Science Discovery
2017; 5(4): 287-292
http://www.sciencepublishinggroup.com/j/sd
doi: 10.11648/j.sd.20170504.18
ISSN: 2331-0642 (Print); ISSN: 2331-0650 (Online)

Research on Data Mining Technology Based on Weka Platform

Wang Pan Zao

Department of Information and Engineering, Sichuan Tourism University, Chengdu, China

Email address:
644464113@qq.com

To cite this article:
Wang Pan Zao. Research on Data Mining Technology Based on Weka Platform. *Science Discovery*. Vol. 5, No. 4, 2017, pp. 287-292.
doi: 10.11648/j.sd.20170504.18

Received: March 21, 2017; **Accepted**: May 18, 2017; **Published**: June 8, 2017

Abstract: Research uses Weka big data mining technology platform to analyze the data.The association rules miningmethods of discrete sample data by Weka technology,using SimpleKMeans clustering algorithm for clustering analysisof thesimulated sample data mining, common features of each type of data and the difference data between differentclusters from where, for a variety of data region division, analysis of different regions of the data distribution. For example, mining research project, a school in the college entrance examination scores data for the simulation sample, in Chinese, math and English college entrance examination scores as the object of analysis of large data mining, thepaper in science classes, the language, the total scores were compared with the distribution.The integrated use of statistical analysis and data mining technology, mining analysis on college entrance examination data deeply, get useful informationwith performance clustering, has strong theoreticalvalue, can help to the college entrance examination reform, give some guidance to high school education.

Keywords: Data Mining, Cluster Analysis, Weka Platform, College Entrance Examination

基于Weka平台的大数据挖掘技术研究

王攀藻

信息与工程学院，四川旅游学院，成都，中国

邮箱
644464113@qq.com

摘要：研究使用Weka大数据挖掘技术平台对数据进行挖掘分析。采用Weka的离散化技术对样本数据进行关联规则挖掘的方法，使用SimpleKMeans聚类算法对模拟样本数据进行聚类分析，从中挖掘每一类数据的共同特征以及不同簇间数据的区别所在，对各种数据区间划分，分析不同区域的数据分布。举例挖掘研究项目，以某校高考成绩数据为模拟样本，以语文、数学和英语高考成绩为分析对象，进行大数据挖掘研究，得出在文理科分班下，语、数、外总分成绩分别对比分布结果。综合运用统计分析和数据挖掘技术，深入地对高考成绩数据进行挖掘分析，获得以成绩聚类为主的潜在有用信息，具有较强的理论价值，能对高考模式改革起到帮助作用，对高考教育起到一定的指导作用。

关键词：数据挖掘，聚类分析，Weka平台，高考成绩

1．引言

当今社会是一个信息爆炸的时代。随着网络和信息技术的不断普及，人类产生的数据量正在呈指数级增长，大约每两年翻一番，根据监测，这个速度在2020年之前会继续保持下去，这意味着人类在最近两年产生的数据量相当于之前产生的全部数据量，大量数据源的出现导致了非结构化、半结构化数据爆发式的增长，信息数据的单位由TB-PB-EB-ZB的级别暴增，这些由人类创造的信息背后产生的这些数据早已经远远超越了目前人力所能处理的范畴，如何管理和使用这些数据，逐渐成为一个新的领域，这就是大数据[1]，大数据是一股席卷所有行业，领域和经济体的“破坏性力量”，不仅行业信息技术体系结构需要改变适应它，而且企业内部的部分也需要调整以便它能够提供信息和揭示洞察力。

教育、银行、电信和零售业等各行业已经对信息和数据的采集与分析格外看重。这些收集来的数据需要进行一定深度和广度的挖掘才能将信息数据转化为有用的知识和经济效益[2]。而单靠人力是很难完成对庞大的信息和数据进行分析的。因此，数据挖掘技术应运而生，该技术就是帮助人们从海量数据中提取有效的、隐含的、潜在有用的知识以优化和促进相应行业的信息化管理和发展。对于教育行业而言，教育信息化是教育教学模式转变和教育改革的重要依托和重要手段。而教育考试是教育教学过程中考核学生的重要环节，是学生学业评价的重要组成部分。因此，教育考试的信息化水平从某些层面上也体现着教育信息化的水平。据教育部报道，每年大量考生参加高考形成了海量考试数据，这些数据的背后隐藏着重要的信息和潜在的知识经济，如何有效挖掘这些信息和知识已成为教育研究者日益关注的问题。

针对这种情况，本文提出通过对高考成绩进行大数据挖掘分析方案，研究不同类别的高考学生的成绩，最终是否会呈现出，与我们主观分析判别不一样的结果，为高考教育改革，提供支持。本方案是通过系统得到模拟样本数据，以中国成都部分高中中学高三学生的数学、语文、英语三门高考成绩数据为分析对象，利用数据挖掘平台Weka 进行挖掘。主要进行关联规则挖掘和聚类分析，从一个新的角度尝试分析和挖掘高考数据，获取可用知识。

2．大数据挖掘技术

数据挖掘技术的飞速发展始于20世纪90年代中期，是代表美国及欧洲等发达国家社会进步的一个重要因素。近十几年数据挖掘在中国国内也逐渐兴起，已经在诸多领域得到了广泛的应用[3]，同时对数据管理、分析和信息智能采用了全盘考虑的方法，发挥出了大数据的巨大潜能。在各个行业，领先利用大数据的企业已能开创新的运营效率、新的收入流，差异竞争优势及全新的业务模式。企业已开始从战略角度考虑如何利用大数据为其发展提供支撑。

大数据的算法首先是J. B. MacQueen在1967年提出的K-means算法，是一种被广泛应用于科学研究和工业应用中的经典聚类挖掘算法[4]。其核心思想就是分类与距离，首先将样本中的n个数据对象按照一定的的方法先分成k份，再通过使用算法，确定一种使得每一个分类中的对所有的对象能够实现到该聚类中心的距离的平方和最小，这样就达到了聚类的目的[5]。K-means算法的工作流程如图1所示，过程是，输入：这需要自行输入聚类个数k，n个数据对象。输出：已经分好类的k个聚类。（1） 任意选取k个对象，这些选定的k个对象每个对象都要作为单独一个聚类的起始点，并且成为中心；（2）对于其他没有选定的对象，重复地逐次地正交计算其到第一步中选定的k个聚类中心所代表的类的距离，计算后进行比较，将对象分配到距离最近的聚类中；（3）在（2）完成以后，重新计算k个聚类的中心；（4）聚类中心结果与前一结果进行对照，如果聚类中心不同，那么下一步执行（2），否则下一步执行（5）；（5）向外输出程序执行的结果。

首先从n个数据对象中任意选择k个对象作为初始聚类中心；而对于所剩下的其它对象，则要一一计算他们与已经选定的中心的距离，并且根据这些计算结果将这些剩下的对象分到与他们最为相近的，即距离最小的聚类中心所在的类中。然后再计算每个所新聚类的聚类中心（该聚类中所有对象的均值）。不断重复这一过程，直到标准测度函数开始收敛为止。一般都采用均方差作为标准测度函数，具体定义如下：

$$E=\sum_{i=1}^{k}\sum_{p\in C_i}|p-m_i|^2 \qquad (1)$$

其中：

E—数据库中所有对象的均方差之和；

P—对象的空间中的某一个特定的点；

m_i—聚类C的均值（P和m_i均是多维的）。

公式（1）使得各聚类的各个元素之间获得最大的相似性，也能使得不同的类群之间有较大的差异。

3．Weka大数据挖掘平台

大数据需要一种可让业务和技术都获得竞争优势的新型分析平台，新平台借助于对海量数据集的更高级别处理能力，不仅能让企业不断发现大数据内含的深藏的具有可操作性的见解，还能实现与用户之间在网络环境里的无缝连接。

Weka数据挖掘平台是目前最为主流的通用数据挖掘工具之一，其强大的数据挖掘功能和集成的众多通用挖掘算法得到了许多理论研究者和应用开发人员的青睐。Weka全名也叫怀卡托智能分析环境（Waikato Environment for Knowledge Analysis），是现在几种主流的开源通用数据挖掘平台之一，也是现在发展最为完备的数据挖掘工具之一，其是基于JAVA环境下开源的机器学习（machinelearning）以及数据挖掘（datamining）软件，存储数据的格式是ARFF（Attribute-Relation File Format）文件，这是一种ASCII文本文件，二维表格存储在ARFF文件中。这也就是WEKA自带的“weather. arff” 文件，在WEKA安装目录的“data”子目录下可以找到。其优势在于包括多种数据预处理、分类、聚类、回归、关联规则、属性选择和

交互式数据可视化算法和技术，具有通用性和强大的算法合成等优势。

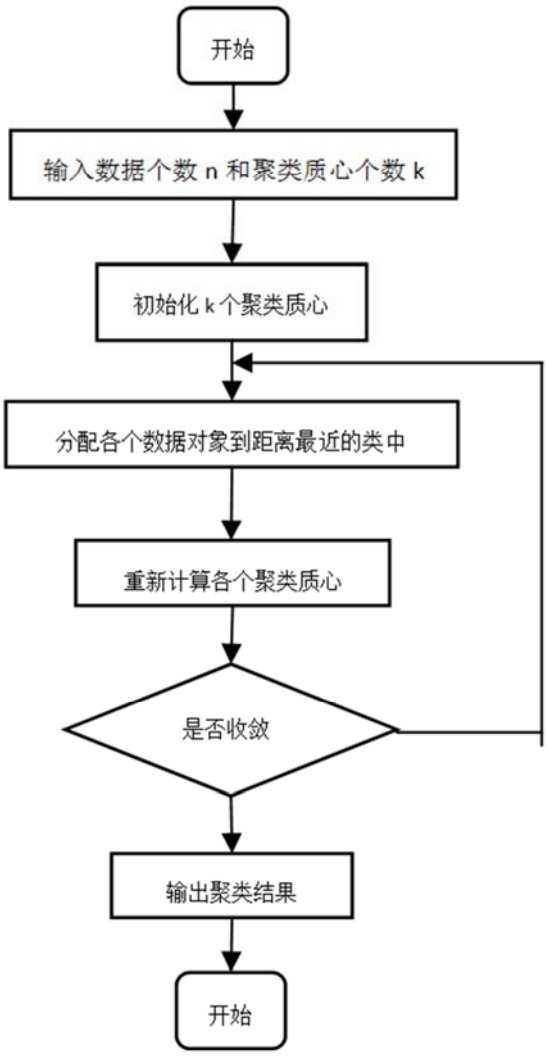

图1 K-means流程图。

研究的应用环境为Intel(R)Core(TM) i5 -3210MCPU @2.50GHz, 4G 内存，Windows7 专业版操作系统。安装 Weka3.6.9 版本。WekaGUIChooser提供了四个常用的应用入口：Explorer、Experimenter、Knowledge Flow和SimpleCLI。其中Explorer 是 Weka 挖掘数据的主要应用环境，提供了包括数据预处理、分类、聚类分析[6]、关联规则挖掘[7]、属性分析等在内的数据处理和挖掘过程。通过 Explorer 应用，结合系统算法过滤功能，可以对挖掘算法进行选择，配置算法参数，最后获得挖掘结果并用可视化工具对数据集进行可视化处理和呈现[8]。Experimenter是一个运行算法试验管理工具，该平台允许用户修改算法并进行测试验证。Knowledge Flow是一个“知识流”型的图形化数据挖掘接口，通过拖动工具栏中的各功能模块进行流程组合，最终形成数据挖掘知识流设计并实施挖掘过程，它能够完成部分 Explorer 无法实现的功能[9]。SimpleCLI提供了简单的命令行界面，允许用户通过命令来完成 Weka 所有的操作，它能够对 Weka 中的所有Java包和类进行操作[10]。

4．研究项目举例

Weka 数据挖掘平台默认数据存储的格式是 ARFF 文件，同时也支持 CSV数据格式。本项目研究挖掘分析中国成都高中高三年级高考语文、数学和英语三门成绩，在文理科分班的情况下，总分成绩各个分布情况。根据数据处理模式和需求，采用CSV格式数据，对语文、数学和英语高考数据进行挖掘和分析[11]。

4.1．数据导入

首先进入Explorer 应用模块，打开“高考数据.CSV”文件，将答题样本数据导入Explorer 进行预处理和挖掘，Filter 区域是Weka的过滤器，通过“Choose”选择相应过滤和数据变换算法并在右侧参数框中进行参数的设置，可对数据进行过滤和变换。

Explorer的当前数据关系（Currentrelation）区域给出了当前数据集的基本情况，包括实例个数和属性（字段）个数等，本例中有 416 条记录，176个属性。左下方的处理属性（Attributes）区域列出了从CSV文件中读取的所有属性。右侧的当前属性（Selectedattributes）区域则提供了当前选中属性的摘要，并将属性分布可视化显示出来以供参考。对于数值属性和分类属性给出的摘要也不同，数值属性摘要包括最小值、最大值、平均值和标准差等，而分类属性则会给出分类属性中的每个可能的取值和相应实例的总数。

4.2．深入挖掘

数据挖掘是从大量的、不完整的、有噪音的、模糊的、随机的数据中提取出隐含在其中的，事前不知道的，但又是潜在的有用的信息和知识的过程，致力于数据分析和理解。其处理对象是大量的日常业务数据，目的是为了从这些数据中抽取一些有价值的知识或信息，提高信息利用率，原始数据是形成知识的源泉。数据挖掘包括数据描述、聚类、分类、预测、孤立点分析、关联规则等多方面。其中数据描述又称为数据总结，目的是对数据进行浓缩，给出它总体的综合性描述，实现对原始数据的总体把握。常用数据描述方法是统计学的传统方法，如计算机数据项的总和、均值、所在比例、方差等基本描述统计量，或是绘制直方图、折线图等统计图形，研究已分类资料的特征，分析对象属性，据此建立一个分类函数或是分类模型，然后运用该函数或模型计算总结出据特征，将其他未经分类或新的数据分派到不同的组中，计算结果通常简化为及格离散值，常用来对资料做筛选工作[12]。

研究项目中导入的数据数学、语文、英语字段（Maths, Chinese, English）为例字段，在没有进行离散化处理前它们都是数值型属性，摘要区域显示：

数学最低分为50，最高分为118，平均得分 97.825，标准差为11.638。

语文最低分为53，最高分为125，平均得分103. 894，标准差为12. 426。

英语最低分为56，最高分为132，平均得分109. 454，标准差为13. 072。

右下方的数据可视化模块将数据、信息和知识转化为一种形象化的视觉形式，侧重人对数据、信息和知识自上而下的加工处理。相对于繁杂的数据，图表不仅能更加简洁地表述信息，还适用于大量信息的描绘，即对大量数据的承载。将各门课程得分分布以数据可视化柱状图的形式显现出来。以数学课程为例，如图2所示。

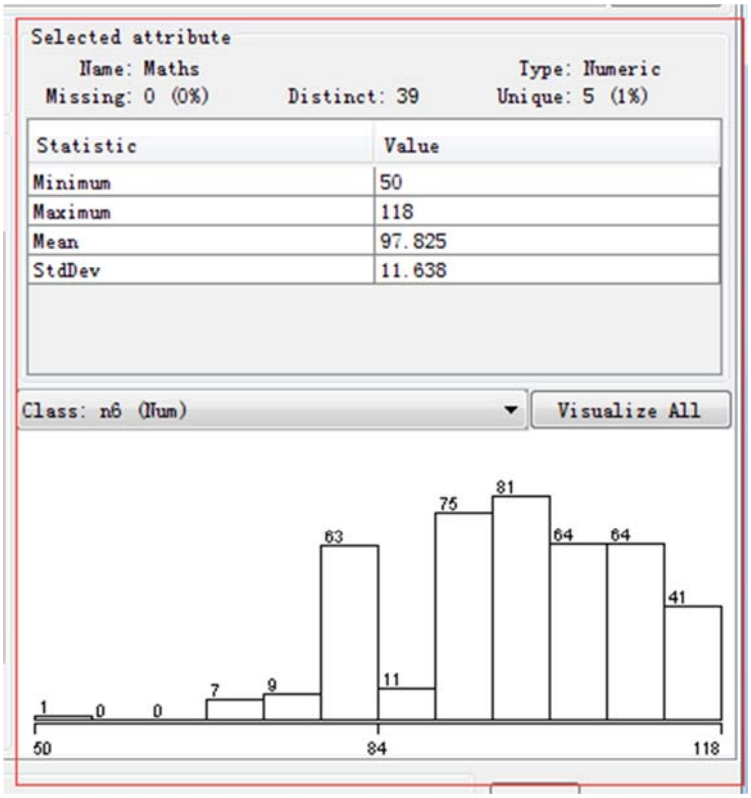

图2 柱状图呈现数学得分。

若数据集的最后一个属性是分类变量，直方图中的每个长方形就会按照该变量每个分层的比例分成不同的颜色段，在“Class”选择框中可以选择不同的分类变量。点击“VisualizeAll”按钮，可以得到所有属性字段的分布图。如图3所示。

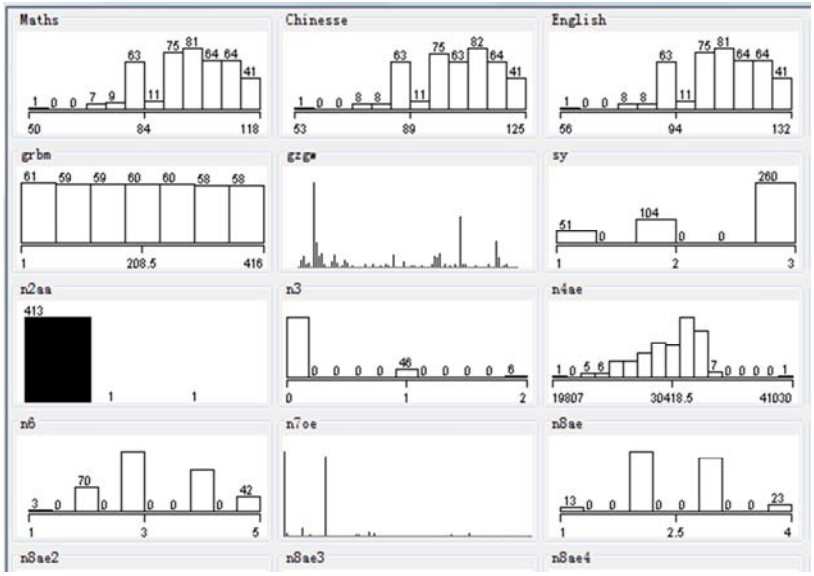

图3 所有属性字段分布图。

4.3．数据离散化

对于数值型属性，需要对数据做离散化处理，离散化又分为等距离散化和等频离散化，等距离散化是将连续变量划分为取值范围均匀、等距的n份，例如每10分一个得分区间，将学科总分划分为10个段；等频离散化则是从实例分布的角度来看，把某属性划分为均匀的区间，每个区间中分布的实例数相同。如将分数划分为五段，每个分数段中的考生数一致。本项目采用等频离散化方式，属性离散化处理后，之前的连续属性值将被所属的区间标记替代，数据离散化技术将使连续属性值的个数减少，这种效果非常很明显，例如在分数分布中，150 分的高考总分连续属性很可能拥有150个连续属性值，而完成 5 个区间的数据离散化处理后，仅仅需要 5个属性值来描述。通过收集较高层次的概念（如“优秀”、“良好”、“中等”、“合格”或“不及格”）并用以替换较低层次的属性值（如得分的数值）可以进行数据规约。通过这种处理，能够使数据更有意义、更容易理解，尽管此过程会丢失一些数据的细节。

在数据离散化模式下，以高考成绩数学（Maths）字段为例，采用无监督离散化处理，此时再来观察数学分数属性，可以发现相应实例已经被系统离散化到对应的五个容器。这五个容器实为 5 个离散数值区间，分别为“(-inf-64)”、“(64-77)”、“(77-91)”、“(91-104)”、“(104-inf)”，“inf”为上下限，也就是离散化之前的最大最小取值（分别为最小值50，最大值118）。也就是说，经过SimpleKMeans 聚类算法的处理，已将样本数据划分为5个簇类。输出结果中显示这些簇类的各种属性和参数，包括每个簇包含的实例数，每个簇中的样本所具有的高相似数据取值以及每个簇所含实例比率等。如图4所示。

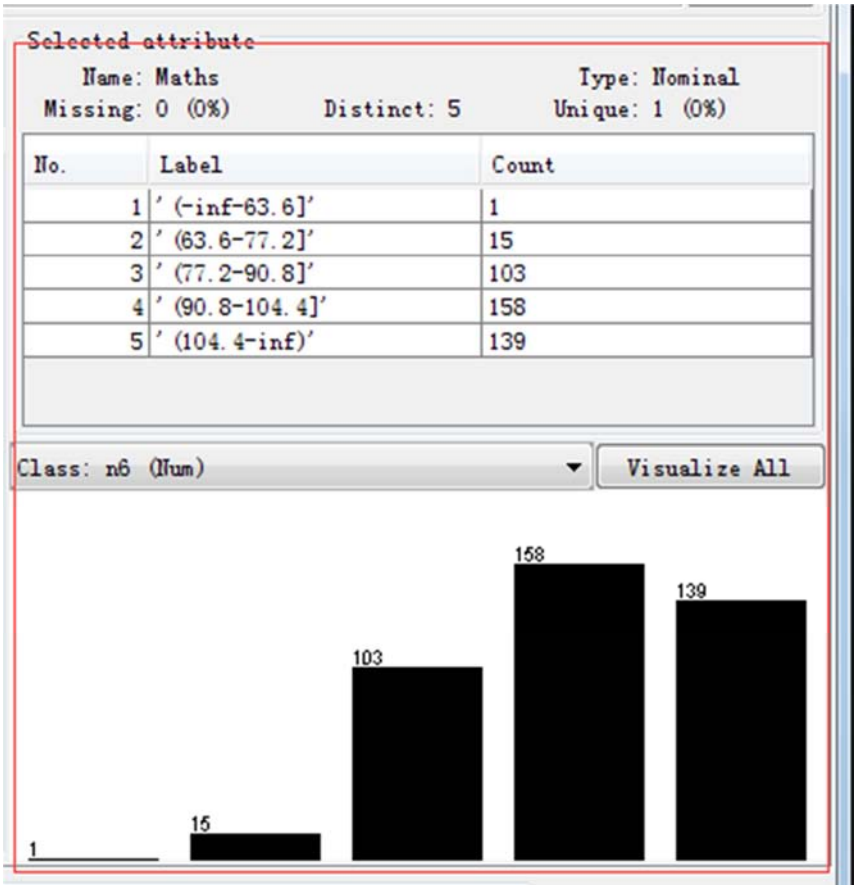

图4 离散化处理后的数学分数分布。

4.4．聚类分析

数据挖掘的关键突破口，将物理或抽象对象的集合分组为由类似的对象组成的多个类的分析过程，其目的是在相似的基础上收集数据来分类。聚类类似于分类，但与分类的目的不同，是针对数据的相似性和差异性将一组数据分为几个类别。属于同一类别的数据间的相似性很大，但不同类别之间数据的相似性很小，跨类的数据关联性很低。聚类与分类的不同还在于，聚类所要求划分的类是未知的。

使用SimpleKMeans聚类算法对模拟样本数据进行聚类分析，从中探索每一类考生的共同特征以及不同簇间考生的区别所在[13]。是一种典型的划分聚类算法，它用一个聚类的中心来代表一个簇，即在迭代过程中选择的聚点不一定是聚类中的一个点，该算法能处理数值型数据，利用Class分类变量的设置和Weka直方图的显示可以直观地对语文、数学和英语成绩总分进行分析。例如：选择文理属性（文理分班），得到如下图5所示的结果：

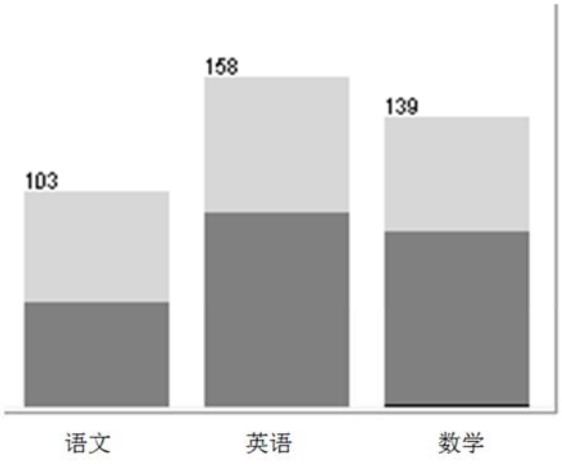

图5 以文理分班为分类属性的成绩分析。

图5中黑色代表理科班，白色代表文科文科班。从图中数值分布可看出，文、理科班的语文、数学和英语分布是不一样的。①理科班的语文成绩分数和文科班的语文成绩分数是非常接近的，说明在语文成绩这块，并不像人们主观认为文科生的语文成绩就一定远高于理科生的语文成绩。②理科班的英语成绩分数略高于文科班的英语成绩分数，英语成绩的挖掘分析结果，已经超出意料之外，根据以前的经验是文科的英语成绩是优于理科生的。③理科班的数学成绩分数高于文科班的数学成绩分数，这个符合我们平时的经验。数据挖掘成绩分布图，反映出每一类考生的共同特征以及不同簇间考生的区别所在。呈现出语、数、外高考成绩关系数据集的整体情况，通过挖掘结果分析，可以从中发现我们平时不易察觉的结果，为正确决策提供支撑保障。

5．结论

大数据挖掘技术为时下的许多重要领域提供了数据分析业务解决方案。考试数据管理是学校教育管理和实施的重要组成部分之一，是评估学校教育成果的重要方式，高考成绩数据的分析和信息化管理将极大地促进教育信息化。

本次大数据挖掘项目研究，应用考试数据分析系统获取模拟样本数据，主要获得中国部分高考成绩数据集合总成绩数据集，并进行初步预处理，为之后在系统中进行数据离散化等处理做出准备。在Weka中快速将全体考生的得分分布划分到不同区间（例如数学成绩的五个区间），以全面了解高考考试成绩分布状况。这种划分能够通过Weka工具的参数设置，选择不同的区间设置，进行综合对比。使用数据挖掘工具Weka的等频离散化技术，对离散化后的数据进行聚类分析。利用Class分类变量，文理分班的设置和Weka直方图显示直观地对成绩进行分析。

无监督离散化处理高考成绩样本数据后，SimpleKMeans聚类算法对模拟样本数据进行聚类分析，从中探索出每一类考生的共同特征以及不同簇间考生的区别所在。进行了关联规则挖掘和聚类分析，从一个新的角度尝试分析和挖掘高考数据，获取可用知识。应用考试数据分析系统获取模拟样本数据，主要获得高考成绩数据集合总成绩数据集，在挖掘平台中将考生的成绩划分到各个区间，从高考数据分析的角度来看，通过数据挖掘平台的此项功能能够快速对考生成绩情况进行区间划分。以考察文理不同属性的成绩分布，获得分析结果，以点带面从而了解高考考试成绩分布状况，对高考模式改革具有一定的帮助作用。而通过基于数据挖掘的高考数据分析应用，将能够对全省高考乃至更大范围的成绩数据进行全面分析应用。

本次挖掘研究因获取的考生信息有限，仅仅只能从部分信息（如文理分班等）进行分析。今后结合更多考生的属性，如性别、家庭情况、年龄等等进行挖掘分析，更全面地了解每一类考生的共同特征，从而发现哪些因素影响着考生的表现。

致谢

本文为四川省教育厅重点科研项目(编号16ZA0352)的阶段性成果之一。

参考文献

[1] 胡志伟. 关于大数据治理研究与分析 [J]. 物联网世界，2014（08）：58。

[2] 纪希禹. 数据挖掘技术实例 [M]. 机械工业出版社，北京：机械工业出版社，2009:102。

[3] 陶雪娇，胡晓峰，刘洋．大数据研究综述 [J]．系统仿真学报，2013(01):142-146。

[4] 张引，陈敏，廖小飞．大数据应用的现状与展望 [J]．计算机研究与发展，2013(02):216-233。

[5] 邬贺铨．大数据思维 [J]．科学与社会，2014(01)：76-77。

[6] JeffreyDean, Sanjay, Ghemawat.MapReduce:simplified data processing on large cluster [J]. Communications of the ACM, 2013, 51(1):107–113.

[7] Daniel J.Power.UsingBig Data for analytics and decision support [J]. Journal of Decision Systems, 2014(2): 78。

[8] Shi XiaLiu,Michelle Xzhou, ShimeiPan. TIARA [J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2012 (2):108。

[9] 程学旗，靳小龙，王元卓．大数据系统和分析技术综述 [J]．软件学报，2014(09)：123-124。

[10] 耿直. 大数据时代统计学面临的机遇与挑战 [J]. 统计研究，2014(02):89。

[11] 李金昌. 大数据与统计新思维 [J]. 统计研究，2014(01):167。

[12] 姚徐. 数据挖掘在计算机等级考试中的应用 [J]. 计算机光盘软件与应用，2013（01）：55。

[13] 柳玉巧．聚类分析和关联规则技术在成绩分析中的研究及应用 [D]．武汉：华中师范大学，2014。