基于灰色关联分析的 Apriori 算法的研究及应用

赵永进,林 卫,贺娜娜,王振华
(河南师范大学 计算机与信息工程学院,河南 新乡 453007)

摘 要:关联分析是一种重要的数据挖掘技术。文中结合房地产行业的特点,将关联分析方法应用于对消费者购房行为的研究中。传统的关联规则挖掘算法-Apriori 算法在实际应用中存在着计算量大、挖掘效率低、产生大量不相关的关联规则等问题。为了减少计算量、提高挖掘效率、发现有价值的关联规则,提出了一种灰色关联度分析算法和 Apriori 算法结合的研究方法。首先采用灰色关联度分析算法得出影响消费者购房需求和偏好的关键因子,然后采用 Apriori 算法对关键因子和目标因子之间进行关联规则挖掘。以某市问卷调查的消费者信息记录进行建模,结果表明该关联分析方法具有较高的挖掘效率并且研究结果具有合理性和准确性。

关键词:购房行为;灰色关联度分析;Apriori 算法
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2013)11-0255-03
doi:10.3969/j.issn.1673-629X.2013.11.062

Research and Application of Apriori Algorithm Based on Gray Relational Analysis

ZHAO Yong-jin, LIN Wei, HE Na-na, WANG Zhen-hua
(College of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China)

Abstract: Association analysis is an important data mining techniques. In this paper, combined the characteristics of the real estate industry, the association analysis method is applied to the study of consumer purchase behavior. The traditional association rule mining algorithm-Apriori algorithm has large calculation, the low efficiency of mining, resulting in a large number of irrelevant association rules in practical applications. In order to reduce computation, improve the efficiency of mining and find valuable association rules, the combination method of gray relational analysis algorithm and Apriori algorithm is presented. First of all, gray relational analysis algorithm to draw the key factors affecting consumer purchase requirements and preferences, and then use the Apriori algorithm for mining association rules between the key factor and the target factor. The modeling results show that the association analysis method has a high efficiency of mining and the findings has the reasonableness and accuracy of consumer information records of a city survey.

Key words: purchase behavior; gray relational analysis; Apriori algorithm

0 引 言

Apriori 算法是一种以概率为基础的挖掘关联规则频繁集的算法。该算法利用由多到少、从简单到复杂的循序渐进方式,搜索数据库的项目之间的相关关系,并利用概率的表示形成关联规则。但是 Apriori 算法存在着计算量大,挖掘效率低,产生大量不相关、不正确甚至是没有价值的关联规则等问题[1-5]。灰色关联度分析用于定量地描述灰色系统中各因素之间,在发展过程中随时间而相对变化的情况[6-11]。文中针对这些问题提出灰色关联度分析算法和 Apriori 算法结合挖掘关联规则的方法,并将该方法应用于对消费者购房行为的研究中,发现数据变化的规律,减少了算法的计算量、提高了挖掘效率、发现有价值的关联规则,从而得出有价值的结论。

1 基于灰色关联度分析的 Apriori 算法

针对单一的 Apriori 算法可能产生庞大的候选项目集,挖掘出没有价值的冗余规则。文中提出将灰色关联度分析算法和 Apriori 算法结合以提高算法效率的方法,从而减少算法的计算量、挖掘用户感兴趣的知识、提高了挖掘效率。

首先,对数据进行灰色关联分析。灰色关联度分

收稿日期:2012-12-04 **修回日期**:2013-03-12 **网络出版时间**:2013-07-24
基金项目:河南省基础与前沿技术研究计划项目(102300410102);河南省教育自然科学研究计划项目(2010B520011)
作者简介:赵永进(1978-),男,河南新乡人,硕士,讲师,研究方向为数据挖掘;林 卫,博士,副教授,研究方向为机器学习、模式分析。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20130724.0953.016.html

析算法的输入是选择需要进行关联分析的样本矩阵 $\boldsymbol{X}=\{x_0,x_1,x_2,\cdots,x_m\}$，其中 x_0 是要进行分析的目标项。它的数据结构采用样本矩阵 = 样本点个数 × 属性集项数，记为 $\boldsymbol{X}=\{x_0,x_1,x_2,\cdots,x_m\}$，以二维数组的形式 x 出现。而相应的输出是各个项对目标项的关联度由小到大排序的结果，找出 $\boldsymbol{X}_{\max}$。

用灰色关联度分析算法对相应数据进行以下处理：首先对矩阵 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_m\}$ 中求出各列的均值，分别用以上均值去除矩阵中各原始数列得均值化数列。然后计算各比较数列 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_m\}$ 同目标数列 x_0 在同一时期的绝对差的参考数列为：$\{x_0(t)\}=\{x_{01},x_{02},\cdots,x_{0m}\}$，将第 k 个比较数列（$k=1,2,\cdots,m$）各时间点的数值与参考数列对应时间点差值的绝对值进行比较，得到所有 m 个比较数列在各时间点的绝对差值中的最小者 $\Delta(\min)$ 和最大者 $\Delta(\max)$，再根据 $\zeta_{0k}=\dfrac{\Delta(\min)+\rho\Delta(\max)}{\Delta_{0k}(t)+\rho\Delta(\max)}$ 计算第 k 个比较数列与参考数列在 t 时期的关联系数，并从中找出最大值 max 和最小值 min。然后，当参考数列的长度为 n 时，由 m 个比较数列共可得到 $n\times m$ 个关联系数，构成相应的矩阵，在此基础上分别求各个数列每个时期的关联系数的平均值即 $r_{0k}=\dfrac{1}{n}\sum_{i=1}^{n}\xi_{0k}(t)$ 可得关联度。最后进行关联排序，找出关联度最大的项，记为 $x_{\max}$。

在灰色关联度分析算法得出影响目标项的首要影响项的基础上接下来对新的项集 $X_1=\{x_0,x_{\max}\}$ 采用 Apriori 算法挖掘关联规则。Apriori 算法的输入是数据集 $X_1=\{x_0,x_{\max}\}$，最小支持度阈值为 min_sup。它的输出是 X_1 中的频繁项目集 L。相应的处理步骤如下：首先是读入所有事务数据 X_1，得出候选 1_ 项集合 C_1 和相应的支持度数据；然后是将每个 1_ 项集的支持度和 min_sup 比较，得出频繁 1_ 项集合 L_1；最后是将频繁 1_ 项集两两进行连接，产生候选 2_ 项集合 C_2，与 min_sup 比较，得出频繁 2_ 项集合 L_2。通过以上步骤可以看出灰色关联度分析产生的关联规则频繁集大大减少了。

2 算法在研究消费者购房行为中的应用

文中对消费者住房需求问卷调查进行了分析。实例中所用指标主要分为两大类：消费者基本信息指标包括性别、年龄、婚否、文化程度、职业、平均月收入等；消费者需求及偏好指标包括未来 3 年有无购房打算、购房类型、购房目的、住房最关注问题、购房面积、购房类型等。

Apriori 算法对样本的属性要求为非数值型，故首先对建模的 16 个属性指标进行离散化处理。

由于此次数据采集采用问卷调查的方式进行，故收集的原始数据都是以字母选项的形式出现，然而文中采用的灰色关联度分析算法仅能处理数值问题，与 Apriori 算法数据类型产生矛盾，为了一致所以将字母编码对应为整数值编码，即 A—1、B—2、C—3…以此类推，这样在处理过程中不同的整数值就代表不同的属性的类型。

分析中需要的指标经过初步处理后的部分数据如表 1 和表 2 所示。

表 1 部分数据预处理（消费者基本信息指标）

编号	Q1.1	Q1.2	Q1.3	Q1.4	Q1.5	Q1.6	Q1.7	Q1.8	Q1.9	Q1.10	Q1.11
1	1	1	2	2	6	2	2	6	1	3	2
2	1	2	1	3	4	5	1	3	3	3	1
3	2	3	1	3	4	5	1	3	2	4	3
4	2	3	1	2	8	1	2	7	3	3	1
5	1	4	1	5	4	5	1	3	3	2	3
6	1	2	3	5	3	3	4	5	3	3	1
7	2	1	2	3	4	5	3	2	1	3	5
8	2	2	3	3	5	5	2	3	3	3	3
9	2	2	1	4	6	2	2	4	2	5	3
10	2	3	1	1	1	4	1	5	3	3	1

表 2 部分数据预处理（消费者需求及偏好指标）

编号	Q2.1	Q2.2	Q2.3	Q2.4	Q2.11
1	3	5	1	3	3
2	3	2	1	2	3
3	1	2	1	5	3
4	3	3	1	4	1
5	3	2	1	4	1
6	1	2	2	6	2
7	2	1	1	3	3
8	3	2	1	1	3
9	1	2	3	5	3
10	1	4	1	2	1

经过预处理的样本数据，运用灰色关联度分析算法，以购房类型为目标项，以 Q1.1 ~ Q1.11 预处理结果为影响因子项，作为输入，计算这 11 方面对目标项的关联度如表 3 所示。通过算法运行可以得到影响消费者选择何种购房类型的最大的因素是消费者月平均收入即 Q1.8。

表 3 Q1.1—Q1.11 对购房类型的关联度

因子项	关联度
R01	0.758 0
R02	0.772 5
R03	0.725 7
R04	0.687 5
R05	0.737 1
R06	0.698 2
R07	0.683 0
R08	0.795 6
R09	0.774 5
R010	0.726 5
R011	0.652 8

Apriori 算法的数据结构分为如下几个方面:

第一部分为数据项的处理。采用字符表示形式存储项 x_0 和 x_{max} 中对应的具体元素 $x_{01},x_{02},\cdots,x_{0n}$ 和 $x_{max1},x_{max2},\cdots,x_{max}$。如消费者现阶段居住情况($C$、已有房)、购房目的($A$、自住) 可以用字符表示为 '$c$'、'1';

第二部分记录的处理。由于每一个数据项由字符组成,每一条记录由若干数据项组成,所以每一条记录的表示用两个字符的集合表示如$\{x_{01},x_{max1}\}$。如某位消费者现阶段居住情况(C、已有房) 和购房目的(A、自住) 可以表示为{'c','1'};

第三部分为数据集合的处理。数据集合由若干条记录组成,用一个二维字符数组来表示,其中行数代表记录条数即事务数目,列数表示数据项个数,如 char[n][2] = {{'x_{01}',' x_{max1}'},{'x_{02}',' x_{max2}'},…,{'x_{0n}',' x_{max}'}};

第四部分为频繁集的处理。关联规则的形式化表示为 $X => Y$,一条规则无非就是包含这么几个成员,规则的条件、结论、支持度等。在该算法中,用输出形式控制产生频繁集的表示,如 1_频繁集{1} = 14,表示选择数据项'1' 的支持度为 14,2_频繁集{1,c} = 13,表示关联规则 1 => c,支持度为 13。然后以消费者平均月收入 Q1.8 和购房类型 Q2.2 建立事务集,其中 Q1.8 的数据用字母字符表示,Q2.2 的数据用数字字符表示,带入 Apriori 算法,设置 min_sup = 4,得到如下频繁集,如表 4 和表 5 所示。

表 4　1_项频繁集

频繁项集	支持度计数
1	5
2	35
3	8
5	6
6	5
C	22
D	16
E	6
F	8
G	7

表 5　2_项频繁集

频繁项集	支持度计数
< 2,c >	21
< 2,d >	6
< 3,g >	4

结合实际情况分析可以知道:月收入水平决定着一个人购房类型,一般而言,收入水平高的会选择配置较高的高档住宅或者别墅,而收入属于一般水平的会选择大众类型的普通住宅。

这就说明运用灰色关联度分析算法和 Apriori 算法结合可以挖掘出比较准确、有价值、用户感兴趣的知识。

3　结束语

文中将基于灰色关联度分析算法的 Apriori 算法结合挖掘关联规则应用于对消费者购房行为的研究中,得到了一些有价值的结果,可以对消费者购房起到一定的指导意义。

由于房地产市场是一个复杂的系统问题,还需要结合实际的情况进行分析与决策。

参考文献:

[1] Aggarwal C, Yu P. Data mining techniques for personalization [J]. IEEE Data Engineering Bulletin, 2000, 23(1): 4-9.

[2] 侯雪波,田　斌,葛少云,等. 关联规则技术在电力市场营销分析中的应用[J]. 电力系统及其自动化学报, 2005, 17(2): 67-72.

[3] 朱晓峰,李玲娟,徐小龙,等. 基于 MapReduce 的关联规则增量更新算法[J]. 计算机技术与发展, 2012, 22(4): 115-118.

[4] 陆　楠,王　喆,周春光. 基于 FP-Tree 频集模式的 FP-Growth 算法对关联规则挖掘的影响[J]. 吉林大学学报:理学版, 2003, 41(2): 180-185.

[5] 刘　辛,杨素锦. 基于数组的 Apriori 算法在体质测试数据分析中的应用[J]. 山东理工大学学报:自然科学版, 2011, 25(5): 55-58.

[6] 孙芳芳. 浅议灰色关联度分析方法及其应用[J]. 公路与管理, 2010(17): 364-366.

[7] 刘　宏,吴　江,耿国华,等. 基于灰色关联分析的高感兴趣度数据挖掘算法研究[J]. 计算机工程与设计, 2008, 29(16): 4242-4244.

[8] 李　楠,宁燕子,杨存志,等. Apriori 关联规则算法的 C 语言实现[J]. 大连民族学院学报, 2011, 13(1): 52-55.

[9] 杨　峰,吴明慧. 多维多层次挖掘关联规则在商品房交易中的应用[J]. 信阳师范学院学报:自然科学版, 2004, 17(3): 345-347.

[10] Han J, Kamber M. DataMining: Concepts and techniques[M]. San Francisco: Morgan Kaufmann Publisher, 2001.

[11] Weiss A. Computing in clouds[J]. ACM Networker, 2007, 11(4): 18-25.

基于灰色关联分析的Apriori算法的研究及应用

作者： 赵永进， 林卫， 贺娜娜， 王振华， ZHAO Yong-jin， LIN Wei， HE Na-na， WANG Zhen-hua

作者单位： 河南师范大学 计算机与信息工程学院,河南 新乡,453007

刊名： 计算机技术与发展

ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2013(11)