第 33 卷 第 2 期 计 算 机 工 程 2007 年 1 月
Vol.33 No.2 Computer Engineering January 2007
·软件技术与数据库· 文章编号：1000—3428(2007)02—0056—03 文献标识码：A 中图分类号：TP311

一种用于挖掘正负关联规则的可量化标准

赵 亮[1]，萧德云[1]，刘震涛[2]
(1. 清华大学自动化系，北京 100084；2. 清华大学台湾研究所，北京 100084)

摘 要：传统的关联规则数据挖掘的支持度-置信度框架存在着弊端。一方面，它不能发现同时满足支持度和置信度而其前提和结论却相互独立的规则；另一方面，也不能区分正负关联规则。该文提出了一种评价关联规则的可量化的标准，进一步提出一种同时挖掘正负关联规则的框架，用此框架来分析调研问卷。

关键词：负关联规则；相关性；数据挖掘；问卷

Quantitative Criterion for Mining Both Positive and Negative Association Rules

ZHAO Liang[1], XIAO Deyun[1], LIU Zhentao[2]
(1. Department of Automation, Tsinghua University, Beijing 100084; 2. Institute of Taiwan Studies, Tsinghua University, Beijing 100084)

【Abstract】 The conventional framework for mining association rules is the support-confidence framework which has some limitations. For one thing, it can not prune such useless rules as satisfy both minimum support and minimum confidence with their antecedents and consequents independent. For another thing, it can neither separate the negative association rules from the positive ones. The purposes of this paper are to find a quantitative criterion for mining both positive and negative association rules, and further put forward a novel framework whose efficiency is tested in analyzing a questionnaire.

【Key words】 Negative association rules; Correlation; Data mining; Questionnaire

关联规则的数据挖掘是从大量数据中挖掘出内在联系的方法，常用在购物篮分析中。在调研问卷分析中也常有类似的需求：哪些调查题目之间存在着内在的联系？本文尝试用关联规则数据挖掘方法来分析一份甲地区对乙地区看法的调研问卷。现在对于关联规则的挖掘多是侧重于基于支持度-置信度框架的正关联规则的挖掘。设有事件 A 和 B，正关联规则类似于“事件 A 导致事件 B”，可以写成 $A \Rightarrow B$ 的形式。如果得到这样的规则“事件 A 将导致 B 不发生”，这就是负关联规则，记作 $A \Rightarrow \neg B$(符号“$\neg$”表示“非”或“不”等否定含义)。对于负关联规则相关研究却比较少，但是负关联规则同样可以给我们带来有价值的规则。本文提出一个分类正、负关联规则可量化的标准，并藉此提出频繁集中挖掘有用正负关联规则的框架。

1 负关联规则挖掘

1.1 支持度-置信度框架

评判关联规则是否有用的框架中，一般比较流行的是R. Agrawal提出的支持度-置信度框架(support-confidence)[1]。

设 $I = \{i_1, i_2, \ldots, i_m\}$ 是项的集合。设任务相关的数据 D 是数据库事务的集合，其中每个事务 T 是项的集合，使得 $T \subseteq I$。每个事务有一个标识符，称作 TID。设 A 是一个项集，事务 T 包含 A 当且仅当 $A \subseteq T$。关联规则是形如 $A \Rightarrow B$ 的形式，其中 $A \subset I$，$B \subset I$，并且 $A \cap B = \varnothing$。规则 $A \Rightarrow B$ 在事物 D 中成立，具有支持度 s，其中 s 是 D 中事务包含 $A \cup B$ 的百分比。它的概率为 $P(A \cup B)$。规则 $A \Rightarrow B$ 在事务集 D 中具有置信度 c，如果 D 中包含的事务同时也包含 B 的百分比是 c。这是条件概率 $P(B|A)$，即

$$supp(A \cup B) = P(A \cup B)$$

$$conf(A \Rightarrow B) = P(B|A)$$

关联规则满足最小支持度，如果 s 频率大于或等于 min_conf；关联规则满足最小置信度，如果 c 频率大于或等于 min_conf。同时满足最小支持度(min_supp)和最小置信度(min_conf)的规则称为强规则。

1.2 支持度-置信度框架的缺陷

在调研问卷的数据挖掘中，可以发现仅仅考虑正关联规则是不够的。假设所考虑集合中有 10 000 个数据项，其中包含事件 A 和 B，同时包含事件 A 和 B 二者记为 $A \cup B$。设包含 A 的有 6 000 项，包含 B 的有 5 000 项，包含 $A \cup B$ 的有 2 500 项，min_sup=0.2，min_conf=0.3。那么 $supp(A \cup B)$ = 0.25>min_supp，$conf(A \Rightarrow B)$ =0.42>min_conf，得到 $A \Rightarrow B$ 是强关联规则。再考虑 $A \Rightarrow \neg B$。$supp(A \cup \neg B)$ = 0.35>min_supp，$conf(A \Rightarrow \neg B)$=0.58>min_conf，这说明 $A \Rightarrow \neg B$ 也是强关联规则。然而这与 $A \Rightarrow B$ 相矛盾，同时由于 $conf(A \Rightarrow \neg B) > conf(A \Rightarrow B)$，$A \Rightarrow \neg B$ 应该更可靠一些。

这是由于 $conf(A \Rightarrow B) = P(B|A) < P(B)$，说明由于 A 的发生 B 发生的概率反而下降了，因此 A 和 B 应该是相互削弱的关系。然而这是支持度-置信度框架所无法识别的。

1.3 负关联规则

设任务相关的数据 D 是数据库事务的集合中有项集 A 和

作者简介：赵 亮(1980－)，男，硕士，主研方向：数据挖掘；萧德云，博士、教授、博导；刘震涛，教授

收稿日期：2006-01-29 **E-mail：**zhaoliang@sjtu.org

项集 B。形如 $A \Rightarrow \neg B$，$\neg A \Rightarrow B$，$\neg A \Rightarrow \neg B$ 的关联规则称为负关联规则(negative association rules)。

负的关联规则的支持度和置信度的定义和正关联规则相似，不过是分别用 $\neg A$ 和 $\neg B$ 分别代替原式中的 A 和 B。它们的支持度可用如下方法计算：

$$supp(A \cup \neg B) = supp(A) - supp(A \cup B)$$
$$supp(\neg A \cup B) = supp(B) - supp(A \cup B)$$
$$supp(\neg A \cup \neg B) = 1 - supp(A) - supp(B) + supp(A \cup B)$$

它们的置信度可用如下方法计算：

$$conf(A \Rightarrow \neg B) = \frac{supp(A) - supp(A \cup B)}{supp(A)}$$
$$conf(\neg A \Rightarrow B) = \frac{supp(B) - supp(A \cup B)}{1 - supp(A)}$$
$$conf(\neg A \Rightarrow \neg B) = \frac{1 - supp(A) - supp(B) + supp(A \cup B)}{1 - supp(A)}$$

2 支持度-有效度框架

2.1 相关研究

早在 1991 年Piatetsky Shapiro[2]就指出当支持度满足式(1)时，$A \Rightarrow B$ 没有意义。

$$supp(A \cup B) \approx supp(A) \times supp(B) \quad (1)$$

在 1997 年，Brin[3]等人使用A和B的相关性的概念来剔除没有意义的挖掘结果和负相关结果，但是他们没有注意到保留负相关规则。后面也有这方面的研究[5,6]，但大都基于式(2)相关性概念：

$$corr_{A,B} = \frac{P(A \cup B)}{P(A)B(B)} \quad (2)$$

$corr_{A,B}$ 的取值范围是 $[0, \infty)$。当 $corr_{A,B} < 1$ 时，A 和 B 为负相关，表示二者相互降低了对方出现的可能性；当 $corr_{A,B} > 1$ 时，A 和 B 是正相关，表示二者相互提升了对方出现的可能性；当 $corr_{A,B} = 1$ 时 A 和 B 相互独立，表示 A 和 B 概率上没有关系。

可见 $corr_{A,B}$ 可以判断 A 和 B 是否相互影响，并且 $corr_{A,B}$ 的大小可以判断相互影响的程度。

在文献[4]中已经证明，当 $corr_{A,B} > 1$ 时，$corr_{\neg A,\neg B} > 1$，$corr_{A,\neg B} < 1$，$corr_{\neg A,B} < 1$。也就是说当 $corr_{A,B} > 1$ 时，只考虑 A 和 B，$\neg A$ 和 $\neg B$ 之间的关联规则；当 $corr_{A,B} < 1$ 时，只考虑 $\neg A$ 和 B，A 和 $\neg B$ 之间的关联规则。

2.2 有效度

$corr_{A,B}$ 的一个缺点是取值在临界值 1 两侧不对称，而且取值范围不确定(随频繁项的不同而不同)，这对于合理的设置门限值造成了困难。所以把 $corr_{A,B}$ 加以改造，将其映射到[-1,1]这个区间，得到的新的评价标准称之为有效度。定义如下：

$$crit(A \cup B) = \begin{cases} corr_{A,B} - 1, & corr_{A,B} < 1\text{时} \\ 1 - corr_{A,\neg B}, & corr_{A,B} \geq 1\text{时} \end{cases} \quad (3)$$

当 $corr_{A,B} < 1$ 时，A 和 B 负相关，$-1 \leq corr_{A,B} - 1 < 0$，并且当 A 和 B 负相关度越大，即 $corr_{A,B}$ 越小，而 $crit(A \cup B)$ 越趋近于-1，特别是当 A 和 B 负相关度最大时，$crit(A \cup B) = -1$；当 $corr_{A,B} \geq 1$ 时，A 和 B 正相关，$0 < 1 - corr_{A,\neg B} \leq 1$，并且当 A 和 B 正相关度越大，即 $corr_{A,B} = P(B|A)/P(B)$ 越大，而 $corr_{A,\neg B} = [1 - P(B|A)]/[1 - P(B)]$ 越小，$crit(A \cup B)$ 越趋近于 1，特别是当 A 和 B 正相关度最大时，$crit(A \cup B) = 1$，当 A 和 B 相独立时，$crit(A \cup B) = 0$。

将 $corr_{A,B}$ 改写成

$$crit(A \cup B) = \begin{cases} corr_{A,B} - 1, & corr_{A,B} < 1\text{时} \\ \frac{P(B)}{1 - P(B)}(corr_{A,B} - 1), & corr_{A,B} \geq 1\text{时} \end{cases} \quad (4)$$

设 $P(B)$为常数，则 $corr_{A,B}$ 的取值范围是[0, $1/P(B)$]，根据式(4)得到图 1。图 1 反映了 $crit(A \cup B)$ 和 $corr_{A,B}$ 的映射关系。

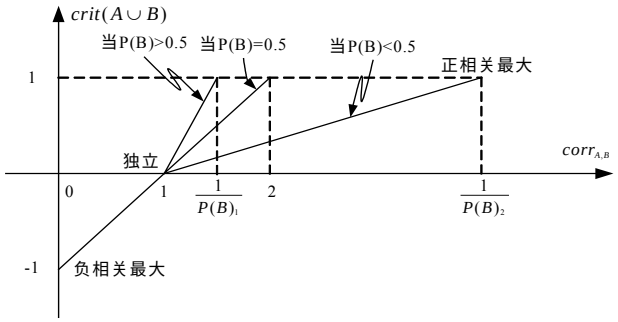


图 1 crit(A∪B)**和** $corr_{A,B}$ **的映射关系**

具体计算时的 $crit(A \cup B)$ 可写成如下形式：

$$crit(A \cup B) = \begin{cases} \frac{supp(B \cup A)}{supp(A)supp(B)} - 1, & supp(B \cup A) < supp(B)supp(A)\text{时} \\ \frac{supp(B \cup A) - supp(B)supp(A)}{supp(A)[1 - supp(B)]}, & supp(B \cup A) \geq supp(B)supp(A)\text{时} \end{cases} \quad (5)$$

当 $crit(A \cup B) \in [-1, 0)$ 时 A 和 B 负相关，当 $crit(A \cup B) \in (0, 1]$ 时 A 和 B 是正相关，当 $crit(A \cup B) = 0$ 时 A 和 B 相互独立。实际上当 $|crit(A \cup B)|$ 接近于 1 时关联项可以舍弃。即设置 $0 < min_crit_P < 1$ 和 $0 < min_crit_N < 1$，关联项需满足以下最小相关度要求

$$\begin{cases} -crit(A \cup B) \geq min_crit_N, \text{当} crit(A \cup B) < 0 \\ crit(A \cup B) \geq min_crit_P, \text{当} crit(A \cup B) > 0 \end{cases} \quad (6)$$

为方便，也可取 $min_crit_P = min_crit_N = min_crit$，则上述条件可写成 $|crit(A \cup B)| \geq$min_crit。

2.3 支持度-有效度框架

在生成相关规则时使用有效度的概念来代替置信度概念，这就是支持度-有效度框架(support-crit)。首先计算支持度，设置 min_supp、min_corr_P 和 min_corr_N，选出 support ≥min_supp 的频繁项集。然后在生成关联规则时，根据式(5)计算 $crit(A \cup B)$，再根据式(6)判断是否满足最小相关度要求。如果满足，判断是正相关还是负相关。正相关考虑(A,B)和($\neg A$, $\neg B$)的关联规则，负相关考虑(A, $\neg B$)和($\neg A$,B)的关联规则。

3 算法

记 FS：频繁项集；min_supp：最小支持度；min_conf：最小置信度；min_crit_P：最小正相关支持度；min_crit_N：最小负相关支持度；AR_P：正关联规则集；AR_N：负关联规则集。

(1)使用 Apriori 算法生成满足最小支持度 min_supp 的 FS。

(2)AR_N=Φ; AR_P=Φ;

for any itemset X in FS do begin

  for any itemset A∪B=X and A∩B=Φ do begin

    if supp(A∪B)<supp(A)supp(B) then

      if -crit≥min_crit_N then

        conf ($A \Rightarrow \neg B$) =supp ($A \cup \neg B$) /supp(A)

conf ($\neg A \cup B$) =supp ($\neg A \cup B$) /supp ($\neg A$)
if conf ($A \Rightarrow \neg B$) ≥min_conf
AR_N=AR_N∪($A \Rightarrow \neg B$)
end
if conf ($\neg A \Rightarrow B$) ≥min_conf
AR_N=AR_N∪($\neg A \Rightarrow B$)
end
end
else
if crit≥min_crit_P then
conf ($A \Rightarrow B$) =supp(A∪B)/supp(A)
conf ($\neg A \Rightarrow \neg B$) =supp(A∪B)/supp(A)
if conf ($A \Rightarrow B$) ≥min_conf
AR_P=AR_P∪($A \Rightarrow B$)
end
if conf ($\neg A \Rightarrow \neg B$) ≥min_conf
AR_N=AR_N∪($\neg A \Rightarrow \neg B$)
end
end
end
end
end
(3)return AR_P andAR_N

4 调研问卷中的关联规则挖掘

在调研问卷中的数据和购物篮数据有以下两点不同：(1)数据中有连续值，而不仅仅是 0 或 1 这样的离散值，这就要求首先要将连续值变成 0 和 1 这样的离散值；(2)调研问卷中一般目的明确，往往是为了找原因而有针对性地设计选项，因此可能的原因项和可能的结果项是先已分类，挖掘的目的是为了找出真正的原因项和结果项。这样只须将频繁项中含有可能结果项的频繁项集挑出即可。

这里仅拿出一个挖掘子任务的部分数据和挖掘结果来说明算法的有效性。略去项目细节，将数据抽象成仅包含 0 和 1 的向量。选了 4 个原因候选项(前 4 项)，2 个结论候选项(后两项)。组成 10 个 6 维向量形式项集。设 min_crit=0.1，min_supp=0.2，为了显示有效度的作用在新算法中让 min_conf=0，在原框架中让 min_conf=0.4。表 1 是测试用的数据项集。

表 1 数据集

TID	项集向量		TID	项集向量	
	1234	56		1234	56
T1	1101	00	T6	0101	01
T2	1111	00	T7	1000	11
T3	0101	00	T8	0010	01
T4	0111	10	T9	0110	01
T5	1010	10	T10	1111	01

表 2 和表 3 是结论为候选结论项的相关规则挖掘结果，为了对比我们把满足最小支持度的要求的所有结果列出。分别用置信度和有效度对结果筛选的得到的统计结果见表 4。

可见使用 support-confidence 框架无法区分关联关系的正负，而且错把独立的项作为有效项；而 support-crit 解决了这些问题。

表 2 由 2 项频繁集生成的挖掘结果

2 项频繁集				
频繁项集		支持度	置信度	有效度
1	5	0.2	0.4	0.143
1	6	0.2	0.4	-0.2
2	6	0.3	0.429	-0.143
3	5	0.2	0.333	0.111
3	6	0.3	0.5	0
4	6	0.2	0.333	-0.333

表 3 由 3 项频繁集生成的挖掘结果

3 项频繁集					
频繁项集			支持度	置信度	有效度
2	3	6	0.2	0.5	0
2	4	6	0.2	0.333	-0.333

表 4 分别用置信度和有效度对结果筛选的得到的统计结果

support-confidence 框架	support-crit 框架		
不区分正负	正	负	独立
5	2	4	2

5 结论

通过调研问卷的关联规则挖掘的计算结果，表明原先的支持度-置信度框架不能识别无效(即因果项独立)的关联规则，同时不能明确区分正负关联规则；本文所提出的支持度-有效度框架，通过有效度这一概念，定量描述关联规则的因果项的相互独立程度，并区分出正负关联规则。

参考文献

1 Agrawal R, Imielinski T, Swami A. Mining Association Rules Between Sets of Items in Large Databases[C]. Proceedings of the ACM SZGMOD International Conference on the Management of Data, 1993-05.

2 Piatetsky-Shapiro G. Discovery, Analysis, and Presentation of Strong Rules[C]. Proceedings of Knowledge Discovery in Databases, Menlo Park, Calif., USA: AAAI/MIT, 1991.

3 Brin S, Motwani R, Silverstein C. Beyond Market: Generalizing Association Rules to Correlations[C]. Proceedings of the ACM SIGMOID Conference, 1997.

4 董祥军，王淑静，宋瀚涛. 基于两级支持度的正负关联规则挖掘[J]. 计算机工程, 2005, 31(10).

5 Savasere A, Omiecinski E, Navathe S. Mining for Strong Negative Associations in a Large Database of Customer Transactions[C]. Proc. of ICDE, 1998.

6 Antonie M, Zaïane O R. Mining Positive and Negative Association Rules: An Approach for Confined Rules[C]. Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, Pisa, Italy, 2004-09.