《智能信息处理》课程作业

基于形式概念分析的柔性决策规划

智泽镕

作业	分数[20]
得分	

2021 年 11 月 29 日

基于形式概念分析的柔性决策规划

智泽镕

（大连海事大学 信息科学技术学院，大连 116026）

摘 要 关联规则获取是知识发现和数据挖掘中的核心问题之一。对超市来讲，从交易数据中挖掘出的关联规则有两点重要意义：一是有助于设计商品的摆放位置；二是帮助商品进货搭配规划，为更好利用关联规则进行进货搭配规划，知识工程师不仅需要考虑关联规则的可信度、支持度和兴趣度，更需要考虑支持集对关联规则的贡献度和关联规则自身的平衡度和复杂度。本文首先采用形式概念分析理论挖掘交易数据中的关联规则，这些规则具有 100%的可信度。然后，在关联规则柔性筛选的基础上进行商品进货决策规划。所谓柔性是指用户可自己定义规则的不同阈值组合（例如析取和合取）选择规则。

关键词 形式概念分析，关联规则，柔性筛选方法

Flexible Decision Planning Based on Formal Concept Analysis

ZHI Ze-Rong

(School of Information Science and Technology, Dalian Maritime University, Dalian 116026)

Abstract One of the core tasks of knowledge discovery and data mining is the mining of association rules. For some supermarket, they mainly focus on the following two important significances of association rules acquired from a lot of trade-off data: one is that they can be used to arrange the places of correlative products on the shelves and the other is that they can be used to plan harmonious stocks of the correlative products- In order to settle the aims, knowledge engineers not only should take into account confidences, supports and interest degree of association rules but focus on contribution degree of support set and balance/complexity of association rules. In the paper, the theory of Formal Con-cept Analysis are firstly used to acquire association rules with confidence 100% from a large of tradeoff formal context, and then discussed the products planning based on flexible filtering association rules Flexibility means that users can define different thresholds combination of rules (e-g- disjunction and conjunction) for choosing satisfactory rules.

Keywords Formal concept analysis, Association rule, Flexible filtering method

1 引言

关联规则获取是知识发现和数据挖 掘的核心内容之一[1-6]。关联规则是类似于“80%的买可乐的顾客都买了牙膏和香皂”40%的顾客同时买了这三样物品）的陈述句。其中,80%和 40%分别称为该规则的可信度和 支持度。形式上，关联规则是形如“A-B”的条件式。规则获取的任务就是从大量历史数据中挖掘出用户感兴趣的大于最小可信度和最小支持度的关联规则。许多学者给出各种各样的关联规则挖掘算法，例如联想规则发现 Apriori 算法网和基于形式概念分析的挖掘方法[4,7,8]。

对大型超市来讲，从交易数据中挖掘出的关联规则有两点重要意义:一是有助于设计商品的摆放位置;二是帮助商品进货搭配规划。因此，关联规则挖掘是一个很实用的

研究领域。对于关联规则的评价主要有三个指标:可信度、支持度和兴趣度 13 向。但是，若考虑到商品的搭配进货，这三个指标是不够的，因为这些指标都没有考虑到支持集对规则的贡献，即没有考虑商品的数量。另外，还需要考虑规则本身的复杂性和平衡性。对于规则“A-B”（可信度 10％）,该规则的复杂度和平衡度分别为|AUB|/|M|和|A|/|B|,其中 M 是形式上下文中所有商品 的集合。复杂度和平衡度的直观含义是:在支持度固定的情况下，复杂度越小说明商品搭配进货越简单，越大说明越困难;在支持度和复杂度固定的情况下，平衡度越小说明更少的商品集决定更多的商品集，也就是说商品搭配进货管理越简单，越大说明越困难为了讨论上述问题，我们采用形式概念分析 (Formal Concept Analysis, 简写为FCA) 理论获取关联规则。这些规则具有 100%的可信度。FCA 是数据分析和知识发现的有效工具, 在数据分析中得到广泛应用[7,8]。为了利用 FCA 理论, 我们把超市中的交易事件看作对象, 把商品看成属性, 把交易事件中购买的数量看成是交易事件在属性上的属性值。例如, 可乐 (交易 2) =4 瓶说明 2 号顾客购买了 4 瓶可乐。然后, 用户可以根据自己定义的阈值组合进行规则的柔性筛选。最后, 计算所选规则的质心, 进行商品搭配进货决策规划。

本文第2部分介绍形式概念分析的几个重要概念和抽取算法;第 3 部分讨论关联规则分析和柔性决策设计;最后, 总结全文。

2 形式概念分析的基本概念

为保证论文的可读性和完整性, 本节介绍一些形式概念分析中的重要概念和获取算法, 详细请看文[7]。

定义 1 (形式上下文) 一个形式上下文是一个三元组 (G, M, I) 。其中, G={g1, …, gn}是有限的对象集, M={m1, …, mk}是有限的属性的集合, I⊆G×M 上的二元关系。如果 (g, m) ∈I, 则称对象 g 有属性 m。

为表示超市交易数据样本, 可以用一个交叉表来表示, 其中行表示交易事件 (即对象) 名称, 列表示商品 (即属性) 名称, 行和列的交叉处表示对应的商品量 (见图 2) 。直观含义为“顾客 g 购买了 w? (?表示超市中常用单位) 商品 m”。

定义 2 (形式概念) 设 (G, M, I) 是个形式上下文, 称序对(A, B) 是 (G, M, I) 的一个形式概念, 若满足条件:1) A⊆G, B⊆M; 2) A′={m∈M|∀x∈A (x, m) ∈I}=B;3) B′={g∈G|∀y∈B (g, y) ∈I}=A。

若 (A, B) 是一个形式概念, 则称 A 是概念外延, B 是概念的内涵。

定义 3 (伪内涵) 称 B⊆M 是 (G, M, I) 的一个伪内涵, 若满足条件:1) B≠B″;2) ∀P⊂B⊆M (P″⊆B)

定义 4 (集合间的字母序关系) 设 A, C⊆G, 称 A 是按字母序小于结合 C, 如果能区分集合 A 和 C 的最小元素在集合 C 中。形式上, A<iC:=∃i ∈ C ＼ A (A∩{1, 2, …, i-1}=C∩{1, 2, …, i-1}) 。

定义 5 (集合闭操作) 给定集合 B, 定义闭操作Γ#如下:BΓ#:=B ∪ (∪ {C|A→C∈Γ, A⊂B};BΓ#Γ#:=BΓ# ∪ (∪ {C|A→C ∈ Γ, A⊂BΓ#}, 继续下去能发现一个集合Γ# (B) 满足Γ# (B) =Γ# (B) Γ#。

算法 1 关联规则抽取算法

输入:形式上下文(G, M, I),其中 $G=\{g_1, \cdots, g_K\}$和 $M=\{m_1, \cdots, m_N\}$;$=\{1, 2, \cdots, N\}$
输出:形如 $X\to Y$ 的关联规则集,其中 $X, Y\subseteq M$
过程:
(1) $\Gamma_i:=\{\}$
(2) $B=\Phi$
(3) FOR $i=N$ to 1 DO
(4) IF $B<_i \Gamma^{\#}((B\cap\{1, 2, \cdots, i-1\})\cup\{i\})$
(5) THEN $\Gamma_i\leftarrow\Gamma\cup\{B\to B^{\#\#}\backslash B\}$
(6) $B_i\leftarrow\Gamma^{\#}((B\cap\{1, 2, \cdots, i-1\})\cup\{i\})$
(7) ELSE $i_i\leftarrow i-1$
(8)对 B 和 Γ重复步骤(2)~(7)
(9)当没有新的伪内涵出现时,算法结束,输出 Γ_i /Γ就是关联规则集合/

3 基于关联规则分析的柔性规则筛选

3．1 基于形式概念分析的规则获取

下面举例说明上述算法。假设给定一个超市的销售信息(部分)(见图 1)。首先把信息表转化成对应的形式上下文(见图 2),交叉处是商品的数量(数量单位以超市的销售单位为准),然后利用上述的获取算法获取关联规则集合(此时不考虑商品的数量)。算法开始于空集中,最小的伪内涵是{可乐},{可乐}'={2,5,8,9,13}, {2,5,8,9,13}'={可乐，牙膏，

香皂}，即{可乐}=(可乐，牙膏，香皂)，所以，{可乐｝f{牙膏，香皂｝是一条关联规则，也就是说,所有买可乐的顾客也同时买了牙膏
和香皂。继续进行下去，得到关联规则集(见图 3)。

3．2 关联规则分析

在获取的关联规则中哪些是有意义的?哪些规则对商品搭配进货更有意义?这是很重要的问题。本节就分析这些问题。

设 W={W1, W2, …, WK}是形式上下文中的所有物品集和 I={I1, I2, …, IL}是所有交易事件集。设 X→Y (其中 X, Y⊆W) 是任意关联规则。

交易事件	物品集
I_1	{面包，牛奶，尿布，啤酒，鸡蛋，牙刷，香皂}
I_2	{面包，牛奶，鸡蛋，可乐，牙膏，牙刷，香皂}
I_3	{面包，牛奶，尿布，啤酒，鸡蛋，牙刷}
I_4	{面包，鸡蛋，牙膏，牙刷}
I_5	{牛奶，鸡蛋，可乐，牙膏，牙刷，香皂}
I_6	{牛奶，尿布，啤酒，鸡蛋，牙刷，香皂}
I_7	{面包，牛奶，啤酒，香皂}
I_8	{面包，牛奶，鸡蛋，可乐，牙膏，香皂}
I_9	{可乐，牙膏，牙刷，香皂}
I_{10}	{尿布，啤酒，牙刷，香皂}
I_{11}	{尿布，啤酒，鸡蛋，牙刷}
I_{12}	{面包，啤酒}
I_{13}	{牛奶，可乐，牙膏，香皂}
I_{14}	{面包，牙膏}

图 1 销售信息事件序列

	面包	牛奶	尿布	啤酒	鸡蛋	可乐	牙膏	牙刷	香皂
I_1	3	4	2	3	3			3	2
I_2	5	3			4	4	2	2	2
I_3	6	7	3	5	5			4	
I_4	4				4		3	5	
I_5		5			5	6	4	6	1
I_6		4	3	6	6			2	2
I_7	7	3		5					1
I_8	9	6			8	7	5		1
I_9						4	3	2	3
I_{10}			2	6				3	3
I_{11}			4	7	5			2	
I_{12}	4			6					
I_{13}		6				5	4		5
I_{14}	6						6		

图 2 对应图 1 的形式上下文

我们说规则的可采纳性就是对商品搭配进货管理方便程度的直观描述。

例如 Rule1{可乐}→{牙膏，香皂}比 Rule2{鸡蛋，香皂}→{牛奶}的可采纳性强，因为前者只需考虑可乐的存货情况就可以大致决定牙膏和香皂的存货状态，而后者需要同时考察鸡蛋和香皂的存货情况才能大致了解牛奶的存货情况。

定义 6 (支持度) 定义规则 X→Y 的支持度为:

$$\text{Support}(X \to Y) = \frac{|\{Ii|X \cup Y \subseteq Ii, Ii \in I\}}{L}$$

关联规则:

Rule 1: {可乐}→{牙膏，香皂}
Rule 2: {鸡蛋，香皂}→{牛奶}
Rule 3: {牛奶，牙刷}→{鸡蛋}
Rule 4: {尿布}→{啤酒，牙刷}
Rule 5: {面包，香皂}→{牛奶}
Rule 6: {面包，牙刷}→{鸡蛋}
Rule 7: {啤酒，鸡蛋} →{尿布，牙刷}
Rule 8: {啤酒，牙刷}→{尿布}
Rule 9: {面包，尿布，啤酒，鸡蛋，牙刷}→{牛奶}
Rule 10: {牙膏，香皂}→{可乐}
Rule 11: {牛奶，牙膏}→{可乐，香皂}
Rule12: {面包，牛奶，可乐，牙膏，香皂}→{鸡蛋}

图 3 利用算法 1 得到的关联规则集

命题 1 支持度、兴趣度、复杂度和平衡度具有以下直观性质:

(1) 支持度越大，规则的可采纳性越强;

(2) 兴趣度越大, 规则的可采纳性越强;

(3) 复杂度越大而平衡度越小, 规则的可采纳性越弱;

(4) 在复杂度和支持度相同时, 平衡度越小规则的可采纳性越强;

(5) 支持度越大而平衡度越小, 规则的可采纳性越强。

如规则{可乐}→{牙膏, 香皂}的支持度是 5/14, 因为包含{可乐, 牙膏, 香皂}的交易事件是 2, 5, 8, 9 和 13。同理, 规则{鸡蛋, 香皂}→{牛奶}的支持度是 3/7。很显然, 支持度越大说明物品越畅销, 支持度越小说明物品越滞销。

命题 2 规则 X→Y 的兴趣度=Support (Y)。

兴趣度越大, 说明 Y 的支持度越大, 也就说明有相同阶的前提集合所确定的后件越多, 规则的实际意义就越好。

命题 3 质心到支持集中各支持事件所对应的数量向量的总曼哈顿距离最小。

有了上面的定义, 可以分析柔性决策的思路。决策之前先对规则进行筛选, 选出用户自己满意的规则, 然后根据规则进行商品搭配进货的决策。我们所说的柔性主要是指用户可以自己定义规则的各种属性的不同阈值, 类似于文[8]。下面, 给出规则集合(见图 3) 所对应的形式上下文。假设用户给定阈值:支持度>2/7∧兴趣度≥2∧复杂度<2/3∧平衡度≤5, 则图 5 就是所对应的形式上下文。所对应的概念格见图 6, 注意, 图 6 中的编号 1～12 表示规则 1～规则 12。

	支持度	兴趣度	复杂度	平衡度
Rule 1	5/14	14/5	1/3	1/2
Rule 2	5/14	7/4	1/3	2
Rule 3	5/14	7/4	1/3	2
Rule 4	5/14	14/5	1	1/2
Rule 5	2/7	7/4	1/3	2
Rule 6	2/7	7/4	1/3	2
Rule 7	2/7	14/5	4/9	1
Rule 8	5/14	14/5	1/3	2
Rule 9	5/14	7/4	2/3	5
Rule 10	5/14	14/5	1/3	2
Rule 11	5/14	14/5	4/9	1
Rule 12	5/14	7/4	2/3	5

图 4 图 3 所对应的形式上下文

	支持度	兴趣度	复杂度	平衡度
Rule 1	√	√	√	√
Rule 2	√		√	√
Rule 3	√		√	√
Rule 4	√	√		√
Rule 5			√	√
Rule 6			√	√
Rule 7		√	√	√
Rule 8	√	√	√	√
Rule 9	√			√
Rule 10	√	√	√	√
Rule 11	√	√	√	√
Rule 12	√			√

图 5 (支持度>2/7∧兴趣度≥2∧复杂度<2/3∧平衡度≤5)所对应的形式上下文

从概念格中很容易看出满足用户所有条件的规则有 1, 8, 10 和 11。只满足支持度和兴趣度的规则有 1, 4, 8, 10 和 11。只满足兴趣度和复杂度的规则有 1, 7, 8, 10 和 11。类似地, 用户还可以用阈值的析取形式表示约束条件得到相应的概念格, 在概念格上抽取感兴趣的规则。

从格中选择出满足条件的规则, 用户可以根据规则, 分别找出规则所对应的质心, 按相应的比例进行物品的搭配进货。例如规则 1{可乐}→{牙膏, 香皂}满足用户的阈值条件, 规则 1 的支持集为{I2, I5, I8, I9, I13}, 所对应的质心为 (可乐, 牙膏, 香皂) = (5.2, 3.6, 2.4) 。因为规则的可信度为 100%, 所以, 可以根据大约按可乐:牙膏:香皂=1:9/13:6/13 的比例搭配进货。当多个规则中有商品交叉时, 进货比例按叠加原则处理。

3．3 约束规则格抽取算法

现在, 把上述的思路表示成算法的形式

算法 2 约束规则格抽取算法

输入: 交易数据样本集和阈值组合
输出: 满足阈值条件的关联规则概念格
过程:
步骤 1: 把交易数据样本集转化为形式上下文(G_1, M_1, I_1)的表格表示。// *G_1 是交易事件集, M_1 是商品集合, 表中交叉处表示交易事件中对应商品的数量 *//
步骤 2: 调用算法 1 抽取(G_1, M_1, I_1)中的关联规则集 R;
步骤 3: 计算关联规则 $R \in R$ 的支持度、兴趣度、复杂度和平衡度;
步骤 4: 生成规则集 R 所对应的形式上下文(G_2, M_2, I_2); // *$G_2 = R$, M_2={支持度, 兴趣度, 复杂度, 平衡度}, 表中交叉处表示规则对应属性的取值 *//
步骤 5: 根据用户的阈值约束将(G_2, M_2, I_2)修改为(G_2, M_2, I_3) // * 将不符合用户阈值条件的属性改为空(Null) *//
步骤 6: 调用算法 3 生成约束的规则格

算法 3 概念生成算法[7]

输入：形式上下文(G, M, I)，其中$G=\{g_1, \cdots, g_N\}:=\{1, 2, \cdots, N\}$
和$M=\{m_1, \cdots, m_k\}$
输出：概念序列
步骤 1：$A_1:=\Phi''$，$\Xi_1:=\{\}$// *Ξ 表示概念集 *//
步骤 2：FOR $i=N$ to 1 DO
IF $A<_i((A\cap\{1, 2, \cdots, i-1\})\cup\{i\})''$
THENA$\oplus i=((A\cap\{1, 2, \cdots, i-1\})\cup\{i\})''$
$\Xi\leftarrow\Xi\cup\{(A, A')\}$
$A\leftarrow A\oplus i$
ELSE $i_1\leftarrow i-1$
步骤 3：对 A 执行步骤 2 的循环
步骤 4：输出 Ξ

总结 本文主要讨论了如何用形式概念分析理论进行超市交易数据分析和挖掘。在关联规则柔性选取的基础上，求出规则的质心，然后根据叠加原则进行商品搭配进货决策。柔性体现在用户可以自行定义阈值选取感兴趣的规则，来指导决策规划。

参 考 文 献

[1]Lei Yuxia, Wang Yan, Cao Baoxiang, Yu Jiguo.Concept Inter-connection Based on Many-Valued Context Analysis.ln:Z.-H.Zhou, H. Li, Q.Yang, eds.PAKDD，07, LNAI4426, 2007.623 ∽630

[2] Stumme G, Taouil R, Bastide Y, et al.Intelligent Structuring and Reducing of Association Rules with Formal Concept Analy-sis.ln:F. Baader, G.Brewka, T.Eiter, eds.KI2001, LNAI2174, 2001.335 ∽350

[3]周欣,沙朝锋,朱扬勇,施伯乐.兴趣度——关联规则的又一个阈值[J].计算机研究与发展,2000(05):627-633.

[4]谢志鹏,刘宗田.概念格与关联规则发现[J].计算机研究与发展,2000(12):1415-1421.

[5]周皓峰,朱扬勇,施伯乐.一个基于兴趣度的关联规则采掘算法[J].计算机研究与发展,2002(04):450-457.

[6]胡吉明,鲜学丰.挖掘关联规则中Apriori算法的研究与改进[J].计算机技术与发展,2006(04):99-101+104.

[7]Bernhard G, Wille R.Formal Concept Analysis:Mathematical Foundations.Springer, 1999.

[8] Zhang Ji-fu, Zhang Su-Ian, Hu Li-hua.Constrained concept lat-tice andits construction method (In Chinese) [J].CAAI Transac-tions on Intelligent Systems, 2006, 1 (2) :31 ∽38.