# 高维数据的频繁封闭模式挖掘算法研究综述①

杨风召

(南京财经大学 电子商务系, 南京 210003)
(江苏省电子商务重点实验室, 南京 210003)

**摘 要：**挖掘频繁模式是数据挖掘领域一个重要且基础的问题。频繁封闭项集挖掘可以提供完全的无冗余的频繁模式。随着生物信息学的兴起，产生了一类具有较多列数的特殊数据集，这种高维数据集对以前的频繁封闭模式挖掘算法提出了新的挑战。对高维数据的频繁封闭模式挖掘算法进行了综述，按照算法的特性对这些算法进行了分类，比较了基于行计数的两类挖掘算法，并对能根据数据子集的特性进行列计数和行计数自动转换的混合计数算法进行了讨论,最后指出了该领域的研究方向。

**关键词：**频繁封闭模式；高维数据；数据挖掘；综述

## Mining Frequent Closed Patterns for Very High Dimensional Data: A Review

YANG Feng-Zhao

(E-Business Department, Nanjing University of Finance & Economics, Nanjing 210003, China)
(Jiangsu Key Laboratory of E-Business, Nanjing 210003, China)

**Abstract:** Mining frequent patterns is a fundamental and essential problem in many data mining applications. Mining frequent closed itemsets provides complete and non-redundant results for frequent pattern analysis. The growth of bioinformatics has resulted in datasets with new characteristics. These datasets typically contain a large number of columns. Such high-dimendional datasets pose a great challenge for existing closed frequent pattern discovery algorithms. This paper presents a survey of the various algorithms for mining frequent closed itemsets in very high dimensional data along with a hierarchy organizing the algorithms by their characteristics. We compare two row enumeration-based algorithms, discuss an algorithm which is designed to automatically switch between feature enumeration and row enumeration during the mining process based on the characteristics of the data subset being considered, and finally point out the research direction in this field.

**Key words:** frequent closed pattern; high dimensional data; data mining; survey

## 1 引言

关联规则挖掘的目的在于确定超大型数据库中项集之间的关系。而关联规则的挖掘问题可以简化为确定频繁项集和它们的支持度问题，我们可以称之为频繁模式的挖掘。对频繁模式挖掘的研究已经有很多，比较有名的算法包括 Apriori、FP-growth、H-mine 等。在进行频繁模式挖掘时，最后得到的频繁模式往往有较大的冗余，在文献[1]中作者又提出了频繁封闭模式的概念。利用频繁封闭模式，我们可以在不丢失信息的情况下减少产生的频繁模式的数量，从而也使得产生的关联规则数量得到缩减。

最初的频繁封闭模式主要是针对交易数据，针对这种数据提出的挖掘算法都是基于特征计数的，它们将特征组合作为搜索空间。这种算法可以表示为特征计数树的搜索问题。采用这种算法对行平均长度较短的数据集进行挖掘效果较好，因为如果设 i 是行的最大尺寸，那么可能的频繁项集有 $2^i$，这种数据集一般 i<100。

① 基金项目:国家自然科学基金(71072172);留学人员科技活动择优资助项目(YFZ302002);江苏高校优势学科建设工程资助项目
收稿时间:2011-03-10;收到修改稿时间:2011-04-19

随着生物信息学的发展，微阵列技术产生了许多基因表达数据集，也就是微阵列数据。与交易数据不同，微阵列数据通常行数较少（通常为几百），但特征数非常多（可能几万或几十万）。在这种数据集上，频繁模式挖掘算法可用来发现有共同调控关系的基因或基因组，基于频繁模式的关联规则可以用来构建基因网络。但对这种数据集进行频繁封闭模式挖掘时，搜索特征组合空间变得几乎不可能，因此需要专门研究高维数据频繁封闭模式的挖掘算法。本文对高维数据频繁封闭模式的挖掘算法进行了总结，比较了主要算法的原理和特点，最后指出了该领域的研究方向。

## 2　频繁封闭挖掘算法的分类

### 2.1 基本定义

设 $F=\{f_1, f_2, \ldots, f_m\}$是一个特征（feathers）的集合。数据集 D 由一个行（rows）的集合 $R=\{r_1,r_2,\ldots,r_n\}$组成，这里每个行 $r_i$ 是一个特征的集合，也就是 $r_i \subseteq F$。例如，表 1 中的数据集中有 5 个特征和 5 个行，特征集 F={a, b, c, d, e}，行集 $R=\{r_1, r_2, r_3, r_4, r_5\}$。给定一个特征集 $F' \subseteq F$，我们用 R(F')表示包含 F'的行的最大集合，R(F')称为特征支持集。给定一个行集 $R' \subseteq R$，我们用 F(R')表示 R'中所有行共有的最大特征集，称为行支持集。给定一个特征集 F'，数据集中包含 F'的行的个数，称为 F'的支持度。根据前面的定义，我们可以将 F'的支持度表示为|R(F')|。对于一个特征集 $F' \subseteq F$，当且仅当 F(R(F'))=F'，称 F'为封闭特征集。对于一个特征集 $R' \subseteq R$，当且仅当 R(F(R'))=R'，称 R'为封闭行集。一个封闭特征集可以称为一个封闭模式。如果一个特征集 $F' \subseteq F$ 满足以下两个条件：（1）F'的支持度|R(F')|高于一个给定的最小支持度阈值；（2）F'是一个封闭模式。则称 F'为一个频繁封闭模式。

### 2.2 频繁封闭模式挖掘算法分类

频繁封闭模式的挖掘算法可以划分为三种类型：基于特征计数的算法、基于行计数的算法和混合计数算法。基于特征计数的算法有 A-close[1]、CLOSET[2]、MAFIA[3]、CHARM[4]和 Closet+[5]等。这些算法分别采用宽度优先搜索（BFS）和深度优先搜索(DFS)算法对搜索特征计数树从根部进行搜索，保证所有的特征组合都访问到。基于行计数的算法有 Carpenter[6]、FARMER[7]、TOPKERS[8]、TD-Close[9]等。混合计数算法有 COBBLER[10]等。表 1 列出了各种算法的特征，包括计数方式和搜索策略。

由于特征计数方法用于高维数据频繁封闭模式挖掘时，效率较低，因此下面我们主要讨论行计数方法和混合计数方法的特点。

表 1　频繁封闭模式挖掘算法分类

| 计数方式 | 搜索策略 | 算法 |
|---|---|---|
| 特征计数 | 宽度优先 | Close　A-Close |
| | 深度优先 | CLOSET　Closet+　MAFIA |
| | 混合策略 | Charm |
| 行计数 | 自底向上深度优先 | Carpenter　FARMER　TOPKERS |
| | 自顶向下深度优先 | TD-Close |
| 混合计数 | 深度优先 | COBBLER |

## 3　基于行计数的频繁封闭模式挖掘算法

与基于特征计数的算法类似，基于行计数的算法也可以表示为行计数树的搜索问题。为了构建行计数树，我们要将表 2 中的数据表 T 转换为变换表 TT，如表 3 所示。为清楚期间，以后我们将表 T 中的记录称为行，表 TT 中的记录称为元组。

表 2　数据表 T

| i | $\mathcal{F}(r_i)$ |
|---|---|
| 1 | a,c,d |
| 2 | a,b,d,e |
| 3 | b,e |
| 4 | b,c,d,e |
| 5 | a,b,c,e |

表 3　变换表 TT

| $f_j$ | $\mathcal{R}(f_j)$ |
|---|---|
| a | 1,2,5 |
| b | 2,3,4,5 |
| c | 1,4,5 |
| d | 1,2,4 |
| e | 2,3,4,5 |

图 1 所示为一棵为表 TT 构建的行计数树，节点 12 代表行组合{1,2}，下面的括号{ad}代表特征 ad 同时被行 1 和 2 所包含。也就是 F(12)=ad。

### 3.1 自底向上深度优先搜索算法

Carpenter 算法[6]采用深度优先搜索（DFS）算法对行计数树从根部进行搜索，这样所有的行组合都会按照字典序依次被访问到。如果不考虑任何优化和修剪策略，在图 1 中，节点被访问的次序为{1, 12, 123, …, 45, 5}。

容易证明，每个封闭模式对应一个唯一的特征支持集，也就是说不存在两个不同的封闭模式，它们对

应的特征支持集是相同的。因此通过对行计数树中所有行的组合进行计数，可以保证找到所有的封闭模式。但是对行计数树进行完全搜索的算法效率较低，必须采用修剪技术对没有必要的分支进行修剪。

FARMER[7]和 TOPKERS[8]用来产生关联规则，在通过变换表进行频繁模式挖掘时同时考虑了类标信息，并且引入了有趣规则群的概念。在进行频繁模式挖掘时，他们同 Carpenter 一样采用了行计数树，进行深度优先搜索。TOPKERS 与 FARMER 的不同之处在于使用偏好选择对有趣规则群进行了过滤，在实现时采用了前缀树使算法更有效。

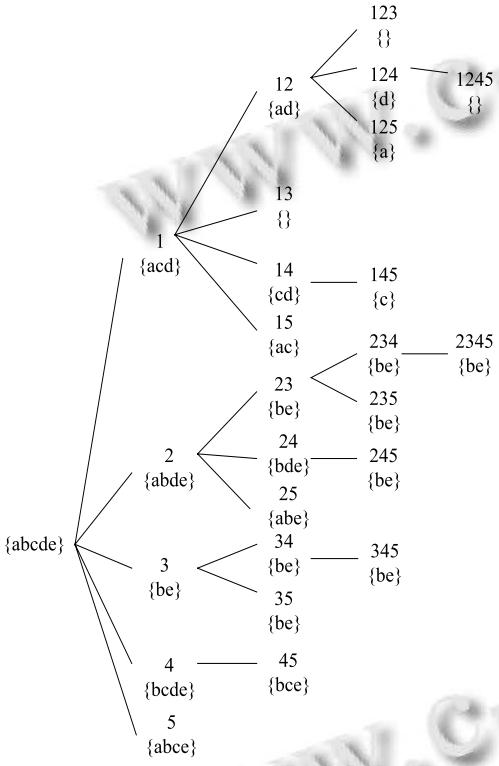


图 1 行计数树

### 3.2 自顶向下深度优先搜索算法

在上面的 Carpenter 算法中，搜索行集按照行组合尺寸由低到高顺序，这样行计数树的搜索空间不能得到较早的修剪。如当支持度为 3 时，显然计数树的前两层不能满足条件，但在 Carpenter 算法中仍然需要对这些结点进行检查，特别当支持度的阈值较高时，这种算法需要搜索的无效结点会非常多。采用自顶向下深度优先搜索算法可以克服上述算法的缺点。图 2 是一颗自顶向下的行计数树，每个结点代表一个行集，我们定义根结点的层次为 0，一个 n 行数据集的最高层次为 n-1。从图中可以看出，如果给定支持度阈值为 minsup，那么没必要搜索那些层次高于 n-minsup 的所有结点，因为这些结点中的行集的尺寸小于 minsup，其所对应的封闭模式也不可能是频繁的。例如当支持度的阈值为 3 时，对表 1 中的数据表，当搜索到层次 2 时就不再继续往下搜索了。

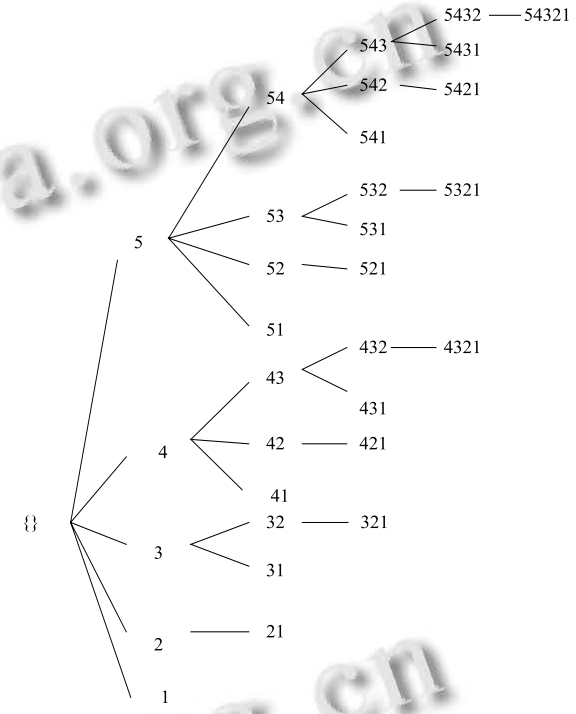


图 2 自顶向下行计数树

在自顶向下的计数树中，每个结点对应一个子表，如根结点代表整个表 TT。树中的其它结点是去除某些行形成的，我们称这些子表为 x-排除变换表。表对应的父结点和子结点，我们分别称之为父表和子表。

图 3 给出了与图 2 的行计数树对应的排除行计数树，在这颗树中显示了排除行之间的父子关系。

从支持度的定义和 x-排除表的计算过程可以看出，由于最小支持度阈值的作用，排除表的尺寸会越来越小，因而搜索空间也会快速收缩。

根据以上思想，采用自顶向下的搜索策略，Hongyan Liu 等设计出一种算法，称为 TD-Close[9]，从表 T 中挖掘所有的频繁封闭模式。为了避免在挖掘的过程中产生所有的频繁特征集，必须尽可能早地进行封闭性检查。

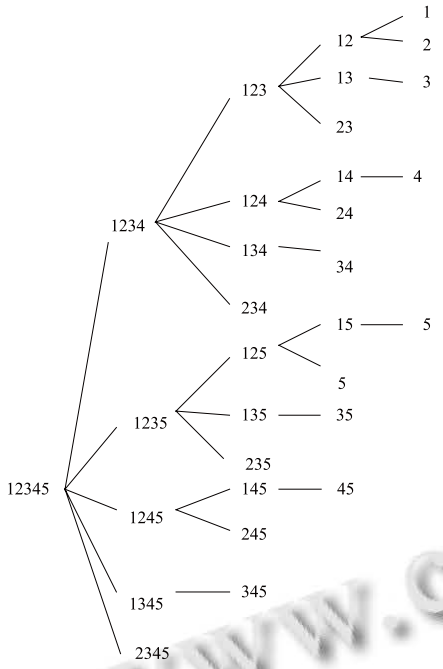


图 3　排除行计数树

## 4　基于混合计数频繁封闭模式挖掘算法

从以上讨论可知，在进行频繁封闭模式挖掘时，特征计数算法对特征数较少的大数据集比较有效，因为在该算法中需要计数的特征组合比较少；而行计数算法则适合于特征数较多的小数据集。但这两种算法都不适合处理特征数较多的大数据集。基于行计数和特征计数的频繁封闭模式挖掘算法在挖掘过程中根据要处理的数据子集的特征自动在特征计数和行计数之间进行切换，每个数据子集都可以用最合适的方法进行处理，实验表明这种方法在处理特征数较多的大数据集时比单纯采用特征计数或行计数都更为有效。

### 4.1　动态计数树

基于行计数或者特征计数的算法中的计数树都是静态的，也就是说，在整个算法中，它们都只使用一种计数方法。而在基于混合计数的算法中采用的计数树是动态的，也就是说，它在执行过程中会根据所处理的表的情况选择不同的算法，比如当特征数小于行数时采用特征计数算法，而特征数大于行数时采用行计数算法。这主要是因为在搜索计数树时产生的条件表的特性与原始的表可能是不同的，原始表适合特征计数算法，而某些条件表可能适合行计数算法。

图 4 是动态计数树的例子。在图 4 中先进行特征计数，然后从特征计数转换成行计数。同理也可以先进行行计数，然后从行计数转换成特征计数。

我们将动态计数树中的所有结点划分成两种类型：行计数结点和特征计数结点。顾名思义，行计数结点是对行子集 R’进行计数，而特征计数结点是对特征子集 F’进行计数。如在图 4 中结点“bc”是特征计数结点，而它的孩子结点“4”却是一个行计数结点。

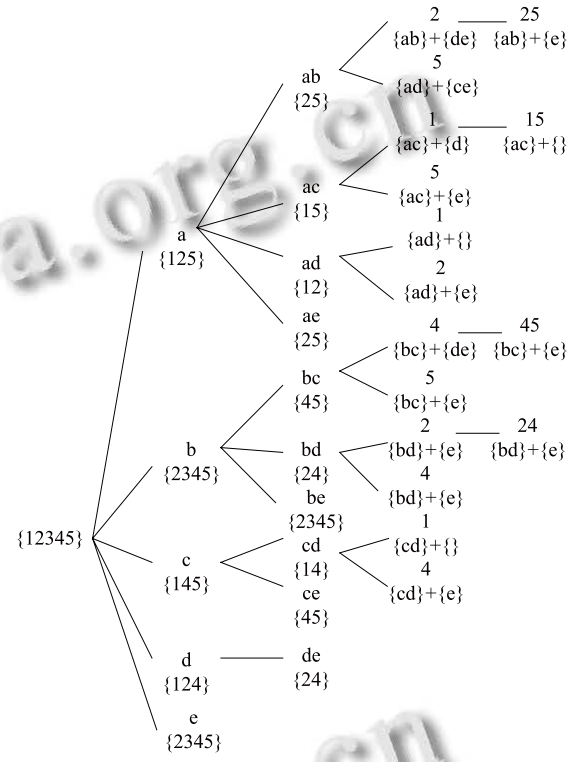


图 4　由特征计数转换成行计数的动态计数树

### 4.2　算法

算法 COBBLER[10]的输入是原始表 T，变换表 TT，特征集 F，行集 R 和支持度 minsup，输出是频繁封闭模式集 FCP。该算法递归计算条件表和条件变换表对动态计数树进行深度优先遍历。每个条件表代表一个特征计数结点，而每个条件变换表代表一个行计数结点。算法首先对频繁封闭模式集合 FCP 初始化，把 FCP 置为空。然后检查转换条件来决定进行行计数还是进行特征计数。根据转换条件调用行计数过程或者特征计数过程。

### 4.3　转换条件

在上述算法中，需要检查转换条件，来决定是否需要从行计数转换到特征计数，或者从特征计数转换到行计数。为了确定转换条件，引入计数代价的概念，在选择计数方法时优先选取特征计数子树和行计数子

树中的计数代价低者。

树的计数代价的估计可以从两个方面考虑：树的大小和在树的每个结点上的计算代价。树的大小可以用树中结点的数量来衡量，而一个结点的计算代价可以用在这个结点上需要处理的行或者特征数目来衡量。

## 5 小结

通过对上述高维频繁模式挖掘算法的比较，我们可以得到如下结论：

① 对行数较少而特征数较多的高维数据集进行频繁模式挖掘时采用行计数方法能有效缩小搜索空间，提高挖掘的效率；

② 对行数和特征数都较多的高维数据集进行频繁模式挖掘时，单纯采用行计数和特征计数的算法效率都较低，应该采用混合计数算法，根据所处理表的特征，动态选择不同的算法，达到提高算法效率的目的。

③ 在行计数算法中，由于支持度阈值的存在，自底向上算法需要搜索较多的无效结点，而自顶向下算法能有效克服这个缺点，快速收缩搜索空间。

尽管迄今为止已经提出了很多频繁封闭模式的挖掘算法，对高维数据的频繁模式挖掘仍然是一个困难的问题。这种困难一方面表现在算法的效率上，对高维大数据集的挖掘仍然效率不高，另一方面表现在应用上，会产生大量的无趣模式。未来的研究方向应该以行计数方法和特征计数方法为基础，将频繁封闭模式的挖掘与领域知识相结合，在挖掘过程中尽早采用领域知识收缩搜索空间，提高挖掘的效率，同时避免产生大量的无趣模式。

## 参考文献


1 Pasquier N, Bastide Y, Taouil R, Lakhal L. Discoverying frequent closed itemsets for association rules. In: Beeri C, Buneman P, eds. Proc. of the 7th International Conference on Database Theory, LNCS 1540. Heidelberg: Springer Berlin, 1999: 398–416.

2 Pei J, Han J, Mao R. CLOSET: An eficient algorithm for mining frequent closed itemsets. In: Chen W, Naughton JF, Bernstein PA, eds. Proc. 2000 ACM-SIGMOD International Workshop Data Mining and Knowledge Discovery. New York: ACM Press,2000:21–30.

3 Burdick D, Calimlim M, Gehrke J. MAFIA: A maximal frequent itemset algorithm for transactional databases. In: Georgakopoulos D, Buchmann A, eds. Proc. of the 17th International Conference on Data Engineering. Heidelberg: IEEE Computer Society, 2001:443–452.

4 Zaki M, Hsiao C. Charm: An efficient algorithm for closed association rule mining. In: Grossman RL, Han J, Kumar V, Mannila H, Motwani R, eds. Proc. of 2002 SIAM International Conference Data Mining. Arlington, VA, 2002: 457–473.

5 Wang J, Han J, Pei J. Closet+:Searching for the best stetegies for mining frequent closed itemsets. In: Getoor L, Senator TE, Domingos P, Faloutsos C, eds. Proc. of 2003 ACM SIGKDD International Conference on Kowledge Discovery and Data Mining. New York: ACM Press, 2003: 236–245.

6 Pan F, Cong G, Tung AK. Carpenter: Finding closed patterns in long biological datasets. In: Getoor L, Senator TE, Domingos P, Faloutsos C, eds. Proc. of 2003 ACM SIGKDD International Conference on Kowledge Discovery and Data Mining. New York: ACM Press, 2003: 637–642.

7 Cong G, Tung AK, Xu X, et al. FARMER: Finding Interesting rule groups in microarray datasets. In: Weikum G, ed. Proc. of the ACM SIGMOD International Conference on Management of Data 2004. New York: ACM Press, 2004: 143–154.

8 Cong G, Tan K, Tung AK, et al. Mining top-k covering rule groups for gene expression data. In: Ozcan F, ed. Proc. of the ACM SIGMOD International Conference on Management of Data 2005. New York: ACM Press, 2005: 670–681.

9 Liu H, Han J, Xin D, Shao Z. Mining frequent. patterns from very high dimensional data: A. top-down row enumeration approach. In: Ghosh J, Lambert D, Skillicorn DB, Srivastava J, eds. Proc. of the Sixth SIAM International Conference on Data. Mining. Bethesda: SIAM, 2006: 20–22.

10 Pan F, Tung AK, Cao G, Xu X. COBBLER: Combining column and row enumeration for closed pattern discovery. In: Hatzopoulos M, Manolopoulos Y, eds. Proc. of 2004 International Conference on Scientific and Statistical Database Management. Washington: IEEE Computer Society, 2004: 21–30.