情况，特别是极端情况。

(3)易度。

通过易度可以分析试题的难易程度，为对试题的评价提供依据。

(4)区分度。

区分度又叫鉴别力，它是测试学生实际水平的区分程度的指标，实际水平高的考生应该得高分，实际水平低的考生应该得低分。

(5)分布形态判断。

可根据试卷分析结果得到学生整体成绩的分布情况，从宏观上对学生的考试情况进行评价。

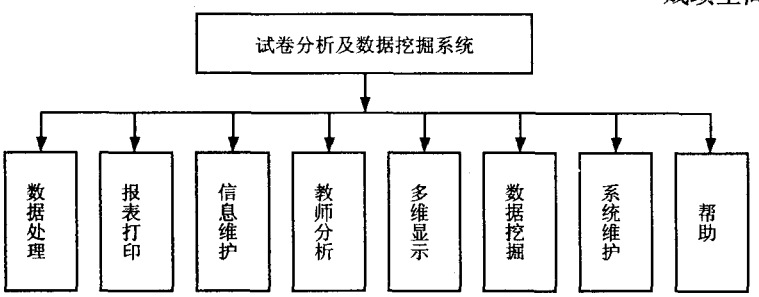


图 1 系统总体结构

2.2 自动评语的生成

程序根据分析得出的若干指标，分别从 5 个方面对试卷下结论，给出对试卷的评语。

(1)给出学生的最高分、最低分、平均分，通过对各分数段人数的分析得到学生成绩“学生成绩较高”、“学生成绩合理”、“学生成绩较低”的结论。

(2)通过对“平均易度”分析，给出试卷难易度结论；通过对“平均区分度”分析，给出试卷是否区分出不同类学生的结论。

(3)根据经验，通过对各小题的(1 - 易度 + 区分度)≤0.6 的判断，判断各小题是否过于简单或漏题。

(4) 正态分布判断成绩呈正态分布，试卷出题合理，难度项比例合适；成绩呈偏正态分布，试卷难度偏高，难度较大的题目较多；成绩呈偏负态分布，试卷难度偏低，难度较大的题目少。

(5)峰值分布判断成绩呈正峰态分布，分布合理；成绩呈高狭峰分布，试卷同难度题目较多，梯度偏小；成绩呈低阔峰分布，试卷难度项目比例接近，梯度较大。

2.3 报表打印

根据分析结果，打印出试卷分析报表，包括试卷基本信息，一般统计指标，自动评语，正态分布图等，如图 2 所示。

2.4 信息维护

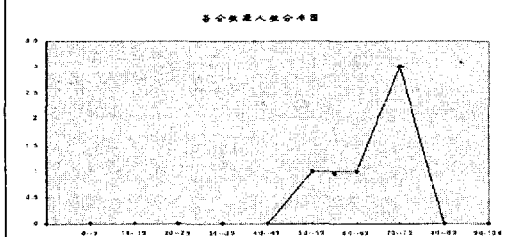
基本信息维护可对“班级信息”、“课程信息”、“课程类别”、“题型信息”、“学期信息”、“学院信息”6 种知识库信息进行浏览、插入、删除和更新操作。

内蒙古民族大学

分析日期：2008.12.01　　试卷质量状况分析表　　学院：数学与计算机科学学院

课程名称		power builder				课程类别		专业必修			
主考教师		张春生				考试班级		03计汉本			
班级人数		5				考试学期		06-07(2)			
成绩分段统计情况	分　数	90-100	80-89	70-79	60-69	50-59	40-49	30-39	20-29	10-19	0-9
	人　数	0	0	3	1	1	0	0	0	0	0
	百分比%	0.0	0.0	60.0	20.0	20.0	0.0	0.0	0.0	0.0	0.0
	最高分	76.0		最低分		57.0		平均分		66.6	
成绩分题统计情况	题　号	一	二	三	四	五	六	七	八	九	十
	题　型	选择	填空	简答	作图						
	标准分	25	25	25	25	0	0	0	0	0	0
	平均分	12.2	17.8	21.8	20.6	0.0	0.0	0.0	0.0	0.0	0.0
	易　度	0.4880	0.7120	0.8720	0.8240	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	区分度	0.0000	-.0667	0.0000	-.0667	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	平均易度		0.7240			平均区分度		-.0333			

成绩分布直方图



计算机分析结论	(注意：计算机分析结论，只提供参考) 1.学生得分最高分：76 最低分：57 平均分：69 学生成绩合理 2.试题比较简单 试题没有区分出不同类学生 3.第1小题过于简单或漏题;第2小题过于简单或漏题;第3小题过于简单或漏题;第4小题过于简单或漏题; 4.成绩呈偏负态分布，试卷难度偏低，难度较大题目少 5.成绩呈高狭峰分布，试卷同难度题目较多，梯度偏小 2008 年 12 月 1日
试卷存在的问题以及进一步改进措施	出题教师签字：　　教研室主任签字： 年 月 日　　年 月 日
院审核意见	教学秘书签字：　　院长签字： 年 月 日　　年 月 日
备注	(对试卷质量等未尽事宜在此说明)

本软件为内蒙古民族大学教务处研制　研制人：数学与计算机科学学院　张春生　　未经授权　不得使用　　联系电话：教务处 (0475)8313154　研制人：(0475)8314849

图 2 试卷分析报表

2.5 教师分析

教师分析主要通过易度、区分度、漏题情况、整体提分情况等对教师出题行为进行判断。

(1)试卷漏题可能性判断。通过易度 + (1 - 区分度)是否接近于 2 来判断是否存在漏题现象,若接近于 2 则怀疑教师漏题。

(2)敏感分数段提分。敏感分数段一般指 50 - 60 之间的分数,若在这个分数段提分,必将是 60 分的人数增多,50 - 60 分人数接近于 0。

(3)整体提分。当学生整体分数偏低时,教师可能采用 $10 * \sqrt{\text{原始分数}}$ 的方法进行整体提分,此时成绩分布呈偏负态,并且分布曲线在 40 分之前的斜率很大,40 分后平稳。

2.6 多维显示

多维显示主要利用圆角坐标和直角坐标进行同班各课程和同课程不同班之间比较,以图形的方式直观显示各种成绩的分布。

2.7 数据挖掘

数据挖掘技术自产生以来,作为一个新的学科,引起人们的重视,特别是 Internet 技术的发展,使得信息的存储量大规模增加,如何在大量的信息中发现潜在的知识,是当前研究的热点内容,目前数据挖掘应用的领域非常广泛,但在试卷分析方面的应用不多,文中在数据挖掘在试卷分析方面进行了以下尝试。

注:下面以本校 2003 级计算机专业 61 名学生成绩为例,table(a1,a2,a3,a4,a5),分别代表学生的数学成绩,专业课成绩,外语成绩,体育成绩,德育成绩。

2.7.1 纵横距离法异常成绩发现

E Knorr 和 R Ng 于 1998 年给出了基于距离的异常定义,一个异常数据是指在一个集合中,远离这个数据的数据较多,因此,所谓异常数据是指相对孤立的数据,即在其邻域内数据较少的数据,文中从学生的纵横成绩(一人不同科,同科不同人)角度进行异常成绩检测,更加合理地发现异常成绩[12]。纵横距离法程序界面如图 3 所示。

2.7.2 缺失成绩的修补

数据库中的数据不一定是完整的,存在一定的数据缺失,数据的不完整性严重影响了数据挖掘的效果,所以缺失数据的修补至关重要。

文中提出用三次参数样条曲线来拟合具有代表性的特征点,然后利用插值的方法实现缺失数据的修补。

2.7.3 关联规则发现

针对经典的 Aprior 算法缺陷,提出了一种基于事务的二元组表示法,该表示法所占内存大小只取决于数据库的基,而与数据库的大小无关,使得整个数据挖掘过程只进行一次数据库扫描,当数据库的基较小时,数据挖掘工作都在内存中完成[13](如图 4 所示)。

2.7.4 决策树分析方法

决策树在数据挖掘中是一个较好的分类挖掘算法,C4.5是决策树中的经典算法,文中以此为工具实

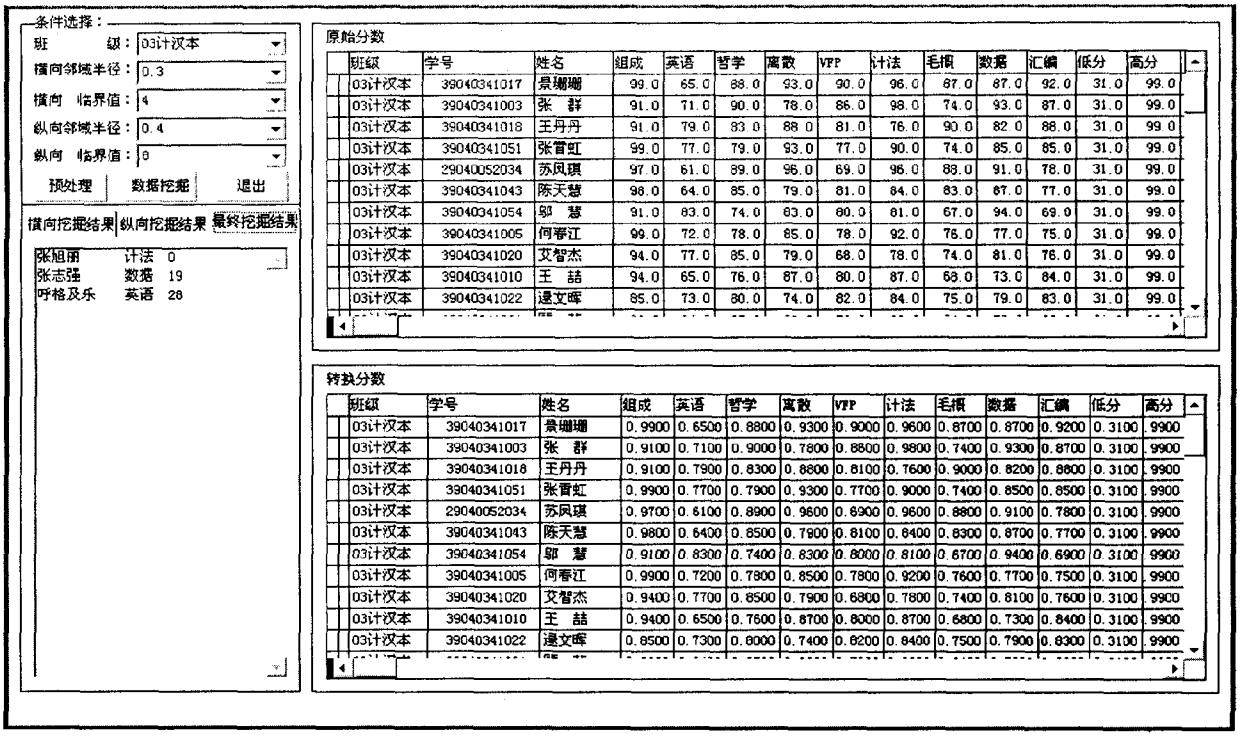


图 3　纵横距离法程序界面

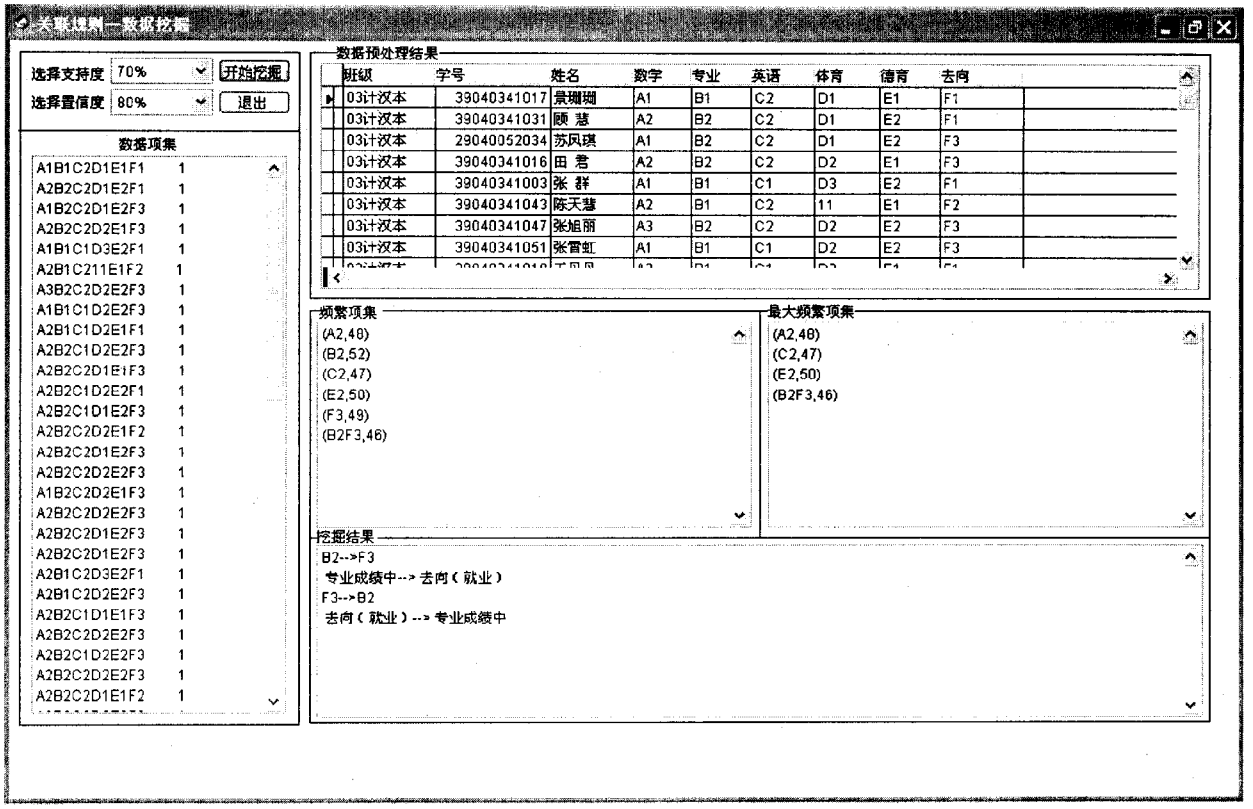


图 4 Aprior 数据挖掘结果

现分类挖掘,结果如图 5 所示。

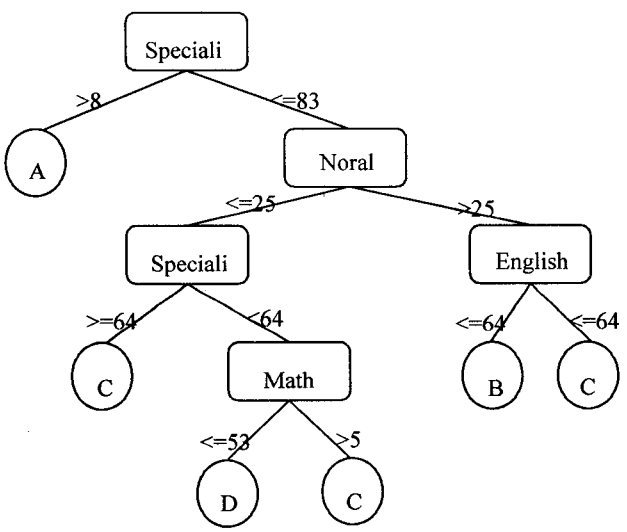


图 5 C4.5 挖掘结果

2.7.5 模糊聚类算法实现

模糊聚类是数据挖掘中常用的一种柔性聚类算法,可有效实现学生成绩的聚类分析,运行结果如图 6 所示。

2.8 系统维护

(1)数据库清空:清空数据库中的所有数据。

(2)数据库的备份:防止遭到数据库破坏,适时备份数据库。

(3)数据库的恢复:当数据库遭到数据库破坏,恢复数据库。

3 结束语

分析了目前试卷分析软件的现状,从试卷的各小题分数出发,通过精细粒度对学生的试卷进行了一般统计学分析,得到了详细的分析结果。同时,通过分析,自动对试卷质量给出评语,打印分析报表,利用多维显示技术,进行同班各科、同科各班的学生成绩比较,并能够进一步地对教师的行为进行分析。

利用数据挖掘技术,对学生进行异常成绩判断、关联规则分析、分类、聚类等等。

本系统分析粒度精细,分析面较广,挖掘层次较深,并充分利用数据挖掘技术,对成绩中的潜在规则进行了深入的挖掘,具有一定的实用价值和较广的应用范围。

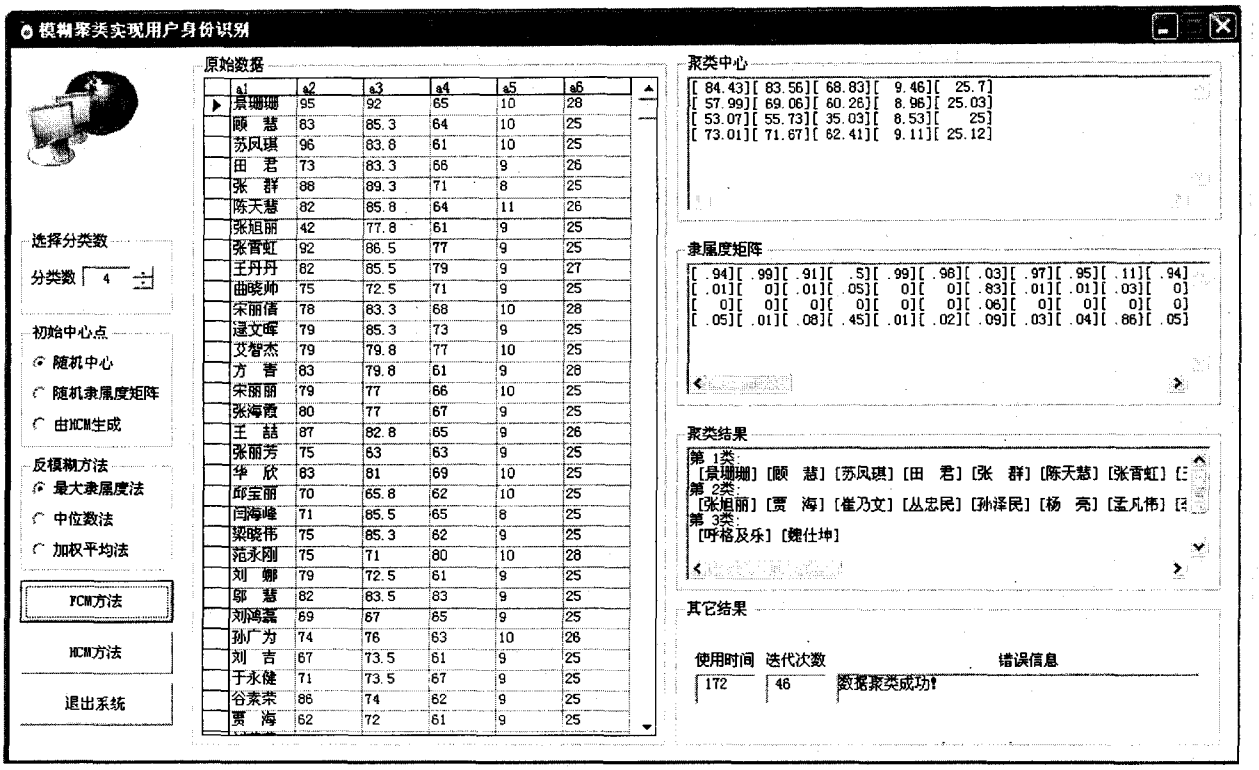
参考文献:

[1] 王孝玲.教育统计学[M].上海:华东师范大学出版社,1986.

[2] 田 文.教育学[M].哈尔滨:哈尔滨工业大学出版社,2000.

[3] 李 艳,张春生. 一种高校试卷分析系统的开发[J].内蒙古民族大学学报:自然科学版,2006(2):148-151.

[4] 刘 瑞.基于经典测试理论的试卷分析系统的设计与实



	a1	a2	a3	a4	a5	a6
▶	景珊珊	95	92	65	10	28
	顾　慧	83	85.3	64	10	25
	苏凤琪	96	83.8	61	10	25
	田　君	73	83.3	66	9	26
	张　群	88	89.3	71	8	25
	陈天慧	82	85.8	64	11	26
	张旭丽	42	77.8	61	9	25
	张雪虹	92	86.5	77	9	25
	王丹丹	82	85.5	79	9	27
	曲晓帅	75	72.5	71	9	25
	宋丽倩	78	83.3	68	10	28
	逯文辉	79	85.3	73	9	25
	艾智杰	79	79.8	77	10	25
	方　言	83	79.8	61	9	28
	宋丽丽	79	77	66	10	25
	张海霞	80	77	67	9	25
	王　喆	87	82.8	65	9	26
	张丽芳	75	83	63	9	25
	华　欣	83	81	69	10	25
	邱宝丽	70	65.8	62	10	25
	闫海峰	71	85.5	65	8	25
	梁晓伟	75	85.3	62	9	25
	范永刚	75	71	80	10	28
	刘　娜	79	72.5	61	9	25
	邬　慧	82	83.5	83	9	25
	刘鸿鑫	69	87	65	9	25
	孙广为	74	76	63	10	28
	刘　吉	67	73.5	61	9	25
	于永健	71	73.5	67	9	25
	谷素荣	86	74	62	9	25
	贾　海	62	72	61	9	25

图 6　FCM 运行结果

现[D].呼和浩特:内蒙古师范大学,2008.

[5] 张祥娟,代　君,张丽芬. 基于.NET 的试卷分析系统[J].九江学院学报,2008,27(3):30－35.

[6] 肖建清,丁德馨,徐　根,等.试卷分析系统开发中的关键技术[J].计算机工程与设计,2008,29(7):1847－1849.

[7] 张　瑶,陈高云,王　鹏.数据挖掘技术在试卷分析中的应用[J].西南民族大学学报:自然科学版,2008,34(4):839－842.

[8] 张顺清.如何进行试卷分析初探[J].化学教学,2008(6):61－63.

[9] 廖云霞.区分度在考试试卷分析中的应用[D].武汉:华中师范大学,2008.

[10] 朱永香,肖赞英,肖丹秦,等.试卷分析指标的选择及其应用[J].医学教育探索,2008,7(3):265－266.

[11] 李圣普,王小辉.智能试卷分析系统设计与实现[J].考试周刊,2008(11):2－3.

[12] 张春生. 基于纵横距离的单纯异常点检测算法及应用[J].内蒙古民族大学学报:自然科学版,2009(4):371－373.

[13] 张春生. 改进的数据库一次扫描快速 Apriori 算法[J].计算机工程与设计,2009(16):3811－3813.

(上接第 240 页)

neering Task Force,1981.

[2] 曾晶萍,杨文俊,彭　力.TCP 友好速率控制协议的分析与应用[J].计算机技术与发展,2007,17(1):210－212.

[3] 赵　飞,叶　震.UDP 协议与 TCP 协议的对比分析与可靠性改进[J].计算机技术与发展,2006,16(9):219－221.

[4] 罗明宇,卢锡城,韩亚欣. Internet 多媒体实时传输技术[J].计算机工程与应用,2000,36(9):119－120.

[5] 杜　恒. SCTP 流媒体传输性能分析研究[J].计算机与信息技术,2007,5(5):36－39.

[6] 尹　浩,林　闯,文　浩,等.大规模流媒体应用中关键技术的研究[J]. 计算机学报,2008,31(5):755－774.

[7] Stream Control Transmission Protocol[S]. RFC4960, Internet Engineering Task Force,2007.

[8] Real Time Streaming Protocol[S]. RFC2326, Internet Engineering Task Force,1998.

[9] Stewart R, Xie Qiaobing. Stream Control Transmission Protocol:A Reference Guide[M]. [s.l.]:Addison Wesley Longman, 2001.

[10] Ong L. An introduction to the transmission protocol[S]. RFC3286, Internet Engineering Task Force,2002.

[11] 朱桂勇,吴庆波.基于 SCTP 多宿特点的多路径同时传输研究[J].计算机技术与发展,2007,17(3):5－9.

[12] 夏　云,孙力娟,叶晓国.SCTP 协议分析与仿真研究[J].计算机技术与发展,2009,19(11):27－30.