

一种新的关联规则挖掘的模型^{*}

A New Model of Mining Association Rules

苏毅娟 严小卫

(广西师范大学计算机系 桂林541004)

Abstract A new algorithm for mining positive and negative association rules is presented. A new confidence is constructed to measure the uncertainty of an association rule based on the probability theory and Piatetsky-Shapiro's model.

Keywords Data mining, Association rule, Algorithm

1. 引言

support-confidence 模型是正关联规则挖掘普遍应用的模型,而如何度量关联规则的不确定性则是正关联规则挖掘中的重要问题之一。在该模型中,用 $\text{supp}(X \cup Y)$ 和 $\text{conf}(X \rightarrow Y)$ 来度量关联规则 $X \rightarrow Y$ 的不确定性。然而,用这一度量标准可能会得到诸如 $X \rightarrow Y$,但 X 与 Y 不相关(或独立)的规则,可见,用 $\text{conf}(X \rightarrow Y)$ 来度量关联规则是不够的。

实际应用中,我们不仅要挖掘正关联规则,而且还要挖掘负关联规则。正关联规则即形如 $X \rightarrow Y$ 的式子,负关联规则即形如 $X \rightarrow \bar{Y}$ 的式子,其中 $X, Y \subseteq I, X \cap Y = \emptyset, I$ 为数据库 D 中的所有项的集合,首先看一个例子,若 $p(c)=0.6, p(t)=0.4, p(t \cup c)=0.05, p(t \cup \bar{c})=0.35, \text{minconf}=0.52$, 有 $p(t \cup c)/p(t)=0.05/0.4=0.125 < \text{minconf}=0.52$, 又 $p(t \cup \bar{c})=0.05$, 这说明 $t \rightarrow c$ 不是一条有效的规则,但 $p(t \cup c)=0.05 \neq p(t)p(c)=0.24$, 即 t 与 c 不是独立的,又由于 $p(t \cup \bar{c})=0.35$, 而 $p(t \cup \bar{c})/p(t)=0.35/0.4=0.875 > \text{minconf}$ 。因此, $t \rightarrow \bar{c}$ 为一有效规则。

注意, $\bar{A}$ 与 $I-A$ 是不同的,其中 I 表示数据库中所有项的集合, $A \subseteq I$ 。若 $\forall i \in A, i$ 都在某事务中出现,则项集 A 被看作事件 A 在该事务中发生。若 $\exists i \in A$, 使得 i 不在该事务中出现,则将 $\bar{A}$ 当作事件 A 在该事务中不发生。

由于 support-confidence 模型存在以上的不足,近来不少研究者提出了一些度量关联规则不确定性的方法。在文[3]中, Piatetsky-Shapiro 提出关系上的规则的概念。文[2]中, Motwani 和 Silverstein 从经典的统计学角度利用 Chi-squared 测试项集的相关性来度量

关联规则的有效性。然而,该模型依然存在着许多的不足:(1)运用 Chi-squared 值只能确定 X 与 Y 是否独立;(2)若 X 与 Y 不独立,则需要一个机制来决定 X 与 Y 是正相关还是负相关;(3)若 X 与 Y 是负相关的,则还需要另外的方法决定是哪种形式。另外,文[5]中, Aggarwal 和 Yu 提出聚集强度模型以度量关联规则的关注程度。然而,该模型并没有考虑 Piatetsky-Shapiro 的观点,而且当挖掘负关联规则时,需重新计算 supp 和 conf 的值。

为了解决上述问题,下面我们将引入一个能挖掘正、负关联规则的新模型。

2. 新模型

本节将在 Piatetsky-Shapiro 模型及概率理论的基础上提出兴趣关联规则挖掘的模型。我们首先建立一个挖掘正、负关联规则的算法,并给出一个产生正大项集、负大项集的方法。

为了能够同时挖掘正关联规则和负关联规则,首先必须产生负关联规则的所有项集,我们将正关联规则中的项集称为正大项集,负关联规则中的项集称为负大项集。形如 $A \rightarrow B$ (或 $\bar{A} \rightarrow B$, 或 $\bar{A} \rightarrow \bar{B}$) 的负关联规则中的项集需满足以下条件:

- (1) $A \cap B = \emptyset$;
- (2) $p(A) \geq \text{minsupp}$ 且 $p(B) \geq \text{minsupp}$;
- (3) $p(A \cup B) \geq \text{minsupp}$ (或 $p(\bar{A} \cup B) \geq \text{minsupp}$, 或 $p(\bar{A} \cup \bar{B}) \geq \text{minsupp}$)。

条件(2)保证了负关联规则的意义,其他条件保证了负关联规则的有效性。

产生所有正大项集、负大项集的过程如下:

procedure1 PNLargeItemsets

^{*} 基金项目:广西自然科学基金(0007038)资助,苏毅娟 硕士研究生。

Input: D: 数据库, minsup: 最小支持度

Output: PLI: 正大项集; NLI: 负大项集

(1) $PLI \leftarrow \emptyset$; $NLI \leftarrow C$; $C \leftarrow \emptyset$

(2) for D 中所有项集 A do

{ $FA \leftarrow$ D 中包含 A 的事务个数;

$p(A) \leftarrow FA/|D|$;

$C \leftarrow C \cup \{A\}$; }

(3) for c 中所有项集 i do

if $p(i) \geq \text{minsup}$ then $PLI \leftarrow PLI \cup \{i\}$;

(4) for PLI 中所有项集 A 和 B do

if $A \cup B \in PLI$ then

if $p(A \cup B) \geq \text{minsup}$ 或 $p(\overline{A \cup B}) \geq \text{minsup}$ then

$NLI \leftarrow NLI \cup \{A, B\}$;

其中第(1)步为初始化,第(2)步产生 D 中所有项集及其支持度,第(3)步产生所有正大项集,第(4)步产生所有的负大项集。

关联规则 $X \rightarrow Y$ 中,我们需要考虑 $p(Y|X)$ 与 $P(Y)$ 的关系。根据条件概率 $p(Y|X)$,我们可以有以下三种情况:

(1) 若 $p(Y|X) = p(Y)$, 则 X 与 Y 独立;

(2) 若 $p(Y|X) = \frac{p(X \cup Y)}{p(X)} > p(Y)$, 那么 Y 与 X 正相关;

(3) 若 $p(Y|X) = \frac{p(X \cup Y)}{p(X)} < p(Y)$, 则 Y 与 X 负相关或者 Y 与 X 正相关。

由于 $-p(Y) \leq p(Y|X) - p(Y) \leq 1 - p(Y)$, 为了反映 $p(Y|X)$ 与 $p(Y)$ 之间的关系,我们定义一种名为概率比率 PR 的测度,以描述 $p(Y|X)$ 相对 $p(Y)$ 的增长率。根据兴趣度的定义,若 $X \rightarrow Y$ 为一关注规则,则有 $|p(X \cup Y) - p(X)p(Y)| \geq \lambda$ (其中 λ 为最小兴趣度)。也就是说,当 $-\lambda < p(X \cup Y) - p(X)p(Y) < \lambda$ 时, $X \rightarrow Y$ 为不感兴趣规则。因此我们用

$PR(Y|X) =$

$$\begin{cases} \frac{p(X \cup Y) - p(X)p(Y)}{p(X)(1 - p(Y))} & \text{当 } p(X \cup Y) - p(X)p(Y) \geq \lambda, \\ & p(X)(1 - p(Y)) \neq 0 \\ \frac{p(X \cup Y) - p(X)p(Y)}{p(X)p(Y)} & \text{当 } p(X \cup Y) - p(X)p(Y) < \lambda, \\ & p(X)p(Y) \neq 0 \end{cases}$$

作为规则 $X \rightarrow Y$ 的置信度。下面是用该方法来度量关联规则不确定性的模型。

定义 (PR 模型) I 为数据库 D 中的所有项的集合, $X, Y \subseteq I, X \cap Y = \emptyset, p(X) \neq 0, p(Y) \neq 0$, minsup 和 minconf 分别是由用户给定的最小支持度和最小置信度, λ 为最小兴趣度, 则:

(1) 若 $p(A \cup B) \geq \text{minsup}, PR(B|A) \geq \text{minconf}$, 则 $A \rightarrow B$ 为一兴趣规则;

(2) 若 $p(A \cup B) \geq \text{minsup}, p(A) \geq \text{minsup}, p(B) \geq \text{minsup}, PR(\overline{B}|A) \geq \text{minconf}$, 则 $A \rightarrow \overline{B}$ 为一兴趣规则;

(3) 若 $p(\overline{A \cup B}) \geq \text{minsup}, p(A) \geq \text{minsup}, p(B) \geq \text{minsup}, PR(B|\overline{A}) \geq \text{minconf}$, 则 $\overline{A} \rightarrow B$ 为一兴

趣规则;

(4) 若 $p(\overline{A \cup B}) \geq \text{minsup}, p(A) \geq \text{minsup}, p(B) \geq \text{minsup}, PR(\overline{B}|\overline{A}) \geq \text{minconf}$, 则 $\overline{A} \rightarrow \overline{B}$ 为一兴趣规则。

例1 有一事务数据库 TD 如表1, A=面包, B=咖啡, C=茶, D=糖, E=啤酒, F=奶油。假设 minsup = 0.2, minconf = 0.4, $\lambda = 0.8$ 。

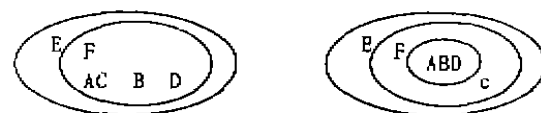
表1 事务数据库 TD 示意图

TransactionID	Items
T1	A, B, D
T2	A, B, C, D
T3	B, D
T4	B, C, D, E
T5	A, C, E
T6	B, D, F
T7	A, E, F
T8	C, F
T9	B, C, F
T10	A, B, C, D, F

运用 PR 模型,可得到如下的关注规则: $B \rightarrow D, A \cup B \rightarrow D, B \rightarrow E, A \cup B \rightarrow E, A \cup D \rightarrow E, B \cup D \rightarrow E, B \cup F \rightarrow E, C \cup F \rightarrow E, D \rightarrow \overline{C \cup F}, A \cup B \cup D \rightarrow E, B \cup D \rightarrow \overline{C \cup F}$ 。

而在 support-confidence 模型中,得到如下的规则: $A \rightarrow B, A \rightarrow C, A \rightarrow D, B \rightarrow C, B \rightarrow D, B \rightarrow F, C \rightarrow D, C \rightarrow F, A \rightarrow B \cup D, A \cup B \rightarrow D, A \cup D \rightarrow B, B \rightarrow C \cup D, B \cup C \rightarrow D, B \cup D \rightarrow C$, 其中 $A \rightarrow C, A \rightarrow D, C \rightarrow F, A \rightarrow B \cup D$ 为独立的规则。

图1根据挖掘的关联规则给出了所有项的布局图,显然(b)的布局比(a)的布局要合理。



(a) 对应 support-confidence 模型的项布局图

(b) 对应 PR 模型的项布局图

图1 项布局图

由以上可以看出,PR 模型比 support-confidence 模型在实际应用中能够为决策者提供更合理的信息。以下我们将看到 PR 模型是度量关联规则的不确定性的有效方法。

定理1 I 为数据库 D 中的所有项的集合, $X, Y \subseteq I, X \cap Y = \emptyset, p(X) \neq 0, p(Y) \neq 0$, minsup 和 minconf 分别是由用户给定的最小支持度和最小置信度, λ 为最小兴趣度, $X \rightarrow Y$ 为兴趣规则, 若

(1) $p(X \cup Y) \geq \text{minsupp}$;

(2) $\text{PR}(Y|X) \geq \text{minconf}$.

证明: 由于 $\text{PR}(Y|X) \geq \text{minconf}$, 根据 PR 的定义, 有

$$\begin{aligned} p(X \cup Y) - p(X)p(Y) &\geq \lambda \\ \text{PR}(Y|X) &= \frac{p(Y|X) - p(Y)}{1 - p(Y)} \geq \text{minconf} \\ p(Y|X) - p(Y) &\geq \text{minconf}(1 - p(Y)) \end{aligned}$$

故 $p(Y|X) \geq \text{minconf} + p(Y)(1 - \text{minconf})$

由于 $1 - \text{minconf} \geq 0$, 因此 $p(Y|X) \geq \text{minconf}$

根据 Piatetsky-shapiro 的观点, $X \rightarrow Y$ 为一感兴趣的规则.

以下定理证明类似, 从略.

定理2 I 为数据库 D 中的所有项的集合, $X, Y \subseteq I, X \cap Y = \emptyset, p(X) \neq 0, p(Y) \neq 0, \text{minsupp}$ 和 minconf 分别是由用户给定的最小支持度和最小置信度, λ 为最小兴趣度, $X \rightarrow Y$ 为兴趣规则, 若

(1) $p(X \cup Y) \geq \text{minsupp}$ $p(Y) \geq \text{minsupp}$;

(2) $\text{PR}(Y|X) < 0$ $\text{PR}(Y|X) \geq \text{minconf}$.

定理3 I 为数据库 D 中的所有项的集合, $X, Y \subseteq I, X \cap Y = \emptyset, p(X) \neq 0, p(Y) \neq 0, \text{minsupp}$ 和 minconf 分别是由用户给定的最小支持度和最小置信度, λ 为最小兴趣度, $X \rightarrow Y$ 为兴趣规则, 若

(1) $p(Y \cup X) \geq \text{minsupp}$ $p(X) \geq \text{minsupp}$;

(2) $\text{PR}(Y|X) \geq \text{minconf}$.

定理4 I 为数据库 D 中的所有项的集合, $X, Y \subseteq I, X \cap Y = \emptyset, p(X) \neq 0, p(Y) \neq 0, \text{minsupp}$ 和 minconf 分别是由用户给定的最小支持度和最小置信度, λ 为最小兴趣度, $X \rightarrow Y$ 为兴趣规则, 若

(1) $p(X - \cup Y) \geq \text{minsupp}$ $p(X) \geq \text{minsupp}$ $p(Y) \geq \text{minsupp}$;

(2) $\text{PR}(Y|X) < 0$ $\text{PR}(Y|X) \geq \text{minconf}$.

基于 PR 模型挖掘关联规则可以分解成以下两步进行:

(1) 产生所有的正大项集 PL 和负大项集 NL.

(2) 根据 PL 和 NL, 产生所有的规则

设 D 为数据库, I 为数据库 D 中的所有项的集合, 用户给定 $\text{minsupp}, \text{minconf}, \lambda > 0$, 我们有算法如下:

Algorithm RuleGeneration

Input: D : 数据库, $\text{minsupp}, \text{minconf}, \lambda$: 阈值

Output: $X \rightarrow Y$: 关联规则

(1) call procedure PNLargeItemsets

(2) for PL 中所有大项集 A do

for 所有满足 $X \cup Y = A$ 和 $X \cap Y = \emptyset$ 的项集 X, Y do

if $p(X \cup Y) - p(X)p(Y) \geq \lambda$ then

if $\text{PR}(Y|X) \geq \text{minconf}$ then

output 规则 $X \rightarrow Y$, 置信度 $\text{PR}(Y|X)$, 支持度 $p(A)$

else if $p(X \cup Y) - p(X)p(Y) \leq -\lambda$ then

if $\text{PR}(Y|X) \leq -\text{minconf}$ then

if $p(X \cup Y) \geq \text{minsupp}$ then

output 规则 $X \rightarrow Y$, 置信度 $-\text{PR}(Y|X)$, 支持度 $p(X \cup Y)$

(3) for NL 中的所有项集 A do

for 所有满足 $X \cup Y = A$ 和 $X \cap Y = \emptyset$ 的项集 X, Y do

if $p(X) \geq \text{minsupp}$ & $p(Y) \geq \text{minsupp}$ & $p(X \cup Y) \geq \text{minsupp}$ then

if $p(X \cup Y) - p(X)p(Y) \geq \lambda$ then

if $\text{PR}(Y|X) \geq -\text{minconf}$ then

output 规则 $X \rightarrow Y$, 置信度 $\text{PR}(Y|X)$, 支持度 $p(A)$

else if $p(X \cup Y) - p(X)p(Y) \leq -\lambda$ then

if $\text{PR}(Y|X) \leq -\text{minconf}$ then

if $p(X \cup Y) \geq \text{minsupp}$ then

output 规则 $X \rightarrow Y$, 置信度 $-\text{PR}(Y|X)$, 支持度 $p(X \cup Y)$

总结 目前的关联规则挖掘模型还存在许多不足, 离实际应用还有一段距离. 本文给出了一个能同时挖掘正关联规则和负关联规则的算法, 并在概率论和 Piatetsky-Shapiro 模型的基础上构建了一个关联规则的置信度以度量关联规则的不确定性.

参考文献

- 1 Agrawal R, Mielinski T, Swami A. Mining association rules between sets of items in large database. In: Proc. of the ACM SIGMOD Conf. on Management of Data. Washington D. C., 1993. 207~216
- 2 Brin S, Motwani R, Silverstein C. Beyond Market Baskets: Generalizing Association Rules to Correlations. In: Proc. of the ACM SIGMOD Intl. Conf. on Management of Data. Montreal, Canada. 1997. 265~276
- 3 Piatetsky-Shapiro G. Discovery, analysis and presentation of strong rules. Knowledge Discovery in Databases, AAAI Press/MIT, 1991. 229~248
- 4 Srikant R, Agrawal R. Mining generalized association rules. Future Generation Computer Systems, 1997, 13: 161~180
- 5 Aggarawal C, Yu P. A new framework for itemset generation. In: Proc. of the ACM PODS, Santiago, Chile, 1998