Operations Research and Fuzziology 运筹与模糊学, 2016, 6(4), 115-121
Published Online November 2016 in Hans. http://www.hanspub.org/journal/orf
http://dx.doi.org/10.12677/orf.2016.64015

Hans汉斯

Application of Fuzzy Clustering and Data Mining in Data Analysis

Dongsheng Zhang*, Yongqiang Wang, Jing Su, Qingbo Wang

Software Collage, Henan University, Kaifeng Henan
Email: *act@henu.edu.cn

Received: Oct. 17th, 2016; accepted: Nov. 8th, 2016; published: Nov. 11th, 2016

Abstract

In view of the deficiency of general statistical methods, this paper presents the application of fuzzy clustering and data mining methods in data analysis. This paper introduces the basic principles of the two kinds of intelligent analysis methods, gives the experimental results of two kinds of analysis methods through specific cases, and compares their characteristics. In this paper, we put forward the advantages of data mining based on clustering results. It has a strong heuristic and engineering application reference value.

Keywords

Fuzzy Clustering, Data Mining, Data Analysis, Artificial Intelligence, Application Research

模糊聚类与数据挖掘在数据分析中的应用

张东生*，王永强，苏　婧，王青博

河南大学软件学院，河南　开封
Email: *act@henu.edu.cn

收稿日期：2016年10月17日；录用日期：2016年11月8日；发布日期：2016年11月11日

摘　要

针对一般统计方法的不足，提出模糊聚类和数据挖掘方法在数据分析中的应用。概要介绍了两种智能分

*通讯作者。

文章引用: 张东生, 王永强, 苏婧, 王青博. 模糊聚类与数据挖掘在数据分析中的应用[J]. 运筹与模糊学, 2016, 6(4): 115-121. http://dx.doi.org/10.12677/orf.2016.64015

析方法的基本原理，并通过具体案例给出两种分析方法的实验结果，比较了各自的特点。特别提出本文所作的先做聚类分析，再在聚类结果的基础上进行数据挖掘的优势。具有较强的启发性和工程应用参考价值。

关键词

模糊聚类，数据挖掘，数据分析，人工智能，应用研究

1. 引言

随着计算机与网络应用的日益普及和深入，各类大量数据被收集存储在不同的信息管理系统中。但“数据海量，信息缺乏”是这些管理系统普遍存在的尴尬现象，多数管理系统只能实现数据的录入、查询、统计等较低层次的功能，却无法发现数据中存在的各种有用的信息和知识。如果对这些数据进行更深层的分析，就可能发现其内在本质特征、所表达的模式、相互之间的关系及其发展变化趋势。人工智能技术中的聚类分析方法和数据挖掘方法能够很好地发现和揭示这些隐含在大量数据中的信息和知识，使数据发挥作用。

人工智能技术体系中，有许多技术方法可用于数据分析，本文选择使用模糊聚类和关联规则挖掘技术。这两种技术均具有较好的数学模型支撑和算法的科学性、稳定性，不同领域的研究者也提出许多改进方法。但在我国目前众多的信息管理系统中实际应用的情况并不多，其工程应用价值和方法还有待挖掘和推广[1]。

模糊聚类和关联规则挖掘技术已经被广泛应用在发达国家的军事、经济、金融、教育、和社会事务等多个行业中。譬如，可以成功预测银行客户需求，改善营销策略，提高综合收益；电子商务和网络购物网站使用关联规则进行挖掘，然后设置用户有意要一起购买的捆绑包，或设置相应的交叉销售；电信业通过聚类和挖掘技术帮助理解客户商业行为，识别电信模式，捕捉违法行为；教育领域则通过上述深层分析，发现教学过程存在的问题，以利于改善教学效果，提高教学水平[2]。

本文通过一个具体案例，给出模糊聚类和关联规则具体的应用方法和效果。

2. 案例数据与分析方法

2.1. 案例数据

实验数据来自某大学某课程考试。试卷包括 4 题(实验题号分别为 A、B、C、D)，每题 25 分。全体考生平均成绩 78 分，符合正态分布。为说明智能分析的意义，本文从分布于均值附近位置抽取 20 名考生的考试数据进行分析[3]。数据见表 1。

从数据情况来看，20 名考生卷面得分比较接近平均值 78 分，如果按照一般的统计方法分类，这些学生的学习情况属于一个大类——中等类。但是通过模糊聚类分析和数据挖掘分析却可以从中发现更多的信息和知识。

2.2. 模糊聚类分析

模糊聚类是聚类分析方法的一种，其基本方法是根据聚类对象两两之间的相似度组成的模糊相似矩阵，使得当矩阵元素值不低于给定阈值 λ 时转换为 1，否则转换为 0，然后将值为 1 的元素所对应的行、列对象归于一类[4]。模糊聚类法大致可分为三种[5]：一是基于模糊等价关系的传递闭包法；二是基于模糊相似关系的直接聚类法，包括最大树法和编网法；三是基于模糊 C-划分聚类法。三种方法均可较好地

Table 1. Case data
表 1. 案例数据

样本编号	A 题分	B 题分	C 题分	D 题分	总分
01	20	20	20	18	78
02	13	24	22	17	76
03	25	18	24	11	78
04	24	11	16	22	73
05	23	24	12	24	83
06	23	24	10	22	79
07	24	13	16	22	75
08	23	10	17	24	74
09	12	22	23	21	78
10	24	20	18	16	78
11	18	17	19	25	79
12	23	10	16	24	73
13	13	23	22	20	78
14	23	24	10	22	79
15	23	19	23	12	77
16	24	10	23	24	81
17	12	22	23	18	75
18	22	24	13	24	83
19	22	14	23	22	81
20	17	18	17	24	76

实现聚类运算，本文使用传递闭包法。

主要步骤包括：

(1) 整理原始数据。全部 n 个聚类对象的 m 个特征值构成原始数据矩阵 $X=\left(x_{ij}\right)_{n\times m}$。

(2) 原始数据标准化。为使不同量纲的数据可以相比较，通常需要将原始数据 x_{ij} 压缩至[0,1]区间，可通过极差变幻或标准差变幻实现。

(3) 构建模糊相似矩阵。分别计算样本 x_i 与 x_j 的相似度值 r_{ij} 组成模糊相似矩阵 $R=\left(r_{ij}\right)_{n\times n}$。相似度 r_{ij} 的计算方法有欧氏距离法、数据积法、相关系数法、夹角余弦法、最大最小法等多种[5]。本文使用相关系数法，其计算方法为：

$$r_{ij}=\frac{\sum_{k=1}^{m}\left|x_{ik}-\overline{x}_k\right|\left|x_{jk}-\overline{x}_k\right|}{\sqrt{\sum_{k=1}^{m}\left(x_{ik}-\overline{x}_k\right)^2}\sqrt{\sum_{k=1}^{m}\left(x_{jk}-\overline{x}_k\right)^2}}$$，其中 $\overline{x}_k=\frac{1}{n}\sum_{i=1}^{n}x_{ik}$。

(4) 计算传递闭包。多次进行矩阵 R 的自乘运算 $R\cdot R$，直到 $R^{2k}=R^k$ 为止，此时 R^k 称为 R 的传递闭包 $t(R)$。$t(R)$ 是模糊等价矩阵[6]。

(5) 截集取得聚类矩阵。取适当阈值 $\lambda\left(\lambda\in[0,1]\right)$，对模糊等价矩阵 $t(R)$ 作截集处理，求出聚类矩阵 $R''=\left(r_{ij}''\right)_{n\times n}$，其中：$r_{ij}''=\begin{cases}0, & r_{ij}'<\lambda\\ 1, & r_{ij}'\geq\lambda\end{cases}$。

将 r'_{ij} 为 1 的相应样本聚合为同一类，聚类完成。不难看出，λ 值选取越大，聚合出的类别数越多，选取得越小，则聚合出的类别数越少。但聚类结果并不矛盾：较粗类别是较细类别的上位类，利用 λ 取值不同，可获得不同程度的聚类，形成多层次分类结构。

2.3. 数据挖掘分析

数据挖掘是指从大量数据中通过一定算法搜索隐藏于其中信息的过程。本文使用关联规则 Apriori 挖掘方法。

Apriori 算法根据频繁项集的先验知识，使用迭代的方法通过逐层搜索实现挖掘的目标，即找出频繁项集，如超市购物中啤酒和尿不湿被同时购买的记录。基本原理是用 i 项集探索$(i+1)$项集。通过扫描事务(交易记录)，首先找到所有的频繁 1 项集，该集合记做 $L1$，然后利用 $L1$ 找频繁 2 项集的集合 $L2$，再用 $L2$ 找到 $L3$……直到 $L(k-1)$，却不能再找到任何 k 项集。最后再在所有的频繁集中找出强规则，即产生用户感兴趣的关联规则。

在关联规则挖掘方法中，通常会遇到一个总体性能的瓶颈，即搜索到的不总是真正的频繁项。Apriori 算法采用“连接步”和“剪枝步”两种方式来找出实际的频繁项集。

(1) 连接步。连接步的原则是保证前 $k-2$ 项相同，并按字典顺序连接。为找出所有的频繁 $k-1$ 项集的集合 $L(k-1)$，通过将 $L(k-2)$ (所有的频繁 $k-2$ 项集的集合)与自身连接产生候选 k 项集的集合。候选集合记作 Ck。设 $I1$ 和 $I2$ 是 $L(k-1)$中的成员。记 $Ii[j]$表示 Ii 中的第 j 项。由于 Apriori 算法对事务或项集中的项按字典次序排序，即对于$(k-1)$项集 Ii，$Ii[1] < Ii[2] < \ldots\ldots < Ii[k-1]$。将 $L(k-1)$与自身连接，如果$(I1[1] = I2[1])\&\&(I1[2] = I2[2])\&\&\ldots\ldots\&\&\ (I1[k-2] = I2[k-2])\&\&(I1[k-1] < I2[k-1])$，那认为 $I1$ 和 $I2$ 是可连接的。连接 $I1$ 和 $I2$ 产生的结果是$\{I1[1], I1[2], \ldots\ldots, I1[k-1], I2[k-1]\}$。

(2) 剪枝步。Ck 是 Lk 的超集，也就是说，Ck 的成员可能是也可能不是频繁的。通过扫描所有的事务(交易记录)，确定 Ck 中每个候选的计数，判断是否小于最小支持度计数，如果不是，则认为该候选是频繁的。为了压缩 Ck，可以利用 Apriori 性质：任一频繁项集的所有非空子集也必须是频繁的，反之，如果某个候选的非空子集不是频繁的，那么该候选肯定不是频繁的，从而可以将其从 Ck 中删除。

总起来讲，挖掘的过程为：(1) 扫描，(2) 计数，(3) 比较，(4) 产生频繁集，(5) 连接－剪枝，产生候选集。重复步骤(1)~(5)，直到不能发现更大的频繁集[7]。

3. 实验结果与分析讨论

本文所采用的两种分析方法都具有较好的稳定性和较快的收敛速度。两种分析的结果如下。

3.1. 模糊聚类实验结果

根据上节步骤进行仿真实验的聚类结果为：第 I 类：{02,09,13,17}，第 II 类：{04,07,08,12,16,19}，第 III 类：{05,06,14,18}，第 IV 类：{03,15}，第 V 类：{11,20}，第 VI 类：{01,10}。见表 2。

每一类簇特征项的值分布有明显不同。见图 1。

容易发现，第 I 类的特点是 A 题得分较低；第 II 类的特点是 B 题得分较低；第 III 类的特点是 C 题得分较低；第 IV 类的特点是 D 题得分较低；第 V 类的特点是 D 题得分较高；第 VI 类的特点是 4 个题得分比较均衡。

3.2. 数据挖掘实验结果

利用 Apriori 方法对表 2 数据进行挖掘分析(实验时从原始数据中添加了教师编号字段)。挖掘结果见图 2。

Table 2. Cluster analysis table
表 2. 聚类分析表

样本编号	A 题分	B 题分	C 题分	D 题分	所属类别
2	13	24	22	17	Clust-1
9	12	22	23	21	Clust-1
13	13	23	22	20	Clust-1
17	12	22	23	18	Clust-1
4	24	11	16	22	Clust-2
7	24	13	16	22	Clust-2
8	23	10	17	24	Clust-2
12	23	10	16	24	Clust-2
16	24	10	23	24	Clust-2
19	22	14	23	22	Clust-2
5	23	24	12	24	Clust-3
6	23	24	10	22	Clust-3
14	23	24	10	22	Clust-3
18	22	24	13	24	Clust-3
3	25	18	24	11	Clust-4
15	23	19	23	12	Clust-4
11	18	17	19	25	Clust-5
20	17	18	17	24	Clust-5
1	20	20	20	18	Clust-6
10	24	20	18	16	Clust-6

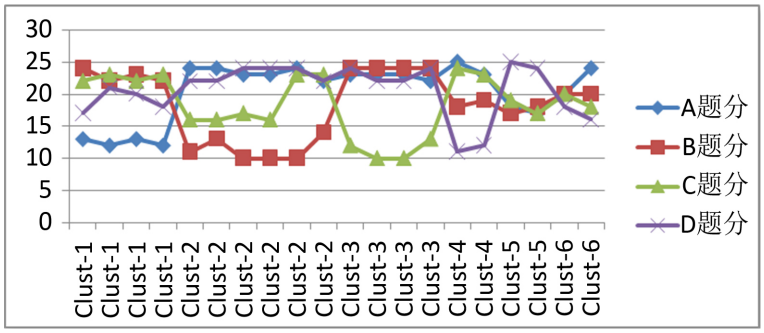

Figure 1. Cluster analysis
图 1. 聚类分析图

从图 2 看到，基于聚类的数据挖掘结果，共找到 20 条关联规则，排名靠前的规则是支持度和可信度较高的。以前 4 条关联规则为例，告诉了我们如下信息：

ques-B = 14.8-16. 7 ==> Clust = clust-2
Clust = clust-2 ==> Teacher = D6203

解读为：第 II 类考生 B 题得分偏低；而这类学生的任课教师为 D6203 (教师编号)。这一信息反映出 D6203 老师对于 B 题相关的教学内容的教学方面存在一定问题。

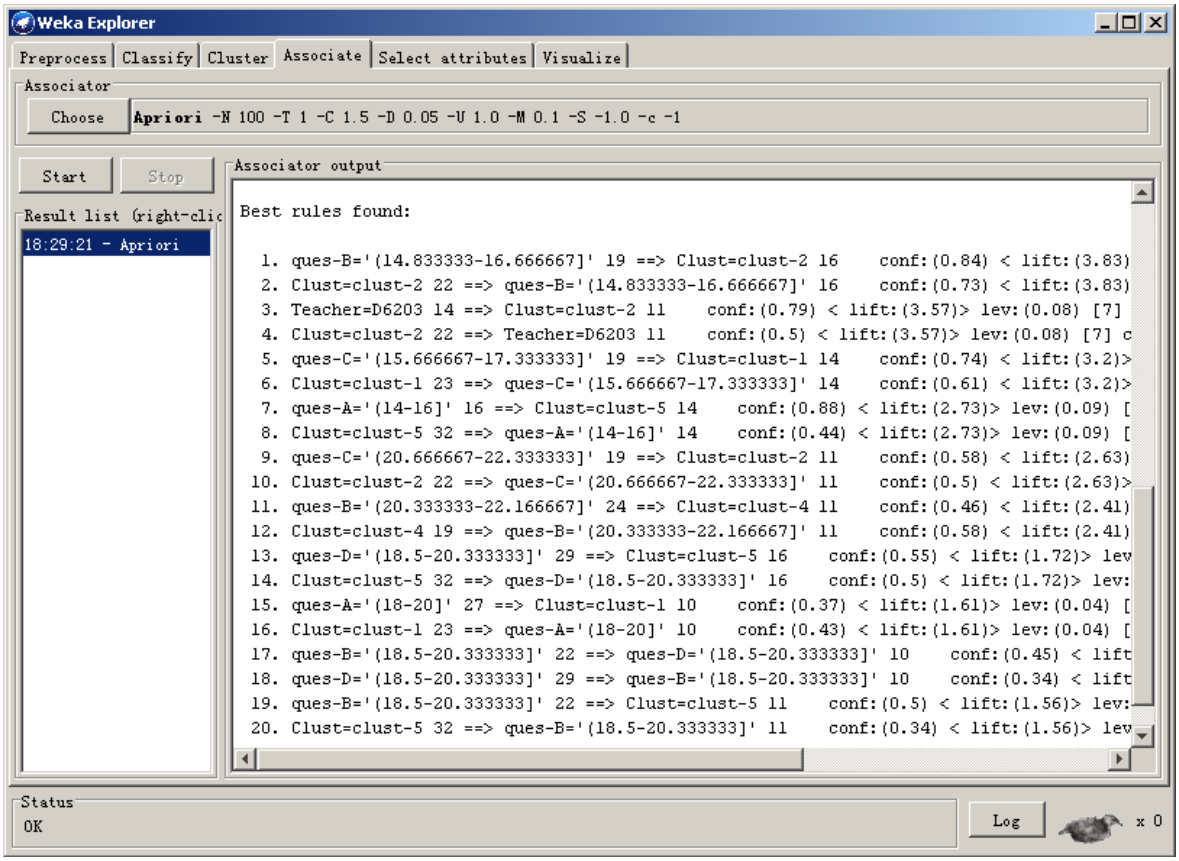

Figure 2. Data mining analysis
图 2. 数据挖掘分析图

3.3. 分析讨论

上述两种智能分析方法显示，聚类分析把看起来是同一类的考生区分出 6 类，6 个类各有不同的特征；而数据挖掘更是找到了每一类形成的一个或多个原因。这些分析功能都是一般统计学分析不能得到的。另外，本文中的关联规则挖掘是在样本取得聚类的基础上进行的，因此，不仅使挖掘得到有效的降维，降低了计算复杂性，而且挖掘的目标更为明确，所挖掘到的规则直接关联具体的类别，其指示意义更为明显和直接。这是不进行聚类分析而直接进行数据挖掘所不能达到的。

4. 结语

通过上述分析可知，一般统计分析方法，只能将本文案例中的考生样本归属于同一类(因为其卷面分数都接近均值)。通过智能分析，更加深刻和准确地发现每个考生知识点和能力点掌握情况的差异之处，以及形成的原因。这就为改进教学效果、提升教学水平和管理效能提供了有力帮助。换成其他数据分析，如企业运行数据、交通状况记录数据、患者检查指标数据等，具有同样的分析价值。

致　谢

感谢河南省高等教育教学改革研究项目和全国高等院校计算机基础教育研究会的资助。

基金项目

河南省高等教育教学改革研究项目“大学计算机基础分级分类教学模式研究(2014SJGLX143)”。

参考文献 (References)

[1] 崔妍, 包志强. 关联规则挖掘综述[J]. 计算机应用研究, 2016, 33(2): 330-334.

[2] Keller, A. (2000) Fuzzy Clustering with Outliers. *Proceedings of* 19*th Conference North American Fuzzy Information Processing Society*, Atlanta, IEEE Press, Piscataway, 143-147.

[3] 张东生, 张缨. 一种带有显著特征项的模糊聚类算法[J]. 河南大学学报(自然科学版), 2011, 41(2): 184-187.

[4] Looney, C.G. (1999) A Fuzzy Clustering and Fuzzy Merging Algorithm. University of Nevada, Redo, NV89557.

[5] 杨淑莹. 模式识别与智能计算——Matlab 技术实现[M]. 北京: 电子工业出版社, 2015: 271-298.

[6] 梁保松, 曹殿立. 模糊数学及其应用[M]. 北京: 科学出版社, 2007: 59-62.

[7] 吕安民, 李成名, 林宗坚. 基于空间统计分析的关联规则应用研究[J]. 计算机科学, 2002, 29(4): 53-54.