

负关联规则挖掘算法研究

朱红蕾, 徐志刚, 李 明

(兰州理工大学 计算机与通信学院, 甘肃 兰州 730050)

摘 要: 文中首先描述了负关联规则的基本概念, 接着分析了负关联规则挖掘的特点和相关算法, 并提出基于相关性和剪切策略的改进算法来进行正、负关联规则挖掘, 最后指出了负关联规则挖掘的研究方向。

关键词: 数据挖掘; 负关联规则; 频繁项集; 相关性

中图分类号: TP311

文献标识码: A

文章编号: 1000-7180(2010)05-0167-03

Research on Negative Association Rules Mining Algorithms

ZHU Hong-lei, XU Zhi-gang, LI Ming

(College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, China)

Abstract: This paper presents the basic concepts of negative association rules firstly. Then the characteristic of negative association rules is analyzed and some algorithms are introduced. With a correlation coefficient measure and pruning strategies, we present an efficient algorithm for mining both positive and negative association rules in databases. At last, we point out the research object.

Key words: data mining; NAR; frequent itemset; correlation coefficient

1 引言

关联规则 (Association Rules, AR) 挖掘是 R. Agrawal^[1] 等人首先提出的数据挖掘领域研究的一个重要课题。传统的关联规则挖掘是形如 $X \Rightarrow Y$ 的蕴含关系, 用于挖掘数据集中项集之间的关联关系, 并于 1994 提出了一种快速算法^[2]。以后诸多研究人员对原有的算法进行改进, 以提高算法挖掘规则的效率。在数据集中, 除了形如 $X \Rightarrow Y$ 的蕴含关系, 还存在着形如 $X \Rightarrow \neg Y$ 、 $\neg X \Rightarrow Y$ 、 $\neg X \Rightarrow \neg Y$ 三种形式的关联规则, 这三种形式的蕴含关系称为负关联规则^[3] (Negative AR, NAR)。传统的形如 $X \Rightarrow Y$ 的蕴含关系称为正关联规则 (Positive AR, PAR)。负关联规则挖掘的研究起步较晚。1997 年 Brin 等人首次提到了两个频繁项集间的负相关^[4]。Savasere 等人研究了负相关性的项形成的强负关联规则问题^[5], 文献^[6]提出用遗传算法来进行负关联规则挖掘。诸

多学者对负关联规则进行了相关研究。文中通过分析负关联规则挖掘的特点, 提出了基于相关性和剪切策略的改进算法。该算法能同时进行正、负关联规则挖掘, 而且在很大程度上避免产生相互矛盾的规则, 提高了算法的效率。

2 相关知识

2.1 基本概念

设 $I = \{i_1, i_2, \dots, i_m\}$ 是 m 个不同项目 (item) 的集合, $D = \{T_1, T_2, \dots, T_n\}$ 是事务数据库, 其中每一个事务 T_i 是 I 中一些项目的集合, 即 $T_i \subseteq I$ 。每个事务 T_i 都有一个唯一的标识符, 记作 TID。如果对于 I 中的一个子集 X , 有 $X \subseteq T$, 那么称事务 T 包含 X 。 $\neg X$ 被称为负项集, 表示一个事务中不存在 (Non-existing) 项集 X 。一条负关联规则就是一个形如 $X \Rightarrow \neg Y$ (或 $\neg X \Rightarrow Y$, 或 $\neg X \Rightarrow \neg Y$) 的蕴涵式, 其中 $X, Y \subseteq I$, 而且 $X \cap Y = \emptyset$ 。

收稿日期: 2009-04-07; 修回日期: 2009-07-12

基金项目: 甘肃省自然科学基金 (2007GS04782)

项集 X 的支持度 (support, 简记为 supp) 是指事务集中包含 X 的事务数与所有事务数之比, 记为 $\text{supp}(X)$, 它是概率 $P(X)$; 规则 $X \Rightarrow Y$ 在事务数据库 D 中的支持度是事务集中包含 X 和 Y 的事务数与所有事务数之比, 记为 $\text{supp}(X \cup Y)$, 它是概率 $P(X \cup Y)$, 即

$$\text{supp}(X \cup Y) = P(X \cup Y) = |\{T : X \cup Y \subseteq T, T \in D\}| / |D|$$

其中, $|D|$ 表示事务数据库 D 中的事务数。

规则 $X \Rightarrow Y$ 在事务数据库 D 中的置信度 (confidence, 简记为 conf) 是指同时包含 X 和 Y 的事务数与包含 X 的事务数之比, 记为 $\text{conf}(X \Rightarrow Y)$, 它是条件概率 $P(Y | X)$, 即: $\text{conf}(X \Rightarrow Y) = P(Y | X) = \text{supp}(X \cup Y) / \text{supp}(X)$ 。

负关联规则的支持度和置信度可以利用正关联规则的相关信息来计算^[7]。支持度计算方法如下:

- (1) $\text{supp}(\neg X) = 1 - \text{supp}(X)$;
- (2) $\text{supp}(X \cup \neg Y) = \text{supp}(X) - \text{supp}(X \cup Y)$;
- (3) $\text{supp}(\neg X \cup Y) = \text{supp}(Y) - \text{supp}(X \cup Y)$;
- (4) $\text{supp}(\neg X \cup \neg Y) = 1 - \text{supp}(X) - \text{supp}(Y) + \text{supp}(X \cup Y)$ 。

根据置信度的定义, 很容易得到负关联规则的置信度计算方法。

- (1) $\text{conf}(X \Rightarrow \neg Y) = \frac{\text{supp}(X) - \text{supp}(X \cup Y)}{\text{supp}(X)}$;
- (2) $\text{conf}(\neg X \Rightarrow Y) = \frac{\text{supp}(Y) - \text{supp}(X \cup Y)}{1 - \text{supp}(X)}$;
- (3) $\text{conf}(\neg X \Rightarrow \neg Y) = \frac{1 - \text{supp}(X) - \text{supp}(Y) + \text{supp}(X \cup Y)}{1 - \text{supp}(X)}$ 。

由上述各式可知存在如下关系:

- ① $\text{conf}(X \Rightarrow Y) + \text{conf}(X \Rightarrow \neg Y) = 1$;
- ② $\text{conf}(\neg X \Rightarrow Y) + \text{conf}(\neg X \Rightarrow \neg Y) = 1$ 。

2.2 相关性的定义和定理

当同时进行正、负关联规则挖掘时, 会发现挖掘的规则是互相矛盾的。例如: 可能挖掘出规则 $X \Rightarrow Y$ 和 $\neg X \Rightarrow Y$, 这两个规则是互相矛盾的。为了解决这些矛盾, 可以通过项集间的相关性来判断挖掘哪种类型的关联规则。

关联规则的相关性在文献[1]中进行了定义。设 X 和 Y 为两个项集, 它们的相关性可以由下式来

度量。

$$\text{corr}_{X,Y} = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)\text{supp}(Y)} \quad (1)$$

$\text{corr}_{X,Y}$ 的值有以下三种情况:

- (1) 如果 $\text{corr}_{X,Y} > 1$, 那么 X 和 Y 正相关, 项集 X 出现的越多, 项集 Y 出现的也越多;
- (2) 如果 $\text{corr}_{X,Y} = 1$, 那么 X 和 Y 相互独立, 项集 Y 的出现与项集 X 无关;
- (3) 如果 $\text{corr}_{X,Y} < 1$, 那么 X 和 Y 负相关, 项集 X 出现的越多, 项集 Y 出现的越少。

由相关性的定义, 给出下面的定理:

定理 1 如果项集 X 和 Y 正相关, 那么挖掘的关联规则形式为: $X \Rightarrow Y$ 或 $\neg X \Rightarrow \neg Y$ 。

定理 2 如果项集 X 和 Y 负相关, 那么挖掘的关联规则形式为: $X \Rightarrow \neg Y$ 或 $\neg X \Rightarrow Y$ 。

3 负关联规则挖掘

3.1 Apriori 经典算法

传统关联规则挖掘的经典算法是 Agrawal 等^[1]在 1993 年提出的 Apriori 算法。该算法基于两阶段频繁项集思想, 将关联规则挖掘分解为两个子问题:

- (1) 找到所有支持度大于最小支持度 (min_supp) 的项集 (Itemset), 这些项集称为频繁项集 L 。
- (2) 利用第(1)步找到的频繁项集 L 根据最小置信度 (min_conf) 产生期望的有趣的关联规则。即: 对于每一个 L 找出 L 所有非空子集 l , 如果 $\text{support}(L) / \text{support}(l) \geq \text{min_conf}$ 就生成有趣的关联规则 $l \Rightarrow (L - l)$ 。

其中, 第二个问题较为容易和直观, 目前大量的研究工作集中在如何高效地产生频繁项集。为了生成所有频繁项集, 算法中引入了 Apriori 性质改进生成所有频繁项集算法的性能。即: 任何频繁项集的子集一定是频繁项集。非频繁项集的父集一定是非频繁项集。

利用 Apriori 性质删除具有非频繁子集的候选项集, 降低计算所有候选项集的支持度的代价。另外, 研究人员还提出了很多优化方法, 集中在两个方面: 通过数据结构或存储的格式变换, 重新组织输入数据以便实现高效的计数; 或者通过减少扫描数据集的次数来提高性能。

3.2 负关联规则挖掘的特点

根据关联规则挖掘的经典算法 - Apriori 算法, 负关联规则挖掘就是挖掘出满足最小支持度和最小置信度这两个阈值的规则, 即有趣的负关联规则。

负关联规则挖掘的过程中有很多地方可以利用正关联规则挖掘的理论.但是,负关联规则挖掘也有其特殊性,集中表现在下面两个方面:

当同时进行正、负关联规则挖掘时,在如何判断挖掘出的关联规则是否有趣的度量上出现了分歧.传统的关联规则挖掘,通常使用最小支持度和最小置信度这两个阈值来衡量一条规则的有趣程度.而当同时挖掘正、负关联规则时,经常会出现两条相互矛盾的规则同时满足这两个参数,使得人们无法正常使用所挖掘出的关联规则.

另外,由于负关联规则包含不存在的项集,所以在进行挖掘时,如不加以限制,产生的负频繁项集会呈指数增长.这会影响正常挖掘,甚至会出现负频繁项集爆炸的问题.

因此,当进行负关联规则挖掘时,应该考虑到负关联规则的特殊性.

3.3 负关联规则挖掘的相关算法

负关联规则挖掘的研究热点主要集中在如何对频繁项集生成的正、负关联规则进行度量;如何生成负频繁项集和解决负频繁项集爆炸的问题;以及如何避免挖掘出虚假的关联规则和相互矛盾的关联规则^[8]等.

在进行正、负关联规则挖掘时,可以对正关联规则和各种形式的负关联规则分别设置不同的最小置信度^[9],该算法虽然考虑了正负关联规则的不同,但是如何设置最小置信度的值是个瓶颈;文献^[7]采用对满足最小支持度和最小置信度的规则再进行相关度计算的方法,但是算法的执行过程中需要不断地调整关联系数,影响了算法的效率;文献^[3]提出了一种基于兴趣度的正、负关联规则挖掘算法,该算法在生成候选项目集时采用了类 Apriori 算法,这样可能会丢失部分有用的非频繁项集.这些方法在一定程度上解决了一些问题,但是都存在各自的局限性,容易遗漏部分有价值的关联规则.

对于负关联规则挖掘中产生的负频繁项集,大都采用了正频繁项集的生成方法.为了有效解决负频繁项集爆炸问题,引入过频繁项集^[10]的概念,并对项集中负项目的个数进行限制,较好地解决了负关联规则挖掘中频繁项集数量过快增长的问题.但是,由于必须恰当的设置两个参数,存在一定的局限性.

为了避免挖掘出虚假的关联规则和相互矛盾的关联规则,提出了一种新的关联规则兴趣度度量方

法——提升率^[11]来代替置信度.该方法较好地避免了可能导致误导的负关联规则的产生,但是并不能降低挖掘负频繁项集的时间开销,仍然可能产生无意义的关联规则.

通过分析负关联规则的三种形式和现有的算法,发现形如 $\neg X \Rightarrow \neg Y$ 的规则虽然在负关联规则中占很大比例,但在实际决策时,不会提供太多有价值的信息.例如:在超市购物篮分析时, X 和 Y 分别代表两个物品,那么 $\neg X \Rightarrow \neg Y$ 就表示顾客不买 X 也不会买 Y .这样的规则对现实决策没有实质性的帮助,所以在挖掘负关联规则时不再考虑这种形式的规则.根据这种剪切策略,可以得到如下的推理:

推理 1 如果项集 X 和 Y 正相关,那么挖掘的关联规则形式为: $X \Rightarrow Y$.

通过推理 1 和定理 2,在同时进行正、负关联规则挖掘时,可以避免产生互相矛盾的规则,提高挖掘的效率.

4 结束语

负关联规则挖掘已经成为关联规则挖掘的一个热点问题.现有的算法在一定程度上解决了负关联规则挖掘的问题,但都是考虑数据库不改变的情况.实际上,数据挖掘的过程是一个动态的交互过程.为了发现用户满意的模式,需要不断调整相应的参数.与此同时,数据库也是不断进行更新的.因此,如何对负关联规则进行合理的度量,如何高效即时获得满意的结果,如何对负关联规则进行维护更新将成为今后的研究方向.

参考文献:

- [1] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[C]// Proc. of the ACM SIGMOD Int. Conf. on Management of data (ACM SIGMOD '93). Washington, USA, 1993(3):207-216.
- [2] Agrawal R, Srikant R. Fast algorithms for mining association rules[C]//Proc. of the 20th Int. Conf. on Very Large Databases(VLDB'94). Santiago, Chile, 1994:487-499.
- [3] Wu Xindong, Zhang Chengqi, Zhang Shichao. Mining both positive and negative association rules[C]// Proceedings of the 19th International Conference on Machine Learning (ICML-2002). San Francisco: Morgan Kaufmann Publishers, 2002:658-665.

一同作为模糊聚类的特征向量,并应用于示温漆彩色图像分割中.实验证明该算法取得了较好的分割效果,为示温漆图像分析处理提供了较好的基础.

参考文献:

- [1] Dunn J C. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters[J]. J. Cybernet, 1973,3(3):32-57.
- [2] 王洪春,彭宏.基于模糊C-均值的增量式聚类算法[J].微电子学与计算机,2007,24(6):156-157.
- [3] 李云松,李明.基于灰度空间特征的模糊C均值聚类图像分割[J].计算机工程与设计,2007,28(6):1358-1360.
- [4] 陈武凡.小波分析及其在图像处理中的应用[M].北京:科学出版社,2002:12-35.
- [5] Deng Y, Charles Kenney, Michael S Moore, et al. Peer

group filtering and perceptual color image quantization [C]// Proceeding of IEEE International Symposium on Circuits and Systems. Florida, 1999(4):21-24.

- [6] 龚巍.一种改进的边缘生长彩色图像分割方法[J].微电子学与计算机,2007,24(6):194-197.
- [7] 李世雄.小波变换及其应用[M].北京:高等教育出版社,1997:30-52.
- [8] Rioul O. Fast algorithm for discrete and continuous wavelet transform[J]. IEEE Trons. Information Theory, 1992,38(2):569-595.

作者简介:

蔡茂蓉 女,(1977-),硕士,讲师.研究方向为图形图像处理.

龚巍 男,(1980-),硕士,工程师.研究方向为图形图像处理.

(上接第169页)

- [4] Brin S, Motwani R, Silverstein C. Beyond market: generalizing association rules to correlations[C]// Processing of the ACM SIGMOD Conference 1997. New York: ACM Press, 1997:265-276.
- [5] Savasere A, Omiecinski E, Navathe S. Mining for strong negative associations in a large database of customer transaction[C]// Proceedings of the IEEE 14th Int. Conference on Data Engineering. Los Alamitos: IEEE-CS, 1998: 494-502.
- [6] Boulicaut J F, Bykowski A, Jeudy B. Towards the tractable discovery of association rules with negations[C]// Proceedings of the Fourth International Conference on Flexible Query Answering Systems FQA S'00. Heidelberg: Physica-Verlag, 2000:425-434.
- [7] 董祥军,王淑静.负关联规则的研究[J].北京理工大学

学报,2004,24(11):978-981.

- [8] 张卫国,高贤强,武晓宏.基于支持度-置信度框架的负关联规则研究[J].微电子学与计算机,2009,26(4):102-104.
- [9] 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):16-18.
- [10] 马占欣,陆玉昌.负关联规则挖掘中的频繁项集爆炸问题[J].清华大学学报:自然科学版,2007,47(7):1212-1215.
- [11] 马占欣,王新社.对最小置信度门限的置疑[J].计算机科学,2007,34(6):216-218.

作者简介:

朱红蕾 女,(1977-),硕士,讲师.研究方向为数据挖掘、人工智能.