

三支概念格中的关联规则提取算法

刘美玉, 祁建军, 刘伟

(西安电子科技大学计算机科学与技术学院, 710071, 西安)

摘要: 针对基于形式概念分析的关联规则提取侧重属性之间的正关联、忽略负关联的问题,提出一种基于三支概念分析的关联规则提取算法(3ARM)。利用对象导出三支概念的内涵包括表达“共同具有”语义的正属性子集和表达“共同不具有”语义的负属性子集的特点,结合三支概念格的泛化与例化结构,高效地提取正负关联规则;基于三支概念的闭项集特性,从三支概念格中选出包含频繁项集的候选概念进行挖掘,减少不必要的搜索;通过对三支概念之间的关系进行研究,从父子概念中提取无冗余的正关联规则和负关联规则,再从兄弟概念中提取正负规则对规则集进行补充,充分挖掘三支概念格中的知识。MovieLens 数据集上的实验结果表明:应用 3ARM 算法,在最小支持度为 10% 时,得到正规则 86 027 条,负规则 93 685 条;3ARM 算法得出的正规则数量比 FARM 算法的多出 0.9 倍~1.5 倍,减少了 FISM 算法最多 28.3% 的冗余负规则,分别减少了 FISM 和 FARM 算法 44%~63% 和 27%~62% 的运行时间。

关键词: 三支概念分析;三支概念格;关联规则;正关联;负关联

中图分类号: TP18 **文献标志码:** A

DOI: 10.7652/xjtuxb202109021 **文章编号:** 0253-987X(2021)09-0189-08



OSID 码

Extracting Association Rules in Three-Way Concept Lattices

LIU Meiyu, QI Jianjun, LIU Wei

(School of Computer Science and Technology, Xidian University, Xi'an 710071, China)

Abstract: Aiming at the problem that the association rule extraction based on formal concept analysis focuses on positive association and ignores negative association between attributes, an association rule extraction method based on three-way concept analysis, named 3ARM, is proposed. The intent of object-induced three-way concept is composed of two parts, including the positive attribute subset that expresses the semantics of “jointly possessed” and the negative attribute subset that expresses the semantics of “jointly not possessed”. The 3ARM algorithm uses these characteristics of object-induced three-way concept and the generalization and instantiation structure of three-way concept lattice to efficiently extract the positive and negative association rules. Based on the characteristics of the closed itemsets of three-way concepts, candidate concepts containing frequent itemsets are selected from three-way concept lattices for mining, so unnecessary searches are reduced. By studying the relationship between three-way concepts, extracting non-redundant positive and negative association rules from parent-child concepts, and then extracting the positive and negative rules from sibling concepts to supplement the rule set, the knowledge in three-way concept lattices can be fully dug up. Experimental

results on the MovieLens dataset show that using the 3ARM algorithm, when the minimum support is 10%, 86 027 positive rules and 93 685 negative rules are obtained. The number of positive rules obtained by 3ARM is 0.9–1.5 times more than that of FARM, reducing the redundant negative rules of FISM by up to 28.3%, and reducing the running time of the FISM and FARM algorithms by 44%–63% and 27%–62% respectively.

Keywords: three-way concept analysis; three-way concept lattice; association rule; positive association; negative association

自 Agrawal 等首次从大型数据库中提取出关联规则^[1]以来,关联规则的提取已成为数据挖掘领域中的重要内容。关联规则可以在大量的数据中揭示特征与特征之间的隐含关系,从而达到数据挖掘的目的。

关联规则表示属性集之间的相关性。在形式概念分析(FCA)^[2]中,概念表示对象集和属性集之间的统一,概念格节点之间的关系表示概念的泛化和例化。因此,概念格是非常适合关联规则提取的工具之一。目前,已有一些利用概念格进行关联规则发现的方法。Zaki 等提出 CHARM 算法,利用概念格闭项集的特性和对概念格的深度优先搜索提取了非冗余的关联规则^[3];Stumme 等提出 TITANIC 算法,通过对概念格进行改造,提高了关联规则挖掘的效率^[4];Kim 等设计了一个完整的关联规则挖掘系统 FARM,该系统将关联规则与植物的生长模型联系起来,依据布尔表达式生成只与用户兴趣相关的关联规则,与 CHARM 算法和 TITANIC 算法相比,FARM 有更优的效果^[5]。在国内:谢志鹏等首先发现了概念格在关联规则提取上的优势,提出了概念格节点内涵缩减概念,给出了相应的渐进式生成算法和基于概念格的关联规则提取算法^[6-8];Wang 等利用领域知识来指导概念格上的关联规则挖掘^[9];王志海等基于概念格的闭项集特性,提出了一种更有利于规则提取的格结构,并给出了闭标记格中的规则提取算法^[10]。从前人的工作中可以总结出,利用 FCA 提取规则的优点是只提取最大规模的频繁项集,从而缩小了搜索空间,减少不必要的挖掘,从整体上提高了效率。另外,使用闭项集还可以减少规则的冗余。然而,这些方法主要研究了属性子集之间的“具有”关系,即正相关,忽略了属性子集之间的“不具有”关系,即负相关。

Brin 等首次提出了负关联的概念,负关联规则描述了两个项集之间的反作用,即某个项集的发生会降低另外某个项集发生的概率^[11]。此后,基于 FCA 进行负规则提取的研究也有了进一步发展。

Rodriguez-Jimenez 等在建格算法 NextClosure 的基础上提取了混合属性蕴含^[12];Missaoui 等根据原形式背景和补背景,提出了一种混合蕴含的提取方法^[13]。这些研究仅关注了项集之间的蕴含,并未提取出完整的负规则。

为了更完整地表达形式背景中的信息,Qi 等结合三支决策^[14]和形式概念分析提出了三支概念分析(3WCA)^[15]。三支概念分析是一种进行知识表示与知识发现的理论,可以同时表达“共同具有”和“共同不具有”两种语义,相较于形式概念分析,三支概念分析蕴含了更加丰富的信息。目前,基于三支概念分析理论的研究有了丰富的成果^[16-24]。为了获取规则方面的知识,许多学者进行了深入的研究:文献[20]给出了决策形式背景上基于属性导出三支协调性和对象导出三支协调性的规则提取,并与强协调决策形式背景的规则进行了比较;文献[23-24]给出了决策形式背景在属性导出三支概念格下的规则提取方法,在三支概念框架下提出了负规则。这些研究都是基于决策形式背景进行的。

本文以形式背景作为原始数据信息,基于三支概念分析理论对关联规则的提取进行研究,提出基于三支概念分析的关联规则提取算法(3ARM)。该算法从对象导出概念格中概念之间的父子关系和兄弟关系出发,同时提取正关联规则与负关联规则,获得更丰富的知识。

1 三支概念分析

三支概念格是三支概念分析的主要内容,三支概念格包括对象导出(OE)概念格和属性导出(AE)概念格。本文主要在 OE 概念格上进行研究,因此只介绍 OE 概念格的相关理论。

定义 1^[2] 形式背景是一个三元组 (G, M, I) ,其中: $G = \{x_1, x_2, \dots, x_n\}$,每一个 $x_i (i \leq n)$ 被称为一个对象; $M = \{a_1, a_2, \dots, a_m\}$,每一个 $a_j (j \leq m)$ 被称为一个属性。对于 $x \in G$ 与 $a \in M$,如果对象 x 拥有属性 a ,记为 xIa ,即 $(x, a) \in I$ 。

定义 2^[2] 设 (G, M, I) 为一个形式背景,对于任意的对象子集 $X \subseteq G$ 和属性子集 $A \subseteq M$,一对算子被定义为

$$\begin{aligned} * : P(G) \rightarrow P(M), X^* &= \{a \in M \mid \forall x \in X, \\ &\quad (x, a) \in I\} \\ * : P(M) \rightarrow P(G), A^* &= \{x \in G \mid \forall a \in A, \\ &\quad (x, a) \in I\} \end{aligned}$$

式中 $P(\cdot)$ 表示幂集。该算子定义为正算子。

定义 3^[2] 设 (G, M, I) 为一个形式背景, $X \subseteq G, A \subseteq M$ 。若二元组 (X, A) 满足 $X^* = A$ 且 $A^* = X$,则称 (X, A) 为 (G, M, I) 一个形式概念,简称概念。 X 称为 (X, A) 的外延, A 称为 (X, A) 的内涵。

定义 4^[15] 设 (G, M, I) 为一个形式背景,对于任意的对象子集 $X \subseteq G$ 和属性子集 $A \subseteq M$,一对负算子被定义为

$$\begin{aligned} \bar{*} : P(G) \rightarrow P(M), X^{\bar{*}} &= \{a \in M \mid \forall x \in X, xI^c a\} \\ \bar{*} : P(M) \rightarrow P(G), A^{\bar{*}} &= \{x \in G \mid \forall a \in A, xI^c a\} \end{aligned}$$

式中 $I^c = (G \times M) - I, \times$ 为笛卡尔积。

文献[15]中同时考虑正算子和负算子,从对象子集与属性子集之间“共同具有”和“共同不具有”两个角度出发,给出如下三支算子的定义。

定义 5^[15] 设 (G, M, I) 为一个形式背景,对于任意的对象子集 $X, Y \subseteq G$ 和属性子集 $A, B \subseteq M$,一对对象导出三支算子定义为

$$\begin{aligned} \triangleleft : P(G) \rightarrow DP(M), X^{\triangleleft} &= (X^*, X^{\bar{*}}) \\ \triangleright : DP(M) \rightarrow P(G), (A, B)^{\triangleright} &= A^* \cap B^{\bar{*}} \end{aligned}$$

式中 $DP(\cdot) = P(\cdot) \times P(\cdot)$ 。该算子简称为 OE 算子。

定义 6^[15] 设 (G, M, I) 为一个形式背景,如果 $X^{\triangleleft} = (A, B)$ 与 $(A, B)^{\triangleright} = X$ 同时成立,则称 $(X, (A, B))$ 为 (G, M, I) 的一个对象导出三支概念,简称 OE 概念, X 称为 $(X, (A, B))$ 的外延, (A, B) 称为 $(X, (A, B))$ 的内涵。

对于 $(X, (A, B))$ 和 $(Y, (C, D))$,定义偏序关系为 $(X, (A, B)) \leq (Y, (C, D)) \Leftrightarrow X \subseteq Y \Leftrightarrow (A, B) \supseteq (C, D)$ 。

用 $OEL(G, M, I)$ 表示由形式背景 (G, M, I) 生成的所有 OE 概念的集合,则 $OEL(G, M, I)$ 在定义 6 的偏序关系“ $\leq$ ”下是一个完备格,称为 (G, M, I) 的对象导出三支概念格,简记为 OE 概念格。其中,上确界和下确界分别为 $(X, (A, B)) \vee (Y, (C, D)) = ((X \cup Y)^{\triangleright \triangleleft}, (A, B) \cap (C, D))$ 和 $(X, (A, B)) \wedge (Y, (C, D)) = (X \cap Y, ((A, B) \cup (C, D))^{\triangleright \triangleleft})$ 。

用 $(X, (A, B)) < (Y, (C, D))$ 表示 $(X, (A, B))$

$\leq (Y, (C, D))$ 且 $(X, (A, B)) \neq (Y, (C, D))$ 。如果 $(X, (A, B)) < (Y, (C, D))$ 成立,且不存在使得 $(X, (A, B)) < (Z, (E, F)) < (Y, (C, D))$ 成立的 $(Z, (E, F))$,则称 $(X, (A, B))$ 是 $(Y, (C, D))$ 的子概念, $(Y, (C, D))$ 是 $(X, (A, B))$ 的父概念,记为 $(X, (A, B)) < (Y, (C, D))$ 。

例 1 表 1 表示形式背景 (G, M, I) ,表中 $G = \{1, 2, 3, 4\}, M = \{a, b, c, d, e\}$, $\bigcirc$ 表示对象具有属性,空表示对象不具有属性。图 1 为 (G, M, I) 对应的 OE 概念格 $OEL(G, M, I)$,图中概念外延和内涵都以罗列的元素序列来表示相应集合,如 123 表示 $\{1, 2, 3\}$ 。

表 1 形式背景 (G, M, I)
Table 1 Formal context (G, M, I)

G	a	b	c	d	e
1	$\bigcirc$	$\bigcirc$		$\bigcirc$	$\bigcirc$
2	$\bigcirc$	$\bigcirc$	$\bigcirc$		
3	$\bigcirc$	$\bigcirc$	$\bigcirc$		
4				$\bigcirc$	$\bigcirc$

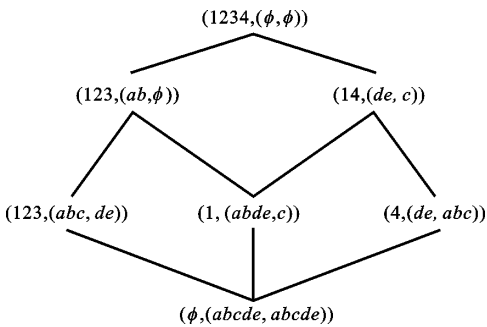


图 1 $OEL(G, M, I)$
Fig. 1 $OEL(G, M, I)$

在 $OEL(G, M, I)$ 中,凡是有边直接相连的两个概念就是一对父子概念。同一个父概念的两个子概念是一对兄弟概念。

2 三支概念格中的关联规则提取

关联规则是形如 $A \Rightarrow B$ 的规则,其中 A 和 B 为项的集合。任意项集 X 的支持度是包含 X 的事务在事务集 D 中所占百分比,记为 $\text{supp}(X) = |X(t)|/|D|$,其中 $X(t) = \{D \text{ 中事务 } t \mid t \text{ 包含 } X\}$ 。支持度大于给定的最小支持度的项集称为频繁项集。关联规则 $A \Rightarrow B$ 的支持度 $\text{supp}(A \Rightarrow B) = \text{supp}(A \cup B)$,即包含 $A \cup B$ 的事务在事务集中所占百分比。关联规则的置信度 $\text{conf}(A \Rightarrow B) = \text{supp}(A \cup B)/\text{supp}(A)$,即事务集中包含 A 的事务同时包

含 B 的百分比。

挖掘关联规则的目标是生成所有支持度大于给定的最小支持度 s 且置信度大于最小置信度 c 的关联规则。一般包括计算所有的频繁项集、从频繁项集中生成置信度大于最小置信度的关联规则两步。本文依据这一思想,对三支概念格进行剪枝,选出候选概念,再根据候选概念间的父子关系和兄弟关系提取出符合条件的关联规则。

在形式概念分析中,形式背景 (G, M, I) 可以理解成事务数据库,其中 G 为数据库中事务的集合, M 为数据库中所有可能项的集合。对于 $x \in G$ 与 $a \in M, xIa$ 则表示 a 属于事务 x 的项集,那么项集之间的关系在概念格中就得到了充分的体现,每个概念就是一个最大频繁项集。概念 (X, A) 中, X 是全体事务集的子集, A 是这些事务共同具有的闭项集。一条关联规则 $A \Rightarrow B$ 成立,对应于内涵包含 A 的概念同样包含 B 。

三支概念分析可以同时表示“共同具有”和“共同不具有”两种语义。对于 $x \in G$ 与 $a \in M, xI^c a$ 则表示 a 不属于事务 x 的项集。在 OE 概念格中, $(X, (A, B))$ 表示事务集 X 中的事务共同具有项集 A , 且共同不具有项集 B 。在三支概念格的两种语义上,除了形如 $A \Rightarrow B$ 的关联规则之外,还可以获取形为 $\neg C \Rightarrow \neg D$ 、 $\neg C \Rightarrow D$ 和 $C \Rightarrow \neg D$ 的规则,分别表示共同不具有 C 的概念同样共同不具有 D 、共同不具有 C 的概念共同具有 D 、共同具有 C 的概念共同不具有 D ,称之为负关联规则,简称负规则。对应地,将 $A \Rightarrow B$ 称为正关联规则,简称正规则。本文主要讨论 OE 概念格中的关联规则提取。

定义 7 设 (G, M, I) 为一个形式背景, $A, B, C, D \subseteq M$, 正规则 $A \Rightarrow B$ 的支持度定义为具有 $A \cup B$ 的对象在 G 中所占的比例,记为

$$\text{supp}(A \Rightarrow B) = \frac{|(A \cup B)^*|}{|G|}$$

负规则 $\neg C \Rightarrow \neg D$ 的支持度定义为不具有 $C \cup D$ 的对象在 G 中所占的比例,记为

$$\text{supp}(\neg D) = \frac{|(C \cup D)^{\bar{*}}|}{|G|}$$

定义 8 设 (G, M, I) 为一个形式背景, $A, B, C, D \subseteq M$, 正规则 $A \Rightarrow B$ 的置信度定义为具有 $A \cup B$ 的对象在具有 A 的对象中占的比例,记为

$$\text{conf}(A \Rightarrow B) = \frac{|(A \cup B)^*|}{|A^*|}$$

负规则 $\neg C \Rightarrow \neg D$ 的置信度定义为不具有 $C \cup D$ 的对象在不具有 C 的对象中占的比例,记为

$$\text{conf}(\neg C \Rightarrow \neg D) = \frac{|(C \cup D)^{\bar{*}}|}{|C^{\bar{*}}|}$$

文献[25]中提出,对于频繁项集 C 和不频繁项集 D ,由于 $\text{supp}(C) > s, \text{supp}(D) \approx 0$,则 $\text{supp}(C \Rightarrow D) \approx \text{supp}(C)$, $\text{conf}(C \Rightarrow \neg D) = \text{supp}(C \cup \neg D) / \text{supp}(C) \approx 1$ 。类似地, $\text{supp}(\neg C \Rightarrow D) \approx \text{supp}(D)$, $\text{conf}(\neg C \Rightarrow D) \approx 1$ 。

在 OE 概念格上提取关联规则的基本思想为:提取每个 OE 概念内涵中的最大频繁项集和最大不频繁项集之间的负规则;依据 OE 概念的直接泛化产生无冗余的正规则和 $\neg C \Rightarrow \neg D$ 形式的负规则;依据概念间的兄弟关系对规则集进行补充,在减少冗余的基础上保证规则的完整。其他的规则可使用文献[13]中的方法由规则集中的规则来推导出。

为了减少不必要的搜索,先从所有的概念中选取候选概念。根据定义 7 可以得出,如要满足 $\text{supp}(A \Rightarrow B) > s$,则应先满足 $|A^*| > s|G|$,即 $\text{supp}(A \Rightarrow B) = |(A \cup B)^*| / |G| > s \Rightarrow |A^*| / |G| > s$ 。类似地,需要满足 $\text{supp}(\neg C \Rightarrow \neg D) > s$,则应先满足 $|C^{\bar{*}}| > s|G|$ 。在 OEL (G, M, I) 中,对于概念节点 $(X, (A, B))$,如果 $|X| > s|G|$,则 $|A^*| > s|G|$ 且 $|B^{\bar{*}}| > s|G|$,该概念为候选概念。从候选概念 $(X, (A, B))$ 出发,先提取概念的内涵中 A 和 B 的关联关系,再利用概念之间的父子关系提取满足条件的关联规则,最后提取兄弟概念间的关联规则。根据以下 3 步来进行:

(1) 对概念 $(X, (A, B))$,可得负规则 $\neg B \Rightarrow A$ 和 $A \Rightarrow \neg B$;

(2) 如果有 $(X, (A, B)) < (Y, (C, D))$,则可以得到正规则 $C \Rightarrow A - C$ 和负规则 $\neg D \Rightarrow \neg(B - D)$;

(3) 如果同时满足 $(X, (A, B)) < (Z, (E, F))$ 且 $(Y, (C, D)) < (Z, (E, F))$,则可以得到正规则 $A \Rightarrow C - A \cap C$ 和负规则 $\neg B \Rightarrow \neg(B - B \cap D)$ 。

根据这 3 步,给出基于三支概念分析的正负关联规则提取算法,即 3ARM 算法,具体如下。

3ARM 算法:从 OEL (G, M, I) 中提取正负关联规则输入: OEL $(G, M, I), s, c$

输出: 正规则集合 R^+ 、正规则对应的支持度集合 S^+ 和置信度集合 C^+ , 负规则集合 R^- 、负规则对应的支持度集合 S^- 和置信度集合 C^-

1 for $(X_1, (A_1, B_1))$ in OEL (G, M, I) ;

2 if $|X_1| < s|G|$ then

3 continue

4 put $\neg B_1 \Rightarrow A_1, A_1 \Rightarrow \neg B_1$ into R^-

```
5  put supp( $\neg B_1 \Rightarrow A_1$ ) into  $S^-$ 
6  put supp( $A_1 \Rightarrow \neg B_1$ ) into  $S^-$ 
7  put conf( $\neg B_1 \Rightarrow A_1$ ) into  $C^-$ 
8  put conf( $A_1 \Rightarrow \neg B_1$ ) into  $C^-$ 
9  for  $(X_2, (A_2, B_2))$  in  $(X_1, (A_1, B_1))$ . Parent-
   concepts:
10  if supp( $A_2 \Rightarrow A_1 - A_2$ )  $> s$  and conf( $A_2 \Rightarrow A_1 -$ 
     $A_2$ )  $> c$  then
11      put  $A_2 \Rightarrow A_1 - A_2$  into  $R^+$ 
12      put supp( $A_2 \Rightarrow A_1 - A_2$ ) into  $S^+$ 
13      put conf( $A_2 \Rightarrow A_1 - A_2$ ) into  $C^+$ 
14  if supp( $\neg B_2 \Rightarrow \neg(B_1 - B_2)$ )  $> s$  and
    conf( $\neg B_2 \Rightarrow \neg(B_1 - B_2)$ )  $> c$  then
15      put  $\neg B_2 \Rightarrow \neg(B_1 - B_2)$  into  $R^-$ 
16      put supp( $\neg B_2 \Rightarrow \neg(B_1 - B_2)$ ) into  $S^-$ 
17      put conf( $\neg B_2 \Rightarrow \neg(B_1 - B_2)$ ) into  $C^-$ 
18  for  $(X_3, (A_3, B_3))$  in  $(X_2, (A_2, B_2))$ . Child-
    concepts:
19  if  $(X_3, (A_3, B_3)) \neq (X_1, (A_1, B_1))$  then
20  if supp( $A_1 \Rightarrow A_3 - A_1 \cap A_3$ )  $> s$  and
    conf( $A_1 \Rightarrow A_3 - A_1 \cap A_3$ )  $> c$  then
21      put  $A_1 \Rightarrow A_3 - A_1 \cap A_3$  into  $R_+$ 
22      put supp( $A_1 \Rightarrow A_3 - A_1 \cap A_3$ ) into  $S^+$ 
23      put conf( $A_1 \Rightarrow A_3 - A_1 \cap A_3$ ) into  $C^+$ 
24  if supp( $\neg B_1 \Rightarrow \neg(B_3 - B_1 \cap B_3)$ )  $> s$  and
    conf( $\neg B_1 \Rightarrow \neg(B_3 - B_1 \cap B_3)$ )  $> c$  then
25      put  $\neg B_1 \Rightarrow \neg(B_3 - B_1 \cap B_3)$  into  $R^-$ 
26      put supp( $\neg B_1 \Rightarrow \neg(B_3 - B_1 \cap B_3)$ )
        into  $S^-$ 
27      put conf( $\neg B_1 \Rightarrow \neg(B_3 - B_1 \cap B_3)$ )
        into  $C^-$ 
28 return  $R^+, S^+, C^+, R^-, S^-, C^-$ 
```

步骤 1~3: 设置循环体, 选择候选概念 $(X_1, (A_1, B_1))$ 。

步骤 4~8: 从概念 $(X_1, (A_1, B_1))$ 的内涵中提取负关联规则。

步骤 9~13: 提取 $(X_1, (A_1, B_1))$ 和父概念之间的正关联规则, 并根据定义 7 和 8 计算正规则的置信度和支持度。

步骤 14~17: 提取 $(X_1, (A_1, B_1))$ 和父概念之间的负关联规则, 并计算负规则的置信度和支持度。

步骤 18~23: 设置循环体, 此时 $(X_2, (A_2, B_2))$ 是 $(X_3, (A_3, B_3))$ 和 $(X_1, (A_1, B_1))$ 的公共父概念, 提取 $(X_1, (A_1, B_1))$ 和兄弟概念 $(X_3, (A_3, B_3))$ 之间

的正关联规则, 计算两个正规则的置信度和支持度。

步骤 24~27: 提取 $(X_1, (A_1, B_1))$ 和兄弟概念 $(X_3, (A_3, B_3))$ 之间的负关联规则, 并计算两个负规则的置信度和支持度。

在第二层循环结束时, 关于概念 $(X_1, (A_1, B_1))$ 生成的规则提取完毕。另外, 从算法中可以分析得出, 随着最小支持度和最小置信度的增大, 候选概念的数量减少, 规则的数量随之减少, 程序运行的时间也随之减少, 然而对非候选概念的判断相对增多, 使得整体效率略微降低。

例 2 根据 3ARM 算法, $s = c = 0.4$ 时, 例 1 中 $OEL(G, M, I)$ 得到关联规则如表 2 所示。

表 2 关联规则及其支持度和置信度
Table 2 Association rules and their support and confidence

序号	规则	支持度	置信度
1	$\phi \Rightarrow ab$	3/4	3/4
2	$\phi \Rightarrow de$	1/2	1/2
3	$ab \Rightarrow c$	1/2	2/3
4	$\neg \phi \Rightarrow \neg c$	1/2	1/2
5	$\neg \phi \Rightarrow \neg de$	1/2	2/3
6	$de \Rightarrow \neg c$	1/2	1
7	$\neg c \Rightarrow de$	1/2	1
8	$abc \Rightarrow \neg de$	1/2	1
9	$\neg de \Rightarrow abc$	1/2	1

规则 1~5 由概念间的父子关系得到, 负规则 6~9 由单个概念生成。特别的, 前件为空集的规则, 如规则 1、2、4、5, 可以根据实际应用选择保留或者删除。例如, 在推荐系统中, 这些规则可以用作为新用户提供推荐的依据, 而后件为空集的规则一般被忽略。

3 实验及分析

实验数据采用推荐系统中广泛使用的 MovieLens 数据集。数据集组成为: User(userID, age, gender, occupation), Movie(movieID, title, genres, tag), Rating(userID, movieID, rating, timestamp)。其中, Rating 含有 610 个用户对 9 742 部电影的 100 836 个评分。

3.1 正规则分析

首先, 把评分信息转化成形式背景, 建立 OE 概念格和经典概念格。OE 概念格分析得到的概念数和父子概念对数分别为 13 583、137 229, 经典概念格的分别为 10 787、87 984。可以看出, 在相同的形

式背景下,OE 概念格中的概念数和父子概念对数均大于经典概念格中的。

然后,对 3ARM 算法提取到的关联规则与其他算法取得的规则进行数量上和分布上的分析。3ARM 算法得到的关联规则中正负规则的分布情况如图 2 所示。可以看出,得到的正规则数量占全部规则数量的 25%~42%,负规则数量占全部规则数量的 58%~75%,是正规则的 1.06 倍~1.26 倍。

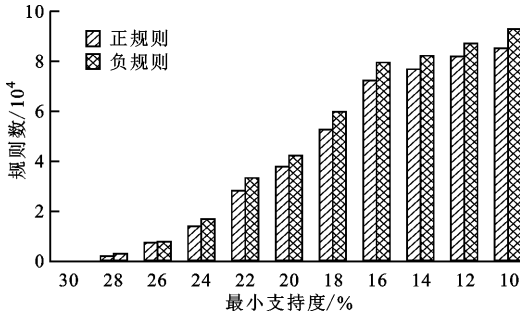


图 2 3ARM 算法得到的正负规则分布情况
Fig. 2 Positive and negative rules distribution from 3ARM

3ARM 算法与 FARM 算法提取出的正规则数对比如图 3 所示。可以看出,两种算法得到的正规则都主要分布在支持度小于 30% 的范围内,且在同一最小支持度阈值下,3ARM 算法得出的正规则数量比 FARM 算法得出的规则数量多出 0.9~1.5 倍。综合分析可知,概念间父子关系数的增多是使得 3ARM 算法得到更多正规则的主要原因。

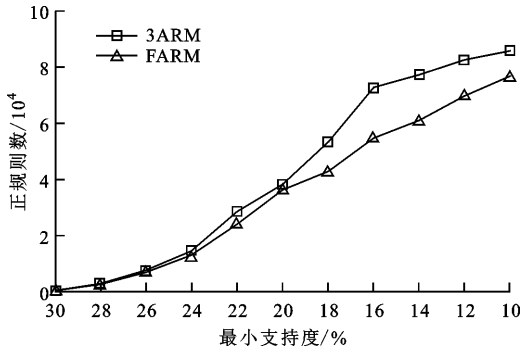


图 3 3ARM 算法与 FARM 算法得到的正规则数对比
Fig. 3 Comparison of positive rules between 3ARM and FARM

3.2 负规则分析

为评估提取到的负规则的可靠性,对 3ARM 算法与 FISM 算法进行对比。FISM 算法是 Balakrishna 等提出的一种基于频繁项集的负关联规则提取算法,利用改进的 FP-tree 生成频繁项集,再从频繁项集中生成负规则^[26]。3ARM 算法与 FISM

算法得到的负规则数对比如图 4 所示。可以看出:支持度为 22%~30% 时,3ARM 算法与 FISM 算法提取到的负规则数量相当;随支持度减小,3ARM 算法提取到的负规则数量逐渐少于 FISM 算法的,且差距逐渐拉大。将两种算法提取到的规则进行对比,可以发现 FISM 算法提取到的规则中存在类似于的冗余规则,而 3ARM 算法利用三支概念的优势,从最大频繁闭项集的泛化与例化关系中提取负规则,大量减少了冗余规则的生成。

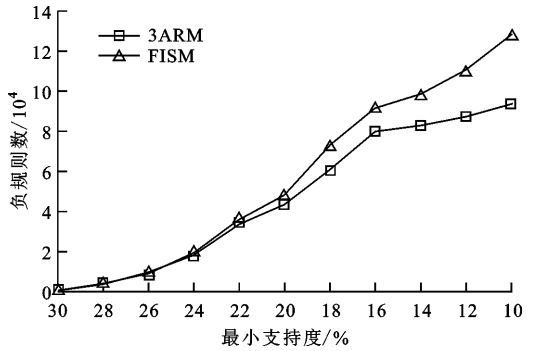


图 4 3ARM 算法与 FISM 算法得到的负规则数对比
Fig. 4 Comparison of negative rules between 3ARM and FISM

3.3 运行时间分析

对 3ARM 算法、FARM 算法和 FISM 算法的运行时间进行对比,实验结果如图 5 所示。需要说明的是,这里的运行时间指的是在概念格上提取关联规则的时间,不包括构建经典格和 OE 概念格的时间,3ARM 算法的运行时间是同时提取正规则和负规则所需时间。为保证公平性,FISM 算法的运行时间也不包含生成 FP-tree 所需的时间,而实际上,3ARM 算法对数据库进行一次扫描就可以建立 OE 概念格,构建 FP-tree 则需要两次扫描数据库。

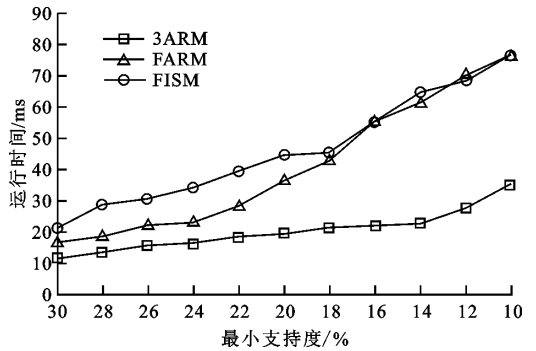


图 5 3ARM、FARM 及 FISM 算法的运行时间对比
Fig. 5 Comparison of running time of 3ARM, FARM and FISM

从图5可以看出,3ARM算法的运行时间明显小于FARM算法和FISM算法的。在同一支持度阈值下,3ARM算法的运行时间比FISM算法减少了44%~63%,比FARM算法减少了27%~62%。与FARM算法相比,3ARM算法省去了布尔表达式的逻辑运算;与FISM算法相比,3ARM算法省去了将频繁项集两两结合生成负规则的时间。3ARM算法利用OE概念格的偏序关系及OE概念闭项集的特性,有规律地提取关联规则,提高了算法的效率。

从第3节实验可以得出,与FARM算法相比,3ARM算法不仅可以延续经典概念格提取非冗余正规则的优点,还可以提取出有价值的负规则。与FISM算法对比,3ARM算法可以减少负规则的冗余。从运行时间上看,3ARM算法减去了复杂的操作,缩短了算法的运行时间。

4 结 论

在关联规则的提取方面,现有许多基于概念格进行正规则提取的研究,取得了较好的效果,并在文本分类^[18]、推荐系统^[27-28]等领域有充分的应用。为更完整地提取关联规则,本文提出基于三支概念分析的关联规则提取算法3ARM,利用三支概念的闭项集特性,从父子关系和兄弟关系中提取了具有置信度和支持度的正关联规则和负关联规则。经过实验对比和分析,得出结论如下。

(1)在同一形式背景上,建立的三支概念格比经典概念格包含更多的概念和父子关系,因此3ARM算法能比FARM算法提取到更多的正规则。

(2)利用三支概念的闭项集特性,能快速得到最大规模的频繁项集,减少了冗余规则的产生;利用3WCA中“共同具有”和“共同不具有”的语义,3ARM算法可以同时提取到正规则和负规则。

(3)根据三支概念间的父子关系和兄弟关系,有规律地进行规则的提取,减少了不必要的操作,使得3ARM算法的运行时间少于FARM算法和FISM算法的。

未来将研究本文算法应用于推荐系统。

参考文献:

[1] AGRAWAL R, IMIELIŃSKI T, SWAMI A. Mining association rules between sets of items in large databases [C]// Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data. New

York, USA: ACM, 1993: 207-216.

- [2] GANTER B, WILLE R. Formal concept analysis: mathematical foundations [M]. Cham, Germany: Springer, 1999: 284.
- [3] ZAKI M J, HSIAO C J. Efficient algorithms for mining closed itemsets and their lattice structure [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(4): 462-478.
- [4] STUMME G, TAOUIL R, BASTIDE Y, et al. Computing iceberg concept lattices with Titanic [J]. Data & Knowledge Engineering, 2002, 42(2): 189-222.
- [5] KIM E H, KIM H G, HWANG S H, et al. FARM: an FCA-based association rule miner [J]. Knowledge-Based Systems, 2015, 85: 277-297.
- [6] 谢志鹏, 刘宗田. 基于概念格的关联规则发现 [J]. 小型微型计算机系统, 2000, 21(10): 1028-1031.
- XIE Zhipeng, LIU Zongtian. Concept lattice based association rule discovery [J]. Mini-Micro Systems, 2000, 21(10): 1028-1031.
- [7] 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, 37(12): 1415-1421.
- XIE Zhipeng, LIU Zongtian. Concept lattice and association rule discovery [J]. Journal of Computer Research and Development, 2000, 37(12): 1415-1421.
- [8] XIE Zhipeng, LIU Zongtian. From intent reducts for attribute implications to approximate intent reducts for association rules [C]// Proceedings of the 5th International Conference on Computer and Information Technology. Piscataway, NJ, USA: IEEE, 2005: 162-169.
- [9] WANG Dexing, HU Xuegang, LIU Xiaoping, et al. Association rules mining on concept lattice using domain knowledge [C]// Proceedings of the 2005 International Conference on Machine Learning and Cybernetics. Piscataway, NJ, USA: IEEE, 2005: 2151-2154.
- [10] 王志海, 胡可云, 胡学钢, 等. 概念格上规则提取的一般算法与渐进式算法 [J]. 计算机学报, 1999, 22(1): 66-70.
- [11] BRIN S, MOTWANI R, SILVERSTEIN C. Beyond market baskets: generalizing association rules to correlations [C]// Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data. New York, USA: ACM, 1997: 265-276.
- [12] RODRIGUEZ-JIMENEZ J M, CORDERO P, ENCISO M, et al. Negative attributes and implications in formal concept analysis [J]. Procedia Computer Science, 2014, 31: 758-765.

- [13] MISSAOUI R, NOURINE L, RENAUD Y. Generating positive and negative exact rules using formal concept analysis: problems and solutions [C]//Proceedings of the 6th International Conference on Formal Concept Analysis. Cham, Germany: Springer, 2008: 169-181.
- [14] YAO Yiyu. Three-way decision: an interpretation of rules in rough set theory [C]//Proceedings of 2009 International Conference on Rough Sets and Knowledge Technology. Cham, Germany: Springer, 2009: 642-649.
- [15] QI Jianjun, WEI Ling, YAO Yiyu. Three-way formal concept analysis [C]//Proceedings of 2014 International Conference on Rough Sets and Knowledge Technology. Cham, Germany: Springer, 2014: 732-741.
- [16] QI Jianjun, QIAN Ting, WEI Ling. The connections between three-way and classical concept lattices [J]. Knowledge-Based Systems, 2016, 91: 143-151.
- [17] 祁建军, 汪文威. 多线程并行构建三支概念 [J]. 西安交通大学学报, 2017, 51(3): 116-121.
QI Jianjun, WANG Wenwei. A multithreaded parallel algorithm for constructing three-way concepts [J]. Journal of Xi'an Jiaotong University, 2017, 51(3): 116-121.
- [18] CHEN Xiaoyu, QI Jianjun, ZHU Xiaomin, et al. Unlabelled text mining methods based on two extension models of concept lattices [J]. International Journal of Machine Learning and Cybernetics, 2020, 11(2): 475-490.
- [19] REN Ruisi, WEI Ling. The attribute reductions of three-way concept lattices [J]. Knowledge-Based Systems, 2016, 99: 92-102.
- [20] WEI Ling, LIU Lin, QI Jianjun, et al. Rules acquisition of formal decision contexts based on three-way concept lattices [J]. Information Sciences, 2020, 516: 529-544.
- [21] 魏玲, 赵思雨. 三支概念分析中的粒与知识结构 [J]. 西北大学学报(自然科学版), 2020, 50(4): 537-545.
WEI Ling, ZHAO Siyu. Granules and knowledge structure in three-way concept analysis [J]. Journal of Northwest University (Natural Science Edition), 2020, 50(4): 537-545.
- [22] 汪文威, 祁建军. 三支概念的构建算法 [J]. 西安电子科技大学学报, 2017, 44(1): 71-76.
WANG Wenwei, QI Jianjun. Algorithm for constructing three-way concepts [J]. Journal of Xidian University, 2017, 44(1): 71-76.
- [23] 刘琳, 钱婷, 魏玲. 基于属性导出三支概念格的决策背景规则提取 [J]. 西北大学学报(自然科学版), 2016, 46(4): 481-487.
LIU Lin, QIAN Ting, WEI Ling. Rules extraction in formal decision contexts based on attribute-Induced three-way concept lattices [J]. Journal of Northwest University (Natural Science Edition), 2016, 46(4): 481-487.
- [24] 刘琳, 魏玲, 钱婷. 决策形式背景中具有置信度的三支规则提取 [J]. 山东大学学报(理学版), 2017, 52(2): 101-110.
LIU Lin, WEI Ling, QIAN Ting. Three-way rules extraction in formal decision contexts with confidence [J]. Journal of Shandong University (Natural Science), 2017, 52(2): 101-110.
- [25] ZHANG Chengqi, ZHANG Shichao. Association rule mining [M]. Cham, Germany: Springer, 2002: 238.
- [26] BALAKRISHNA E, RAMA B, NAGARAJU A. Efficient mining of negative association rules using frequent item set mining [C]//Proceedings of the 1st International Conference on Artificial Intelligence and Cognitive Computing. Cham, Germany: Springer, 2018: 709-716.
- [27] HE Xu, MIN Fan, ZHU W. Top-N recommendation based on granular association rules [C]//Proceedings of 2014 International Conference on Rough Sets and Knowledge Technology. Cham, Germany: Springer, 2014: 194-205.
- [28] 李昌盛, 伍之昂, 张璐, 等. 关联规则推荐的高效分布式计算框架 [J]. 计算机学报, 2019, 42(6): 1218-1231.
LI Changsheng, WU Zhiang, ZHANG Lu, et al. An efficient distributed-computing framework for association-rule-based recommendation [J]. Chinese Journal of Computers, 2019, 42(6): 1218-1231.

(编辑 陶晴)