

基于矩阵的 Apriori 算法改进

宋文慧, 高建瓴

(贵州大学 大数据与信息工程学院, 贵州 贵阳 550025)

摘要:文中介绍了经典 Apriori 算法的原理、思想和步骤,以及基于矩阵的 Apriori 算法。针对 Apriori 算法需要多次扫描数据库和产生大量候选项集的缺点,提出了一种基于矩阵的 Apriori 算法的改进方法。该方法的不同之处在于矩阵的构建方法,通过对事务数据库的一次整体扫描,把事务数据库中的数据转换成一个上三角矩阵,然后通过访问上三角矩阵中的元素就可直接得到频繁 1 项集和频繁 2 项集,再根据经典的 Apriori 算法,利用频繁 2 项集得到频繁 3 项集,依此进行下去。该算法因为有上三角矩阵的引入,故可以适当减少访问事务数据库的次数,同时还减少了大量候选项集的产生,尤其是二次候选项集,节约了存储空间。实验结果表明,该改进算法是有效的,减少了使用扫描数据库的函数的次数,并且保证了频繁项集的准确性。

关键词:关联规则; Apriori 算法; 矩阵; M-Apriori 算法

中图分类号: TP301.6

文献标识码: A

文章编号: 1673-629X(2016)06-0062-03

doi: 10.3969/j.issn.1673-629X.2016.06.013

Improved Apriori Algorithm Based on Matrix

SONG Wen-hui, GAO Jian-ling

(College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China)

Abstract: It introduces the principles, ideas and steps of the classical Apriori algorithm, as well as the Apriori algorithm based on matrix in this paper. In view of the shortcomings for traditional Apriori algorithm of requiring multiple scanning database and producing a large number of candidate itemsets, an improved Apriori algorithm based on matrix is proposed. The difference of this method is the method to construct a matrix, through a whole scan of the transaction database, the transaction data in the database into an upper triangular matrix, and then by accessing the elements in the upper triangular matrix can obtain frequent itemsets 1 and frequent itemsets 2 directly. According to the classical Apriori algorithm, using the frequent itemsets 2 get frequent itemsets 3, proceeding accordingly. Because of the introduction of upper triangular matrix, the improved algorithm can reduce the number of accessing database and the incidence of a large number of candidate itemsets, especially the candidate itemsets 2, saving storage space. The experiment shows that the improved algorithm is effective to reduce the number of functions used to scan the database and to ensure the accuracy of the frequent itemsets.

Key words: association rule; Apriori algorithm; matrix; M-Apriori algorithm

1 概述

数据挖掘,通俗来说,是从大型的数据中找出其内在隐含的信息,而内在的信息可以用关联规则或频繁项集来表示。其中,频繁模式挖掘是关联规则、相关性分析、序列模式、因果关系、情节片段、局部周期性、显露模式等许多重要数据挖掘任务的基础^[1]。关联规则是数据挖掘的众多模式中最重要的一种,通过对数据项集间的关联性进行分析和挖掘,挖掘出在决策制定过程中具有重要参考价值的信息。关联规则经常被用于市场营销中,从交易数据库中可挖掘出不同商品

(项)之间的联系,找出顾客的购买行为模式,再将这些购买行为模式用在营销策略上,从而提高商品销售量,故又称为购物篮分析^[2]。

Apriori 算法^[3]是在 1994 年由 Agrawal 和 R. Srikant 共同提出的,是第一个关联规则挖掘算法,具有很大的影响力,但算法也存在一些不足之处^[4-6]。Apriori 算法是挖掘布尔关联规则频繁项集的算法,利用频繁项集性质的先验知识,通过逐层搜索的迭代方法,即将 k 项集用于探索 $(k+1)$ 项集,来穷尽数据集中的所有频繁项集。算法过程是^[7]:连接→剪枝→生成 C_k →扫

收稿日期:2015-09-16

修回日期:2015-12-18

网络出版时间:2016-05-25

基金项目:贵州省科学技术基金项目(黔科合 J 字[2015]2045)

作者简介:宋文慧(1992-),女,硕士研究生,研究方向为数据挖掘、聚类分析;高建瓴,硕士研究生导师,研究方向为数据挖掘、云计算。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20160525.1709.050.html>

描计数→比较→生成 L_k 。Apriori 算法采用两阶段挖掘的思想,并且基于多次扫描事务数据库来执行。面对存储了海量数据的事务数据库,这无疑将消耗大量的时间和内存空间,也导致了效率较低,这也是 Apriori 算法的瓶颈所在。

文献[8]提出基于粗糙集对 Apriori 算法进行改进。该方法是利用项集分类预处理来对事务数据库中的所有项集进行预处理。文献[9]提出一种 0-1 矩阵的改进算法,改变由低维频繁项目集到高维频繁项目集的多次连接运算。文献[10]改变了频繁 1 项集的排列方式,进而减少了候选项集的产生。文献[11]则消除大量冗余的非频繁项集。

文中在对上述几种算法的研究基础上,通过对矩阵的不同定义,提出了一种新的基于矩阵的 Apriori 改进算法。与文献[9]提到的矩阵的不同之处在于,该改进算法的矩阵是以项集中的项作为矩阵相应的行标和列标,通过一次遍历矩阵即可直接得到频繁 1 项集和频繁 2 项集,从而减少了扫描数据库的次数以及大量候选项集的产生,在时间和内存空间方面有一定的改善。

2 经典的 Apriori 算法

2.1 Apriori 算法的基本思想

Apriori 算法作为一种经典算法,无疑占据重要地位。其基本思想如下:

(1)通过扫描事务数据库中的所有数据项集,得到候选 1-项集,记为 C_1 ,并统计出相应数据项出现的频数。按照预先定义的最小支持度,从 C_1 中筛选出符合要求的频繁 1-项集,记为 L_1 。

(2)通过频繁 1-项集 L_1 的自连接得到候选 2-项集 C_2 ,接着再扫描事务数据库,统计出对应的频数,使之和最小支持度相比较,得到 L_2 。

(3)重复上述步骤,直到不再存在满足要求的频繁 K -项集为止。

从上述可知,Apriori 算法使用逐层搜索的迭代方法,通过低维频繁项目集产生高维频繁项目集^[3]。每次挖掘一层频繁项集 L_k ,就需要扫描一次整个事务数据库。在此过程中,还要生成大量的候选项集,因此 Apriori 算法的效率低。

2.2 Apriori 算法的关键步骤

Apriori 算法是通过迭代的方法,由 L_{k-1} 找 L_k ,关键步骤为:连接、剪枝。

(1)连接步:为得到 L_k ,需要将 L_{k-1} 与自身连接生成候选- K 项集,执行的前提是 L_{k-1} 的元素是可以连接的^[12]。

(2)剪枝步: C_k 的成员可以是也可以不是频繁的,

但所有频繁 K -项集都包含在 C_k 中,通过扫描数据库,确定 C_k 中每个候选的频数,从而确定 L_k ^[12]。

2.3 Apriori 算法的不足

Apriori 算法能够较有效地得到频繁项集,但也存在在一些不足之处。

(1)每当挖掘 K 频繁项集时,都要对事务数据库进行多次扫描以得到相应的支持度计数,不可避免地将导致时间消耗太长。

(2)将会生成很多的候选项集,在 L_{k-1} 通过自连接得到 L_k 时,可能会生成大量的候选项集,特别是 C_2 。

鉴于以上的缺陷,提出了基于矩阵的 Apriori 算法,其核心方法是用矩阵来表示事务数据库中的数据项集。

3 基于矩阵的 Apriori 算法

该算法是把 0-1 矩阵作为辅助元素,只要扫描一次事务数据库,就能得到所有符合条件的频繁项集。

算法过程为^[13](T 代表事务数据库):

(1)构建 $m \times n$ 阶 0-1 矩阵。其中, m 是事务的个数, n 是项集的个数。

- $a_{ij} = 1$, I 项出现在 T_j 中;
- $a_{ij} = 0$, I 项没有出现在 T_j 中。

(2)对 I_i 中 1 的个数记数。

(3)将记数小于 minsup 的 I_i 项(矩阵的行)都删除。

(4)对记数大于等于 minsup 的项做交运算。

(5)用上述的结果项构建矩阵。

- $a_{ij} = 1$, 结果项存在于 T_j 中;
- $a_{ij} = 0$, 结果项不存在于 T_j 中。

(6)对以结果项构建的矩阵中 1 的个数记数。

(7)将记数小于 minsup 的结果项都删除。

(8)返回(4),否则转到(9)。

(9)最终的结果项就是所要求的关联规则。

从上述过程可知,0-1 矩阵改进算法有两个优点:

(1)频繁项集的搜索工作可以仅通过一次扫描数据库完成,减少了访问数据库的次数。

(2)减少大量的候选集的产生,节约存储空间。

4 改进的基于矩阵 Apriori 算法

上述基于矩阵的算法需要产生多个矩阵,并且随着支持度的改变,矩阵需要不断更改。而经典的算法会生成很多的候选项集,主要集中在二项集上,并且需要多次扫描事务数据库。鉴于这些情况,文中采用新的方法来构建相关矩阵。

通过扫描一次事务数据库就可以得到一个上三角矩阵,通过此矩阵,不需要进行自连接便可直接得出符

合条件的频繁 1-项集和频繁 2-项集。然后再按照经典的 Apriori 算法由 L_2 自连接得到 C_3 , 再和 minsup 相比较, 得到 L_3 ……依次进行下去, 直至不存在满足条件的频繁项集。

4.1 矩阵的构成

为方便说明, 设存在下面的数据库, 如表 1 所示。

表 1 某分店的事务数据

TID	商品 ID 列表	TID	商品 ID 列表
T100	I_1, I_2, I_5	T600	I_2, I_3
T200	I_2, I_4	T700	I_1, I_3
T300	I_2, I_3	T800	I_1, I_2, I_3, I_5
T400	I_1, I_2, I_4	T900	I_1, I_2, I_3
T500	I_1, I_3		

上述数据库 D 中总共有 9 个事务, 根据表中的数据建立一个 0-1 矩阵。构建此矩阵时, 以项集 I 的项作为矩阵相对应的行标和列标。在此矩阵中, 某一元素 A_{ij} 的具体值表示相应的二项集 $\{I_i, I_j\}$ 在数据库中总共出现的次数。具体扫描过程为: 如果扫描到一个事务里存在着 $\{I_i, I_j\}$, 那么就在 0-1 矩阵中把和该二项集相对应的 A_{ij} 位置的元素数值加 1, 并且按顺序依次对单个数据项集进行统计, 将结果放在矩阵的主对角线相应位置。

按照上述方法可以得到如下所示的上三角矩阵:

$I_1 \quad I_2 \quad I_3 \quad I_4 \quad I_5$

$$\begin{matrix} I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{matrix} \begin{bmatrix} 6 & 4 & 4 & 1 & 2 \\ & 7 & 4 & 2 & 2 \\ & & 6 & 0 & 1 \\ & & & 2 & 0 \\ & & & & 2 \end{bmatrix}$$

4.2 频繁 1 项集和频繁 2 项集

设最小支持度计数为 2。

频繁 1-项集: 因为上三角矩阵的主对角线的元素代表项集 I 的项在事务数据里出现的次数, 通过与最小支持度相比较, 可以得出 $L_1 = \{I_1:6; I_2:7; I_3:6; I_4:2; I_5:2\}$ 。

频繁 2-项集: 矩阵的上三角元素代表项集 $\{I_i, I_j\}$ 在上述的数据库中总共出现的次数, 通过与 minsup 相比较, 可以得出

$$L_2 = \left\{ \begin{matrix} I_1, I_2:4 \\ I_1, I_3:4 \\ I_1, I_5:2 \\ I_2, I_3:4 \\ I_2, I_4:2 \\ I_2, I_5:2 \end{matrix} \right\}$$

4.3 频繁 3 项集及频繁 K 项集

根据经典的 Apriori 算法对 L_2 进行自连接得到 $C_3 = \{I_1, I_2, I_3; I_1, I_2, I_5\}$, 然后扫描数据库得到相应的频数, 根据 minsup, 筛选出 $L_3 = \{I_1, I_2, I_3:2; I_1, I_2, I_5:2\}$ 。

L_3 再自连接得到 $\{I_1, I_2, I_3, I_5\}$, 因为不符合算法的先验性质, 所以 $C_4 = \varnothing$, 因此算法终止。

4.4 算法描述

对于上述提出的改进的基于矩阵的 Apriori 算法的 (M-Apriori) 描述如下:

输入: 事务数据库 D , 最小支持度 min-sup;

输出: 频繁 1-项集, 频繁 2-项集……

Step1: 首先按照事务数据库构建上三角矩阵。

Step2: 按照上三角矩阵直接得出 L_1 和 L_2 。

Step3: 根据经典 Apriori 算法, 利用 L_2 找到 C_3 , 进而得到 L_3 。按照此方法一直迭代, 直到没有符合 minsup 的频繁项集, 从而找到所有的满足要求的频繁项集。

5 实验与分析

实验环境: 操作系统 Windows 7, CPU 为 3.20 GHz Intel(R) Core(TM) i5-3470, 编译软件为 Matalab2014a。

参数设置: minsup = 0.3。实验采用的数据是超市的购物清单。

实验将经典的 Apriori 算法和文中的 M-Apriori 算法进行比较, 可以发现 M-Apriori 算法比经典的 Apriori 算法使用 count_support 函数的次数减少了 2 次。而 count_support 函数的功能就是通过扫描数据库从大量的候选项集中筛选出符合 minsup 的频繁项集, 从而得到相应的关联规则。另外, M-Apriori 算法首先就需要通过扫描一次数据库生成目标矩阵, 即上三角矩阵。所以, 整体来说 M-Apriori 算法要比经典的 Apriori 算法少扫描一次数据库。此外, M-Apriori 算法减少了候选项集的数目。

针对中小型事务数据库来说, 只要得到目标矩阵就可以很直观地得出频繁 1-项集 L_1 和频繁 2-项集 L_2 。

6 结束语

针对 Apriori 算法需要多次扫描事务数据库这一缺点, 文中在矩阵的基础上改进了 Apriori 算法, 提出了 M-Apriori 算法。实验结果表明, 提出的新算法可以适当减少扫描事务数据库的总次数, 并且保证了频

性评价算法是一种行之有效的确定聚类数的途径。

4 结束语

文中提出一种特征加权的半监督聚类算法,并对该算法在不同模糊聚类有效性评价算法下的聚类结果进行分析。实验结果表明,综合不同的聚类有效性评价结果,能够有效得出输入数据集的最佳聚类数,从而解决大部分聚类算法中聚类数的确定问题,具有良好的应用前景。

参考文献:

- [1] 许海洋,汪国安,王万森. 模糊聚类分析在数据挖掘中的应用研究[J]. 计算机工程与应用,2005,41(17):177-179.
- [2] 高新波,谢维信. 模糊聚类理论发展及应用的研究进展[J]. 科学通报,1999,44(21):2241-2251.
- [3] 高新波. 模糊聚类分析及其应用[M]. 西安:西安电子科技大学出版社,2004:113-119.
- [4] Halkidi M, Batistakis Y, Vazirgiannis M. On clustering validation techniques[J]. Intelligent Information Systems, 2001, 17(2-3):107-145.
- [5] 张惟皎,刘春煌,李芳玉. 聚类质量的评价方法[J]. 计算机工程,2005,31(20):10-12.
- [6] 李 洁,高新波,焦李成. 一种基于修正划分模糊度的聚类有效性函数[J]. 系统工程与电子技术,2005,27(4):723-726.

(上接第 64 页)

繁项集结果的准确性。但由于在生成目标矩阵时,需要统计出所有的项和二次项集在事务数据库中出现的次数,从而增大了统计量,所以在时间方面没有得到优化。如何保证在减少扫描事务数据库次数的基础上提高算法的运行速度,是笔者下一步需要研究的工作。

参考文献:

- [1] Zaïane O R, El-Hajj M. Advances and issues in frequent pattern mining[C]//Proc of eighth Pacific-Asia conference on knowledge discovery and data mining. Sydney, Australia: [s. n.], 2004.
- [2] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[C]//Proceedings of the ACM SIGMOD conference on management of data. Washington, D C: ACM, 1993:207-216.
- [3] Agrawal R, Srikant R. Fast algorithms for mining association rules[C]//Bocca J B, Jarke M, Zaniolo C, et al. Proceeding of 20th international conference on very large data bases. Santiago, CA, USA: Morgan Kaufmann Publishers Inc, 1994:487-499.

- [7] Resson H, Wang D, Natarajan P. Adaptive double self-organizing maps for clustering gene expression profiles[J]. Neural Networks, 2003, 16(5-6):633-640.
- [8] Wu Sitao, Chow T W S. Self-organizing-map based clustering using a local clustering validity index[J]. Neural Processing Letters, 2003, 17(3):253-271.
- [9] Wu Sitao, Chow T W S. Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density[J]. Pattern Recognition, 2004, 37(2):175-188.
- [10] Li Longlong, Jonathan G, He Dongjian, et al. Semi-supervised fuzzy clustering with feature discrimination[J]. Plos One, 2015, 10(9):e0131160.
- [11] 李龙龙,王美丽. 基于加权二叉树的自适应遗传算法研究[J]. 计算机技术与发展, 2010, 20(11):95-99.
- [12] 李 洁,高新波,焦李成. 基于特征加权的模糊聚类新算法[J]. 电子学报, 2006, 34(1):89-92.
- [13] Bezdek J C. Pattern recognition with fuzzy objective function algorithms[M]. [s. l.]:Springer, 1983.
- [14] Bensaid A M, Hall L O, Bezdek J C, et al. Validity-guided (re)clustering with applications to image segmentation[J]. IEEE Transactions on Fuzzy Systems, 1996, 4(2):112-123.
- [15] Xie X L L, Beni G. A validity measure for fuzzy clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1991, 13(8):841-847.

- [4] 程玉胜,邓小光,江效尧. Apriori 算法中频繁项集挖掘挖掘实现研究[J]. 计算机技术与发展, 2006, 16(3):58-60.
- [5] 郭福亮,左凯伶. 关联规则挖掘中 Apriori 算法的一种改进[J]. 计算机与数字工程, 2007, 35(5):3-4.
- [6] Han Jiawei, Kamber M. Data mining: concepts and techniques [M]. 3th ed. Beijing: China Machine Press, 2012.
- [7] 李小兵,吴锦林,薛永生,等. 关联规则挖掘算法的改进与优化研究[J]. 厦门大学学报:自然科学版, 2005, 44(4):468-471.
- [8] 包震宇. 基于粗糙集对 Apriori 算法的改进[D]. 上海:上海师范大学, 2010.
- [9] 马晓辉. 一种基于关联规则 Apriori 算法的改进研究[J]. 现代计算机, 2011(6):6-8.
- [10] 吕桃霞,刘培玉. 一种基于矩阵的强关联规则生成算法[J]. 计算机应用研究, 2011, 28(4):1301-1303.
- [11] 张笑达,徐立臻. 一种改进的基于矩阵的频繁项集挖掘算法[J]. 计算机技术与发展, 2010, 20(4):93-96.
- [12] 芦 洁,刘志健. 挖掘关联规则中对 Apriori 算法的一个改进[J]. 微电子学与计算机, 2006, 23(2):10-12.
- [13] 顾 琳,黎敬涛,张兴涛. 对 Apriori 算法的一种改进—基于 0-1 矩阵处理算法[J]. 电脑知识与技术, 2007, 4(21):814-816.