不确定性数据频繁项集挖掘算法①

张常品, 刘广钟

(上海海事大学 信息工程学院, 上海 201306)

摘 要: 由于不确定性数据大量存在于传感器网络, 移动计算, 军事, 电信等应用领域, 传统的频繁项集挖掘算法难以适用到不确定性数据挖掘. 为了解决这个问题, 本文提出了一种快速有效的算法, 该算法基于可能世界模型, 只需要扫描一次数据库, 且没有建树的过程, 通过实验证明, 我们提出的算法比 UF_Growth 算法效率更高.

关键词: 频繁项集; 不确定性数据; 频繁模式; 关联规则; 可能世界模型

Algorithms of Frequent Item Sets Mining for Uncertain Data

ZHANG Chang-Pin, LIU Guang-Zhong

(Shanghai Maritime University College of Information Engineering, Shanghai 201306, China)

Abstract: Uncertain data exists in many situations, such as sensor networks, mobile computing, military, telecommunications and other applications, which makes it difficult to apply traditional algorithms to mining frequent item sets. To deal with these situations, we propose an efficient algorithm based on possible world model with single scan of database. The algorithm works well without any tree construction. Experimental results show that the efficiency of our algorithm is better than UF_Growth.

Key words: frequent item sets; uncertain data; frequent pattern; association rule; possible world model

关联规则挖掘[1]即从给定的数据库中找出数据项之间有价值的相互关联的知识, 关联规则分两个步骤, 首先找出频繁项集, 然后根据频繁项集找出用户所需要的关联规则.

在传统的事务数据库中, 数据存在与否及其值都是已知的, 然而随着移动网络、军事、传感器、Web 等领域的发展, 不确定性数据普遍存在现实生活中, 其原因可能是原始数据不准确、使用粗粒度的数据集合来装细粒度的数据、采集的数据使用估计值、为了保护隐私需要人为的对数据添加噪音等等, 如何在不确定性数据库中挖掘出有价值的信息, 成为学术界关注的热点.

不确定性数据主要有不确定性数据管理和不确定性数据挖掘两个方面的研究领域. 由于传统的数据记录方式已经不适应不
确定性数据, 为了更好的描述不确定性数据, 我们在数据属性列增加一个概率字段, 这一举措为不确定数据管理带来了广泛的发展.

不确定性数据有存在不确定性和属性不确定性两种存在方式[2]. 存在不确定性是指我们所记录的数据都是以一定的概率形式存在, 并且数据之间可能会相互影响. 属性不确定性是指在数据库中所有的记录都是已知的, 但是属性值以一定的概率形式存在, 我们通常用 PDF(概率密度函数)来表示记录的概率值.

1 不确定性数据频繁模式挖掘

在过去的几十年中, 传统的频繁模式挖掘算法已经相当成熟, 其中最为经典的算法是基于先验的 Apriori 算法和基于树结构的 FP_Growth 算法, 大多数算法都是在这两种算法的基础上提出来的. 随着对不确定性数据研究的深入, 研究者们对不确定性数据频繁项集的相关概念提出了新的定义. 在传统的数据库中, 如果 X 是一个频繁项集, 那么它的支持度 $\sup(x)$ 必须大于用户给定的最小支持度 $\min\_\sup$ 即:

① 收稿时间:2014-03-13;收到修改稿时间:2014-04-14

$\sup(x) > \min\_\sup$. 在不确定性数据库中, 由于项集以概率形式存在, 所以我们用期望支持度 $e\sup(x)$ 代替项集 x 在不确定性数据库中的支持度.

定义 1. 不确定项目和项集, 与传统的数据库不同的是, 不确定性数据库中项集多了一个概率属性, 任一项目 X 都以一定的概率出现, 用 (x_i, p_i) 表示, x_i 表示项目值, p_i 表示对应的概率值.

$I = \{(x_1, p_1), (x_2, p_2), ...(x_k, p_k)\}$ 有 K 个项目组成的集合, 长度为 K 的项集称为 K—项集.

定义 2. 事务与事务数据库, 假设 $I = \{(x_1, p_1), (x_2, p_2), ...(x_k, p_k)\}$ 是由数据库中所有的不确定项目构成的集合, 不确定性数据库 $T = \{T_1, T_2, ...T_n\}$ 表示一次处理中所含不确定性项目的集合, 其中每一个事务 T_i 都是 I 的子集.

定义 3. 期望支持度, 项集在各事务中出现的概率和.

定义 4. 频繁项集, 在不确定性数据集中, 如果项集 X 的期望支持度 $e\sup(x)$ 大于用户指定的最小支持度阈值 $\min\_\sup$ 即: $e\sup(x) > \min\_\sup$, 则 x 为频繁项集, 否则为非频繁项集.

定理 1. 如果项集 I 以期望支持度计算是频繁项集, 那么 I 的所有子集以期望支持度计算也是频繁项集.

2 可能世界模型

目前不确定性数据频繁项集挖掘模型有 3 种, 分别为基于期望支持度和概率频繁模式挖掘技术的可能世界模型、基于估计支持度频繁项集挖掘技术的概率模型、基于信任度的频繁项集挖掘技术的 DS 理论模型. 其他两种模型都可以转换为可能世界模型, 因此, 可能世界模型是不确定性数据频繁模式挖掘的通用模型[3,4]. 在可能世界模型中, 各元组之间任一合法组合都可以构成一个可能世界实例, 实例的概率值可以根据各元组的概率计算得到, 根据可能世界模型的特性, 可以将不确定性数据库分解成许多确定的数据库实例(可能世界实例).

与传统数据库最为明显的区别就是, 不确定性数据库中事务有不确定项集 Y 及其存在概率 $p(Y) \in (0;1)$ 两部分组成, 其中 $(Y \in T)$. 一般来说我们都假设事务之间是相互独立的, 因此可能世界下模型的概率[5-8]如下表示:

$$p(w) = \prod_{t \in I} (\prod_{x \in t} p(x \in t) * \prod_{x \notin t} (1 - p(x \in t))) \quad (1)$$

证明: 不确定性事务数据库 I 中, 事务 t_i 中的每一个项目, 存在两种可能世界 w_1 和 w_2, w_1 为 x_j 在 t_i 中存在的概率($x_j \in t_i$) $p(w) = p_{t_i}(x_j)$, w_2 为 x_j 在 t_i 中不存在的概率($x_j \notin t_i$), $p(w) = 1 - p_{t_i}(x_j)$, 假设 t_i 中包含 x 和 y 两个项目, 则出现 4 中可能世界, 那么在某个世界中出现的概率为 x 和 y 在该可能世界中存在情况为真的概率积, w 为某一可能世界, 则 $p(w) = p_{t_i}(x) * p_{t_i}(y)$, 因为项目之间是相互独立的,所以事务 t_i 中所有项目某一可能世界模型概率 $p(w) = p_{t_i}(x_j) * \prod_{x \notin t_i} (1 - p_{t_i}(x_j))$, 事务 t 中所有项目某一个可能世界的概率 $p(w) = \prod_{x \in t} p(x \in t) * \prod_{x \notin t} (1 - p(x \in t))$, 则事务数据库 I 中某一个可能世界的概率 $p(w) = \prod_{t \in I} (\prod_{x \in t} p(x \in t) * \prod_{x \notin t} (1 - p(x \in t)))$.

假设 $p(w_i)$ 是可能性世界 w_i 的存在概率, $s(x, w_i)$ 是项集 x 在可能世界 w_i 的支持度计数, 那么项集 x 的期望支持度:

$$e\sup(x) = \sum_{i=1}^{|w|} p(w_i) * s(x, w_i) \quad (2)$$

若数据库有 n 项, 那么构成 2^n 个可能世界实例[5][8], 实例的规模是传统数据库的指数倍, 求出 2^n 可能世界的期望支持度这是不可能的, 幸好研究者证明 Chun-Kit[10]等人了以下求期望支持度的公式.

$$e\sup(X) = \sum_{i=1}^{n} (\prod_{x \in X} p(x, t_i)) \quad (3)$$

其中 $p(x, t_i)$ 是项集 x 在事务 t_i 中出现的概率值.

3 不确定性数据频繁模式挖掘算法

不确定性数据频繁模式挖掘算法有基于期望支持度和基于概率频繁模式两种类型. 期望支持度是指项集在各事务中出现的概率和, 当期望支持度大于给定的支持度阈值即为频繁项集. 概率频繁模式指在给定的不确定数据库中, 给定最小支持度和用户定义的频繁概率阈值, 如果项集 X 发生在不少于最小支持度事物中的概率大于用户定义的概率,则称 X 为概率频繁模式.

为了在不确定性数据库中挖掘频繁项集, Chui 等人对 Apriori 算法进行了改进, 提出了 U_Apriori 算法[6]. 与 Apriori 算法类似, U_Apriori 算法也是基于先验性算法, 首先产生候选项集, 然后到数据库中检验是否为

频繁项集, 采用自底向上的模式还搜索频繁项集, 因此有着 Apriori 算法的缺点. 首先, 要得到频繁 1-项集 C_1, 将 C_1 中符合最小支持度计数的项集添加到频繁 1-项集链表 L_1 中, 然后 L_1 自连接得到频繁 2-项集 C_2, 然后在 C_2 中查找符合最小支持度计数的频繁项添加的频繁 2-项集 L_2 链表中, 以此类推, 直到 C_k 为空即没有新的频繁项集产生为止, 该算法时间开销主要用于计算频繁项集和候选项集上. U_Apriori 在频繁模式挖掘过程中需要多次扫描数据库, 并产生大量的候选项集, 因为期望支持度是有概率乘积累加得到的, 所以当候选项集概率比较小时, 会增加很多毫无意义的计算, 并增加了执行时间和系统消耗.

由于 FP_Growth 算法在传统的交易数据库中效率较高, Leung 等人在 FP-Growth 算法[11]的基础上提出一种适用于不确定数据库中挖掘频繁项集的算法 UF-growth[7][12]. UF_Growth 算法相对 U_Apriori 算法性能上进行了改进, 不会产生候选项集, 但是挖掘频繁项集需要两步: 1 建立 UF_Tree 树, 2 挖掘频繁项集. 与传统的 FP_Growth 算法中不同的是, 建树的过程中, 每一个节点需要包含 3 部分的内容: 项目名称、该项的预计支持度和与该项相同的预计支持度出现的次数. 建树具体操作步骤如下: 首先, 第一次扫描数据库, 得到各项的支持度来得到预计支持度, 然后按照预计支持度的大小进行降序排列, 将预计支持度小于用户的阈值的项目即非频繁项删掉. 然后第二次扫描数据库, 建立 UF_tree, 与 FP_tree 不同的是, 只有当事务中项名和预计支持度与 UF_tree 中某个节点的项名及预计支持度完全相同时, 该事务才能与分支合并, 否则另建立分支.

4　改进的算法

我们提出一种基于可能世界模型快速有效的算法, 以期望支持度来进行不确定性数据频繁模式挖掘.

算法描述:

1) 扫描交易数据库, 创建期望支持度表(SCT)和位图表(BFT)

2) 在 SCT 表中找到全连接的节点组合(满足最小支持度计数)

3) 对于所有全连接的节点组合执行以下操作

①在 BFT 表中执行逻辑与操作;

②计算所有逻辑与操作的节点值的总和;

③如果所有执行逻辑与的总和大于给定的最小支持度计数, 将该节点组合添加到频繁项集 L_k 中.

表 1　事务数据库

Trans.ID	Items			
1	a:0.9	b:0.7	c:0.8	d:0.6
2	a:0.9		c:0.7	d:0.9
3		b:0.9	c:0.8	d: 0.8
4	a:0.8	b:0.9	c:0.9	
5	a:0.8		c:0.7	d:0.9

根据我们所提出的算法, 使用数组来创建 SCT 表和 BFT 表, 初始化 SCT 表, 使其各项值为 0, 当扫描第一条交易记录'a:0.9 b:0.7 c:0.8 d:0.6', SCT 表中 ab、ac、ad、bc、bd、cd 中的值会分别加上各自的概率积. 扫描第一条记录后 SCT 表和 BFT 表分别如表 2、表 3 所示:

表 2　SCT

Items	a	b	c	d
a		0.63	0.72	0.54
b			0.56	0.42
c				0.48

表 3　BFT

a	b	c	d
0.9	0.7	0.8	6

扫描第二条交易记录'a: 0.9　c: 0.7　d: 0.9'SCT 表和 BFT 表如表 4、表 5 所示:

表 4　SCT

Items	a	b	c	d
a		0.63	1.35	1.35
b			0.56	0.42
c				1.11

表 5　BFT

a	b	c	d
0.9	0.7	0.8	0.6
0.9		0.7	0.9

扫描最后一条交易记录'a: 0.8　c: 0.7　d: 0.9'后 SCT 表和 BFT 表如表 6、表 7 所示:

表 6　SCT

Items	a	b	c	d
a		1.35	2.63	2.07
b			2.09	1.14
c				2.38

表 7 BFT

a	b	c	d
0.9	0.7	0.8	0.6
0.9		0.7	0.9
	0.9	0.8	0.8
0.8	0.9	0.9	
0.8		0.7	0.9

假设期望支持度为 1.2, 满足最小期望支持度计数的 SCT 表如表 8 所示:

表 8 SCT

Items	a	b	c	d
a		1.35	2.63	2.07
b			2.09	0
c				2.38

根据表 8 找出全连接的项集. 对于 a 行来说, bcd 均满足最小期望支持度, 下面检查 abcd 是否是全连接. 对于 a 行, bcd 没有为 0 的值, b 行 d 值为 0, 所以 abcd 不是全连接的. 然后分别检查 abc 和 acd 是否为全连接的. 经过检查 abc 和 acd 均为全连接的.

在 BFT 表中计算 abc 和 acd 的期望支持度分别为 1.152、1.503, acd 满足最小期望支持度计数, 将 acd 添加到 L_k 频繁项集中去. 我们可以从 SCT 表中直接得出频繁 2 项集. 同理, 在 SCT 中找出全连接的项集, 然后在 BFT 表中根据逻辑与来求得期望支持度, 将满足最小期望支持度计数的项集添加到相应的频繁项集中去. 算法对比如表 9 所示:

表 9 算法对比

算法	多次扫描数据库	实时数据库
U_Apriori	√	×
UF_Growth	×	×
Proposed	×	√

1) 若在一个含有 100 个项目的不确定性数据库中寻找频繁模式, 就需要产生 $2^{100}=10^{30}$ 个候选项集.

2) 需要多次扫描数据库(如果模式的长度为 n, 那么就需要扫描 n+1 次).

3) 虽然 UF_Growth 算法减少了扫描数据库的次数, 但是该算法适用于数据实时更新的数据库, 因为当数据库中有新的交易添加进来的时候, 就要重新构建 UF_tree 树.

4) 我们所提出的算法可以轻松解决以上提出的问题, 只需扫描一次数据库并无需建树, 就可以简单快速的得到频繁项集.

5 实验结果

用不同大小的数据集来测试我们提出的算法和 UF__Growth 算法运行的时间, 如表 10 所示:

表 10 算法运行时间

数据集 ID	数据集大小 (KB)	运行时间(S)	
		Proposed	UF_Growth
1	7.66	2.154	2.185
2	26.4	2.190	2.194
3	33.7	2.159	2.493
4	44.8	2.331	2.827
5	90.7	2.191	2.603
6	92.9	2.159	2.416

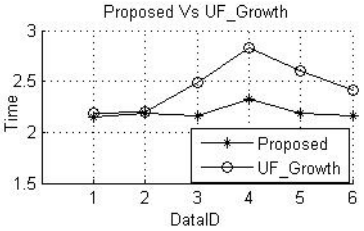

图 1 Proposed Vs UF_Growth

6 总结

由于不确定性数据广泛存在我们的生活中, 目前大多数算法都是应用于传统数据库的, 如何在不确定性数据库中挖掘有价值的信息, 就需要提出一些高效的算法来完成不确定性数据频繁项集的挖掘. 而现有的一些算法都是基于 Apriori 算法和 FP_Growth 算法提出来的, 这些算法的效率并不是很高, 本文中我们提出了一个简单快速的算法来进行不确定性数据频繁项集挖掘, 不仅降低了扫描交易数据库的次数, 而且不是基于树的算法, 可以适用于实时数据库, 通过实验证明我们提出的算法效率更高.

参考文献

1 张友生,徐峰.系统分析师技术指南.北京:清华大学出版社, 2004.

2 Aggarwal CC, Yu PS. A survey of uncertain data algorithms and applications. IBM Research Report, RC24394, 2007.

3 Abiteboul S, Kanellakis P, Grahne G. On the representation and querying of sets of possible worlds. ACM SIGMOD Record, 1987, 16(3): 34–48.

4 Jayram TS, Kale S, Vee E. Efficient aggregation algorithms

for probabilistic data. Proc. of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms. New Orleans, 2007. 346–355.

5 Leung C, Mateo M, Brajczuk D. A tree-based approach for frequent pattern mining from uncertain data. LNAI 5012. PAKDD. 2008. 653–661.

6 Chui CK, Kao B, Hung E. Mining frequent item sets from uncertain data. In: Desai BC, Dacca D, Greco S, eds. Proc. PAKDD. New York. ACM. 2007. 47–58

7 Leung CKS, Carmichael CL, Hao B. Efficient mining of frequent patterns from uncertain data. Proc IEEE ICDM Workshops. 2007. 489–494.

8 Chui CK, Kao B. A decremental approach for mining frequent item sets from uncertain data. LNAI 5012: PAKDD, 2008: 64–75.

9 Leung C, Carmichael C, Hao B. Efficient mining of frequent patterns from uncertain data. Proc IEEE ICDM Work-shops. 2007. 489–494.

10 Aggarwal CC. On Unifying privacy and uncertain data models. 24th International Conference on Data Engineering, Cancun, Mexico, IEEE. April 2008. 386–395.

11 Leung CKS, Carmichael CL, Hao BY. Efficient mining of frequent patterns from uncertain data. Pro. ICDMW. New York. ACM. 2007. 489–494.

12 Leung CKS, Brajczuk AD. Efficient mining of frequent item sets from data streams. LNCS 5071: BNCOD. 2008. 2–14.