Operations Research and Fuzziology 运筹与模糊学, 2017, 7(4), 170-176
Published Online November 2017 in Hans. http://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2017.74018

Hans汉斯

The Combining Technology of Data Mining Based on Clustering and Association Rules

Han Li, Dongsheng Zhang*

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

Received: Nov. 9th, 2017; accepted: Nov. 21st, 2017; published: Nov. 30th, 2017

Abstract

Although clustering analysis and association rules as two main application methods can achieve data mining, but both two methods have three different. The data type of clustering operation is continuous and association rules are discrete. Clustering reflects the description function of the mining and association rules reflect prediction/validation function. The output form of clustering is clusters, and association rules then output the lines of rule. At the same time, both of them have some complementary to each other. So, this paper combined the both methods. The clustering analysis for the set of samples was first executed. This processing will make samples for their respective category entity information. Then, run association rules mining according to the samples what with classification properties. The method show the potential knowledge further including causes of the formation of clustering and the relationship between clusters. The experiment shows that the mining technology has better effect and great value of application.

Keywords

Clustering, Association Rules, Data Mining, Machine Learning

聚类联合关联规则的数据挖掘技术

李　涵，张东生*

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

收稿日期：2017年11月9日；录用日期：2017年11月21日；发布日期：2017年11月30日

*通讯作者。

文章引用：李涵，张东生．聚类联合关联规则的数据挖掘技术[J]．运筹与模糊学, 2017, 7(4): 170-176.
DOI: 10.12677/orf.2017.74018

摘　要

尽管聚类分析和关联规则作为两个主要应用方法都可以实现数据挖掘功能，但两者存在三大差异，聚类的数据类型为连续型，关联规则为离散型；聚类体现挖掘的描述功能，关联规则体现预测/验证功能；聚类的输出形式为类簇，关联规则输出的是规则。两者同时具有一定的互补性。因此，本文将两者结合起来，先对样本集进行聚类分析，使样本实体获得各自的类别信息；再对这些带有分类属性的样本进行关联规则挖掘，使得挖掘运算有效降维且具有更好的挖掘目标，挖掘结果可以清晰地显示聚类形成的原因和聚类之间的关系等潜在知识。实验表明，本文介绍的联合挖掘技术可以取得更好的挖掘效果，具有很大的实用价值。

关键词

聚类，关联规则，数据挖掘，机器学习

1. 引言

数据挖掘是从大量的数据中挖掘出隐含的、未知的、用户可能感兴趣的和对决策者有潜在价值的知识和规则[1]。常用的数据挖掘方法主要有以下几种，分类与聚类分析方法、统计方法、偏差分析方法、决策树与回归树方法、关联规则方法等[1]-[7]。本文讨论将聚类分析与关联规则两种方法结合应用的技术，以实现更好的数据挖掘效果。

聚类分析是研究数据之间物理的或逻辑的相互关系的技术，通过一定的规则将数据集划分为在性质上相似的数据点构成的若干个类簇。聚类分析的结果可以揭示数据之间的内在联系与区别，发现数据库中分布的一些深层的信息与知识，进一步研究，可以概括出每一类的主要特征。也可以把着眼点放在某些特定的类上进行进一步的分析[8] [9]。

关联规则反映一个事物与其他事物之间的相互依存性和关联性。如果两个或者多个事物之间存在一定的关联关系，那么，其中一个事物就能够通过其他事物预测到。关联规则挖掘就是为了在数据集中发现这些关联关系，是数据挖掘技术中最先提出的问题之一，也是数据挖掘的一个主要研究方向。关联规则由 Agrawal、Imielinski 和 Swami 在 1993 年提出[2] [10]，次年，Agrawal 和 Verkamo 提出了关联规则挖掘的经典算法 Apriori [11]。本文后面章节的关联规则实验，即采用了该算法的改进版。

2. 聚类联合关联规则的挖掘技术

聚类分析和关联规则是数据挖掘中两个非常重要且具有各自代表性的典型方法——聚类分析主要实现数据挖掘的描述功能；而关联规则主要实现数据挖掘的预测/验证功能。

2.1. 聚类分析

聚类分析是一种寻求数据的自然聚集结构的重要方法，增强了人们对客观现象的认识。聚类应用的意义，主要表现在处理大量的、繁杂的、属性众多且没有类标志的数据。这些没有类标志的数据经过聚

类处理后，将根据其内在特征的相似性，自动聚集为若干类簇，类内对象相似度较大，而类间对象相似度较小。

聚类分析的基本方法是，同类样本的离差平方和应当较小，而类之间的离差平方和应当较大。假定已经将 n 个样本分成了 k 个类 $C_1, C_2, \cdots, C_k$，用 x_{it} 表示 C_t 中的第 i 个样本的特征值向量，n_t 表示类 C_t 中的样本个数，$\bar{x}_t$ 表示 C_t 的重心，则 C_t 中样本的离差平方和为：

$$S_t = \sum_{i=1}^{n_t} \left(x_{it} - \bar{x}_t\right)' \left(x_{it} - \bar{x}_t\right)$$

全部类内离差平方和为：

$$S = \sum_{t=1}^{k} S_t = \sum_{t=1}^{k} \sum_{i=1}^{n} \left(x_{it} - \bar{x}_t\right)' \left(x_{it} - \bar{x}_t\right)$$

当 n 很大时，要给出全部样本所有可能的聚类，并从中选择出使 S 达到极小的聚类方案是极其困难的。于是，Ward 提出了这种聚类方法，采用离差平方和法，样本之间的距离采用欧氏距离法[12]。聚类分析实现的算法现在已经有很多，本文采用了模糊聚类和人工神经网络聚类等两种方法[13] [14]。

聚类结果是使数据挖掘具备识别群功能。

2.2. 关联规则

关联规则是描述数据库中数据项之间存在潜在关系的规则。设 $I = \{i_1, i_2, \cdots, i_n\}$ 为全体数据项集合，则关联规则可以形式化定义为：$X \Rightarrow Y$，其中 $X \subseteq I, Y \subseteq I$，且 $X \cap Y = \varnothing$。项集之间的关联表示：如果 X 出现在一条交易中，则 Y 在这条交易中同时出现的可能性比较高。

“可能性比较高”的界定方法，则采用支持度和置信度来表述：

规则 $X \Rightarrow Y$ 的支持度定义为 X 和 Y 同时出现的可能性，表示为 $\Pr(X \cup Y)$；规则 $X \Rightarrow Y$ 的置信度定义为全体事务集 D 中包含 X 的同时也包含 Y 的可能性，表示为 $\Pr(X \cup Y)/\Pr(X)$。当支持度和置信度的值都大于给定的相应阈值时的规则称为关联规则[1] [8]。

下面给出关联规则的基本算法 Apriori 的伪代码[15]：

```
L[1]={large 1-itemsets};
for (k=2; L[k-1]≠Φ; k=k+1) do
    C[k]=apriori_gen(L[k-1]);      //构造候选项集
    for all transactions t∈D do
    C[t]=subset(C[k], t);
                //搜索事务 t 中包含的候选项集
    for all C∈C[t] do C.sup=C.sup+1; end for
                //计算支持数
    end for
    L[k]={ C∈C[k] | C.sup>=minsup};
                //得到 K 阶大项集
end for
L=U[k] L[k];
```

其中候选项集的生成是 Apriori 算法的核心，通过 Apriori_gen 函数运算实现。描述如下：

```
insert into C[k]
select P[1], P[2], ···, P[k − 1], Q[k − 1]
```

from L[k − 1] P, L[k − 1] Q

where P[1]= Q[1], ···, P[k − 2] = Q[k − 2], P[k − 1] < Q[k − 1]

对构造的候选项集进行削减：如果 k 阶候选项集 C 的某个 k-1 阶子集不中 L[k − 1]中，那么 C 就不可能是大项集，需要将其从候选项集 C[k]中删除。

for all itemsets C ∈ C[k] do

for all (k − 1) itemsets S of C do

if (S∉L[k − 1]) then delete C from C[k]

关联规则可以发现聚类之间的关系，挖掘出样本和聚类之间的关联规则和潜在知识。

2.3. 联合运用

一般地，聚类分析中，样本的属性值是连续型的；而关联规则挖掘中样本的属性值是离散型的。二者对样本数据的处理方法和分析结果的输出形式有很大差异性和互补性。表 1 对本文所采用的两种聚类方法和一种关联规则方法进行了比较。

从表中容易发现，将聚类分析与关联规则结合起来，可以取得更好的挖掘效果，后面的实验完全证明了这一点。

二者联合运用的具体方法是，先对样本集进行聚类分析，通过聚类把整个样本集分成不同子集，使样本实体获得各自的类别信息；再对这些带有分类属性的样本进行关联规则挖掘，使得挖掘运算有效降维且具有更好的挖掘目标。

3. 实验数据与方法

3.1. 样本数据

用于编程实验的数据来自河南大学本科生的某次考试(http://218.196.195.205/admin/ks/ vbks.asp)。试卷包括 4 个大题(题号分别以 A、B、C、D 标识)，每题满分 25 分，卷面分值 100 分。全体考生平均成绩 77.9 分，符合正态分布。不失一般性，本文实验中随机抽出得分比较接近均值的 100 名考生的考试数据进行挖掘分析。样本数据参见表 2。

Table 1. Function contrast of clustering and association rule
表 1. 聚类与关联规则功能对比

方 法	适合任务	适合数据	可理解性
神经网络	聚类，分类	连续	差
模糊聚类	聚类	连续	中
关联规则	关联	离散	好

Table 2. Sample data
表 2. 样本数据

学号	A 题	B 题	C 题	D 题	总分
01	16	21	23	19	79
02	19	20	17	22	78
03	22	17	21	16	76
⋮	⋮	⋮	⋮	⋮	⋮
99	14	24	20	22	80
100	22	20	19	18	79

3.2. 数据变换

先将样本数据整理成便于聚类的形式，例如，将原始数据中比较复杂的学号和题号替换为容易运算的符号，然后进行标准化变换。本文使用了离差变换和标准差变换[9]。

3.3. 聚类分析

对变换后的样本数据分别进行模糊聚类和自组织神经网络聚类，然后运用 F 检验，自动取得最佳聚类方案[13]。实验中，两种聚类方法获得的最佳聚类结果完全一致，均为 5 类。聚类结果参见图 1。

3.4. 关联规则挖掘

经过 3.3 所述的聚类分析之后，再对已具备类别(图 1 中最右列)的样本数据进行关联规则挖掘分析，将使挖掘运算更为方便，且规则指向性更明确、更容易理解。

本文采用改进的 Apriori 算法进行关联规则分析，输出相应的关联规则，参见图 2。

4. 结果与讨论

聚类结果将 100 个考生样本分为 5 类，其中第 1 类 23 个，第 2 类 22 个，第 3 类 4 个，第 4 类 19 个，第 5 类 32 个。通过表 3 的比较，大致可以了解每类的主要特征。

stuID	Name	Prof	Teach	ques-A	ques-B	ques-C	ques-D	Clust
200912	刘浩	经济	史蒂	21	19	19	20	clust-5
200912	庄明	经济	张东	19	21	20	15	clust-4
200912	李昂	经济	何洪	18	19	21	15	clust-4
200913	郭明	法学	武澎	20	15	17	22	clust-1
200913	张洋	法学	武澎	20	23	15	18	clust-1
200913	王冰	法学	王金	23	24	16	17	clust-4
200913	尹旺	法学	武澎	23	15	21	21	clust-2
200914	陈轩	教育	白晨	18	22	22	18	clust-4
200914	孙少	教育	侯松	17	16	23	24	clust-3
200914	郭延	法学	王金	22	15	18	23	clust-2
200915	周凯	文学	万敏	22	13	22	20	clust-2
200916	刘强	文学	袁洁	20	20	20	20	clust-5
200916	王奇	文学	楚艳	21	15	22	20	clust-2
200916	耿鹏	文学	王红	14	18	20	18	clust-5

Figure 1. Cluster analysis
图 1. 聚类分析图

```
Associator output

Best rules found:

  1. ques-B='(14.833333-16.666667]' 19 ==> Clust=clust-2 16
  2. Clust=clust-2 22 ==> ques-B='(14.833333-16.666667]' 16
  3. Teacher=D6203 14 ==> Clust=clust-2 11    conf:(0.79) < 1
  4. Clust=clust-2 22 ==> Teacher=D6203 11    conf:(0.5) < li
  5. ques-C='(15.666667-17.333333]' 19 ==> Clust=clust-1 14
  6. Clust=clust-1 23 ==> ques-C='(15.666667-17.333333]' 14
  7. ques-A='(14-16]' 16 ==> Clust=clust-5 14    conf:(0.88)
  8. Clust=clust-5 32 ==> ques-A='(14-16]' 14    conf:(0.44)
  9. ques-C='(20.666667-22.333333]' 19 ==> Clust=clust-2 11
 10. Clust=clust-2 22 ==> ques-C='(20.666667-22.333333]' 11
 11. ques-B='(20.333333-22.166667]' 24 ==> Clust=clust-4 11
```

Figure 2. Data mining results of association rules after clustering
图 2. 聚类后进行的关联规则数据挖掘结果

Table 3. Clustering results analysis of ample data
表 3. 样本数据聚类结果分析

类号	A 题均值	B 题均值	C 题均值	D 题均值	4 题均值
1	19.74	19.09	16.17	21.13	19.03
2	20.68	15.86	20.05	19.00	18.90
3	18.36	18.64	20.55	20.82	19.59
4	19.79	20.95	20.05	15.89	19.17
5	15.72	21.68	19.48	20.00	19.22

根据表中数据容易发现，第 1 类考生 C 题得分较低；第 2 类考生 B 题得分较低；第 3 类考生四个题得分均匀；第 4 类考生 D 题得分较低；第 5 类考生 A 题得分较低。但这只是对聚类意义的大致解读，缺乏准确和全面的理解。

与文献[16]等许多基于聚类的分析实验相比较，那些只做到本步骤层面的分析，并不能直接得到具有知识层面的信息和情报，尚需专家对聚类结果进行人工解析才可以理解聚类分析的意义。

基于聚类的关联规则挖掘分析，则将在聚类的基础上得出一系列更为明确和直接的分析结果。例如，在图 2 中，挖掘结果的前 4 条规则就明确给出了如下关联规则：

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

Clust = clust-2 ==> Teacher = D6203

其意义解释为：

第 B 题得分介于 14.8~16.7 (偏低)的考生，被归入“clust-2”类；而“clust-2”类的任课教师是编号为“D6203”的老师。

这一规则明确提示我们，编号为“D6203”的教师在第 B 题的教学方面存在明显问题，需要改正。

如果继续使用关联规则对相关数据集进行挖掘，可能找出“D6203”老师在 B 题教学方面存在问题的原因，从而为督促该教师改善和提高教学效果提供有力的技术依据与支撑。

同时，由于关联规则挖掘是在样本取得聚类的基础上进行的，因此，不仅使得挖掘得到有效降维，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使用关联规则所不得达到的。

5. 结语

按照传统和粗放的考试成绩分析方法，本文所分析的 100 位考生应属于同一类(成绩都接近均值)，但聚类分析却可以通过每个样本属性的特征值，更加深刻和准确地根据每个考生知识点和能力点掌握情况的差异之处，并将其划分为若干类，为进一步挖掘类之间的关系打下基础；而在聚类之后进行的关联规则挖掘，则更进一步发现了聚类形成的原因和聚类之间的关系等潜在的知识。聚类分析和关联规则的联合运用取得了更好的挖掘效果。本文所述实验大部分已经过多个大样本集的实际挖掘应用，实践证明，聚类分析与关联规则联合挖掘技术具有稳定有效的应用价值和非常广阔的应用前景，值得进一步研究推广。

基金项目

感谢河南省教师教育课程改革研究项目(2017)的资助。

参考文献 (References)

[1] 陈安, 陈宁, 周龙骧. 数据挖掘技术及应用[M]. 北京: 科学出版社, 2006.

[2] Agrawal, R., Imielinski, T. and Swami, A. (1993) Database Mining: A Performance Perspective. *IEEE Transactions on Knowledge and Data Engineering*, **5**, 914-925. https://doi.org/10.1109/69.250074

[3] 夏姜虹. 数据挖掘技术的常用方法分析[J]. 云南大学学报(自然科学版), 2011, 33(S2): 173-175.

[4] 张连育, 吕立. 基于策略模式的中医数据挖掘平台的设计与研究[J]. 小型微型计算机系统, 2011, 32(7): 1406-1411.

[5] 孙中祥, 彭湘君, 杨玉平, 贺一. 数据挖掘在教育教学中的应用综述[J]. 2012, 2(1): 78-80.

[6] 戴汝为. 社会智能科学[M]. 上海: 上海交通大学出版社, 2007.

[7] 张东生, 王永强, 苏靖, 等. 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): 7

[8] Agrawal, R. and Srikant, R. (1995) Mining Sequential Patterns. *1995 Proceedings of the Eleventh International Conference on Data Engineering*, Taipei, 6-10 March 1995, 3-14. https://doi.org/10.1109/ICDE.1995.380415

[9] 张东生. 基于模糊聚类的考试分析方法[J]. 电脑知识与技术, 2009, 5(33): 9579-9580.

[10] 李雪梅, 张素琴. 数据挖掘中聚类分析技术的应用[J]. 武汉大学学报(工学版), 2009, 42(3): 396-399.

[11] 徐辉增. 关联规则数据挖掘方法的研究[J]. 科学技术与工程, 2012, 12(1): 60-63.

[12] 王爱平, 王占凤, 陶嗣干, 等. 数据挖掘中常用关联规则挖掘算法[J]. 计算机技术与发展, 2010, 20(4): 105-108.

[13] 张东生, 季超. 动态模糊聚类及最佳聚类效果研究[C]. Proceedings of Chinese Conference on Pattern Recognition (CCPR), Beijing, 4-6 November 2009.

[14] Zhang, D.S., Li, S.Z. and Wei, W. (2010) Visual Clustering Methods with Feature Displayed Function for Self-Organizing. Industrial Mechatronics and Automation. https://doi.org/10.1109/ICINDMA.2010.5538274

[15] 郭涛, 张代远. 基于关联规则数据挖掘 Apriori 算法的研究与应用[J]. 计算机技术与发展, 2011, 21(6): 101-103.

[16] 武森, 俞晓莉, 倪宇, 王瑞峰. 数据挖掘中的聚类技术在学生成绩分析中的应用[J]. 中国管理信息化, 2009, 12(15): 45-47.