

Data Mining

Chapter 5

Association Analysis: Basic Concepts

Introduction to Data Mining, 2nd Edition
by
Tan, Steinbach, Karpatne, Kumar

02/14/2018

Introduction to Data Mining, 2nd Edition

1

02/14/2018 数据挖掘导论，第 2 版 2

关联规则挖掘

- 给定一组交易，根据交易中其他项目的出现情况找到预测项目出现的规则

市场一揽子交易

TID 项目

1 面包，牛奶

2 面包，尿布，啤酒，鸡蛋 3 牛奶，尿布，啤酒，可乐 4 面包，牛奶，尿布，啤酒 5 面包，牛奶，尿布，可乐

关联规则示例

{尿布} → {啤酒}，

{牛奶，面包} → {鸡蛋，可乐}，{啤酒，面包} → {牛奶}，

暗示意味着共同发生，而不是因果关系！

02/14/2018 数据挖掘导论，第 2 版 3

定义:频繁项目集

- 项目集

一个或多个项目的集合

u 示例:{牛奶、面包、尿布}

kitemset

包含 k 个项目的项目集

- 支持计数(σ)

项目集出现的频率，例如 $\sigma(\{\text{牛奶、面包、尿布}\}) = 2$

- 支持

包含项目集的事务的一部分

例如, $s(\{\text{牛奶、面包、尿布}\}) = 2/5$

●频繁项目集

支持大于或等于最小阈值的项集

TID 项目

1 面包, 牛奶

2 面包, 尿布, 啤酒, 鸡蛋 3 牛奶, 尿布, 啤酒, 可乐 4 面包, 牛奶, 尿布, 啤酒 5 面包, 牛奶, 尿布, 可乐

02/14/2018 数据挖掘导论，第 2 版 4

定义:关联规则

示例:

{牛奶, 尿布} {啤酒}

0.4 5

2

| T |

(牛奶尿布、啤酒) = σs

0.67 3

2

(牛奶、尿布)

(牛奶、尿布、啤酒) = σc

●关联规则

形式为 $X \rightarrow Y$ 的蕴涵表达式，其中 X 和 Y 是项集示例:

{牛奶, 尿布} \rightarrow {啤酒}

●规则评估指标

支持

u 包含 X 和 Y 的事务的比例

信心(c)

u 衡量 Y 中的项目在包含 X 的交易中出现的频率

TID 项目

1 面包, 牛奶

2 面包, 尿布, 啤酒, 鸡蛋 3 牛奶, 尿布, 啤酒, 可乐 4 面包, 牛奶, 尿布, 啤酒 5 面包, 牛奶, 尿布, 可乐

关联规则挖掘任务

●给定一组事务 T，关联规则挖掘的目标是找到所有具有支持 $\geq \text{min sump}$ 阈值置信度 $\geq \text{minconf}$ 阈值

●布鲁塞尔方法:

列出所有可能的关联规则

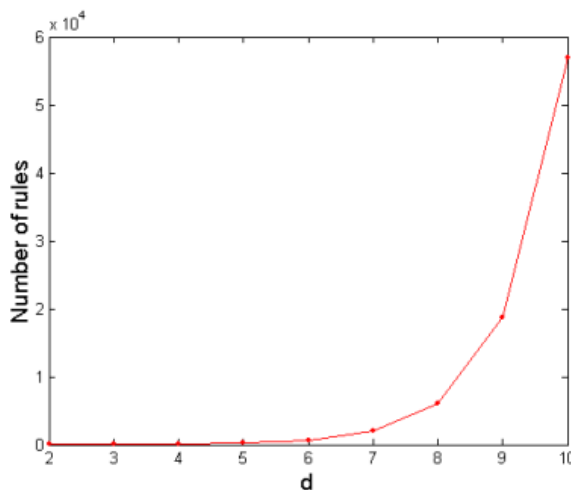
计算每个规则的支持度和可信度，删除未通过 minsup 和 minconf 的规则

计算禁止!

Computational Complexity

● Given d unique items:

- Total number of itemsets = 2^d
- Total number of possible association rules:



$$R = \sum_{k=1}^{d-1} \left[\binom{d}{k} \times \sum_{j=1}^{d-k} \binom{d-k}{j} \right]$$
$$= 3^d - 2^{d+1} + 1$$

If d=6, R = 602 rules

02/14/2018

Introduction to Data Mining, 2nd Edition

6

挖掘关联规则

规则示例:

{牛奶, 尿布} → {啤酒} (s=0.4, c = 0.67) {牛奶, 啤酒} → {尿布} (s=0.4, c = 1.0) {尿布, 啤酒} → {牛奶} (s=0.4, c = 0.67) {啤酒} → {牛奶, 尿布} (s=0.4, c = 0.67) {尿布} → {牛奶, 啤酒} (s=0.4, c = 0.5) {牛奶} → {尿布, 啤酒} (s=0)

TID 项目

1 面包, 牛奶

2 面包, 尿布, 啤酒, 鸡蛋 3 牛奶, 尿布, 啤酒, 可乐 4 面包, 牛奶, 尿布, 啤酒 5 面包, 牛奶, 尿布, 可乐

观察:

以上所有规则都是相同项目集的二进制分区:{牛奶、尿布、啤酒}

源自相同项目集的规则具有相同的支持，但可能具有不同的可信度

因此，我们可以将支持和信心要求分离开来

02/14/2018 数据挖掘导论，第2版 8

挖掘关联规则

●两步方法:

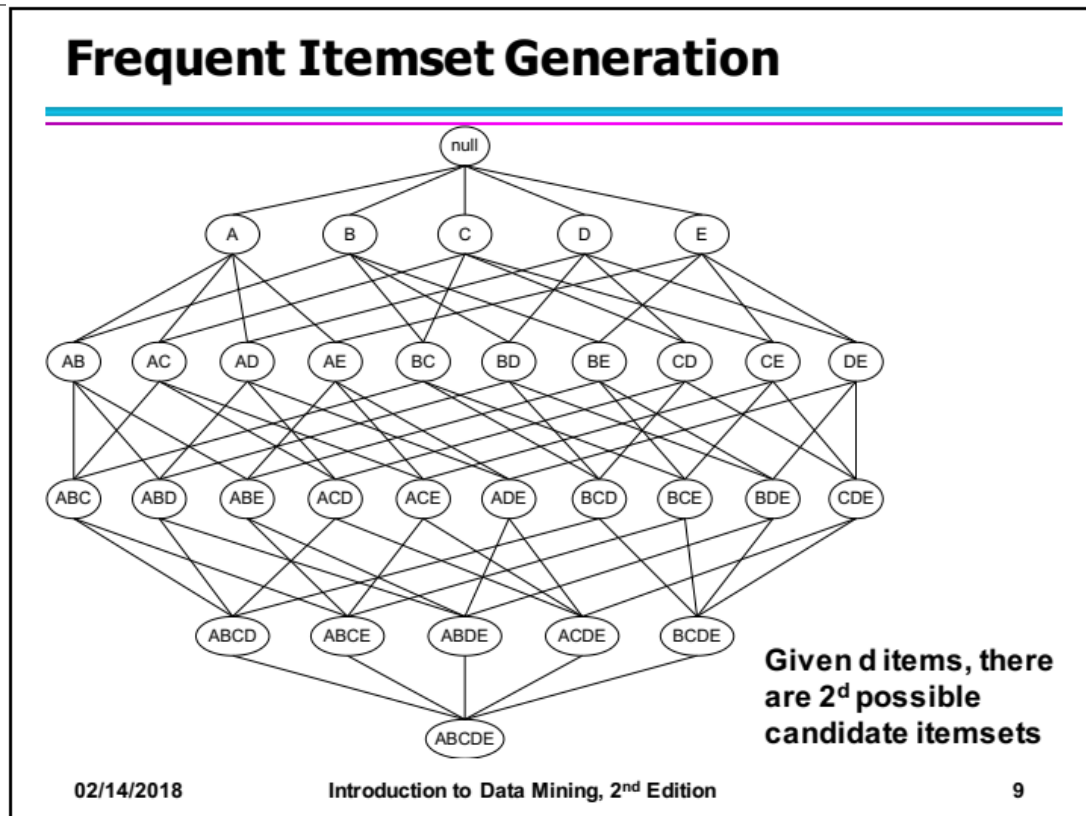
1. 频繁项目集生成

生成支持 $\geq \min \text{sump}$ 的所有项目集

2. 规则生成

从每个频繁项集生成高置信度规则，其中每个规则是频繁项集的二进制划分

●频繁项目集生成的计算成本仍然很高



02/14/2018 数据挖掘导论，第2版 10

频繁项目集生成

●布鲁塞尔方法:

格中的每个项目集都是一个候选频繁项目集通过扫描

资料库

将每个事务与每个候选事务进行匹配，复杂度为 $O(NMw) = >$ 昂贵，因为 $M = 2^d !!!$

TID 项目

1 面包，牛奶

2 面包，尿布，啤酒，鸡蛋 3 牛奶，尿布，啤酒，可乐 4 面包，牛奶，尿布，啤酒 5 面包，牛奶，尿布，可乐

N

交易清单

候选人

M

w

02/14/2018 数据挖掘导论，第 2 版 11

频繁项集生成策略

- 减少候选人数(百万)

完整搜索: $M=2^d$

使用修剪技术来减少 M

- 减少交易数量(N)

随着项目集大小的增加，减少 N 的大小。DHP 和垂直挖掘算法使用

- 减少比较次数(海里)

使用高效的数据结构来存储候选项或事务

不需要将每个候选人与每个交易进行匹配

02/14/2018 数据挖掘导论，第 2 版 12

减少候选人数

- 先验原则:

如果一个项目集是频繁的，那么它的所有子集也必须是频繁的

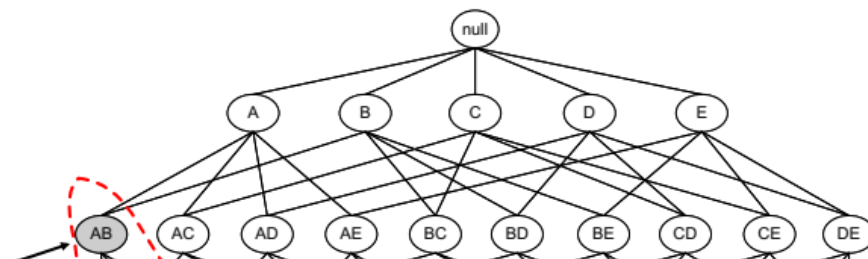
- 由于支持措施的以下属性，先验原则成立:

项目集的支持永远不会超过其子集的支持

这就是众所周知的支持物的反质子性质

$$\forall x, y: (x \subseteq y) s(x) \geq s(y)$$

Illustrating Apriori Principle



Illustrating Apriori Principle

TID	Items
1	Bread, Milk
2	Beer, Bread, Diaper, Eggs
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Bread, Coke, Diaper, Milk



Items (1-itemsets)

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Minimum Support = 3

If every subset is considered,
 ${}^6C_1 + {}^6C_2 + {}^6C_3$
 $6 + 15 + 20 = 41$
With support-based pruning,
 $6 + 6 + 4 = 16$

Illustrating Apriori Principle

TID	Items
1	Bread, Milk
2	Beer, Bread, Diaper, Eggs
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Bread, Coke, Diaper, Milk



Items (1-itemsets)

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Minimum Support = 3

If every subset is considered,
 ${}^6C_1 + {}^6C_2 + {}^6C_3$
 $6 + 15 + 20 = 41$
 With support-based pruning,
 $6 + 6 + 4 = 16$

02/14/2018 数据挖掘导论，第2版 16

阐释先验原则

项目计数

面包 4

可乐 2

牛奶 4

啤酒 3

尿布 4

鸡蛋 1

项目集

{面包, 牛奶}

{面包, 啤酒}

{面包, 尿布}

{啤酒, 牛奶}

{尿布, 牛奶}

{啤酒, 尿布}

项目(1 项目集)

成对(2 个项目集)

(无需生成

涉及可口可乐的候选人

或鸡蛋)

最低支持= 3

如果考虑每个子集，

$C_1 + C_2 + C_3 = 6 + 15 + 20 = 41$

通过基于支持的修剪，

$$6 + 6 + 4 = 16$$

02/14/2018 数据挖掘导论，第 2 版 17

阐释先验原则

项目计数

面包 4

可乐 2

牛奶 4

啤酒 3

尿布 4

鸡蛋 1

项目集计数

{面包, 牛奶} 3

{啤酒, 面包} 2

{面包, 尿布} 3

{啤酒, 牛奶} 2

{尿布, 牛奶} 3

{啤酒, 尿布} 3

项目(1 项目集)

成对(2 个项目集)

(无需生成

涉及可口可乐的候选人

或鸡蛋)

最低支持= 3

如果考虑每个子集，

$$C1 + C2 + C3 \quad 6 + 15 + 20 = 41$$

通过基于支持的修剪，

$$6 + 6 + 4 = 16$$

02/14/2018 数据挖掘导论，第 2 版 18

阐释先验原则

项目计数

面包 4

可乐 2

牛奶 4

啤酒 3

尿布 4

鸡蛋 1

项目集计数

{面包, 牛奶} 3

{面包, 啤酒} 2

{面包, 尿布} 3

{牛奶, 啤酒} 2

{牛奶, 尿布} 3

{啤酒, 尿布} 3

项目集

{啤酒、尿布、牛奶}
 {啤酒、面包、尿布}
 {面包、尿布、牛奶}
 {啤酒、面包、牛奶}
 项目(1 项目集)
 成对(2 个项目集)
 (无需生成
 涉及可口可乐的候选人
 或鸡蛋)
 三元组(3 项集)最小支持度 = 3
 如果考虑每个子集，
 $C1 + C2 + C3 = 6 + 15 + 20 = 41$
 通过基于支持的修剪，
 $6 + 6 + 4 = 16$

02/14/2018 数据挖掘导论，第 2 版 19

阐释先验原则

项目计数

面包 4

可乐 2

牛奶 4

啤酒 3

尿布 4

鸡蛋 1

项目集计数

{面包, 牛奶} 3

{面包, 啤酒} 2

{面包, 尿布} 3

{牛奶, 啤酒} 2

{牛奶, 尿布} 3

{啤酒, 尿布} 3

项目集计数

{啤酒、尿布、牛奶}

{啤酒、面包、尿布}

{面包、尿布、牛奶}

{啤酒、面包、牛奶}

2

2

2

1

项目(1 项目集)

成对(2 个项目集)

(无需生成

涉及可口可乐的候选人

或鸡蛋)

三元组(3 项集)最小支持度 = 3

如果考虑每个子集，

$$C1 + C2 + C3 = 6 + 15 + 20 = 41$$

通过基于支持的修剪，

$$6 + 6 + 4 = 16$$

$$6 + 6 + 1 = 13$$

02/14/2018 数据挖掘导论，第2版 20

Apriori 算法

L_k : 频繁工具包 L_k : 候选工具包

● 算法

设 $k=1$

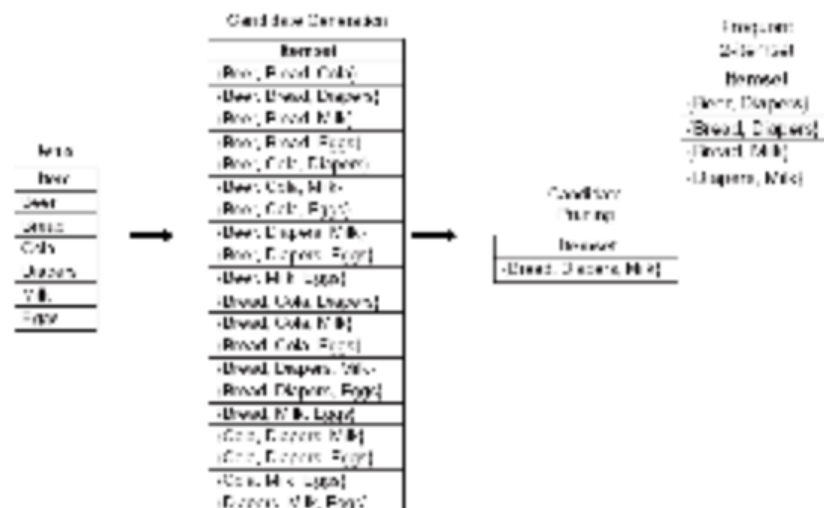
生成 $F1 = \{\text{频繁 1 项目集}\}$ 重复，直到 L_k 为空

候选项生成: 从候选项剪枝中生成候选项: 剪枝候选项集

支持计数: 通过扫描数据库来计算 L_{k+1} 中每个候选的支持度

u 候选人淘汰: 淘汰 L_{k+1} 中不频繁的候选人，只留下频繁的候选人 $\Rightarrow L_{k+1}$

Candidate Generation: Brute-force method



Candidate Generation: Merge F_{k-1} and F_1 itemsets

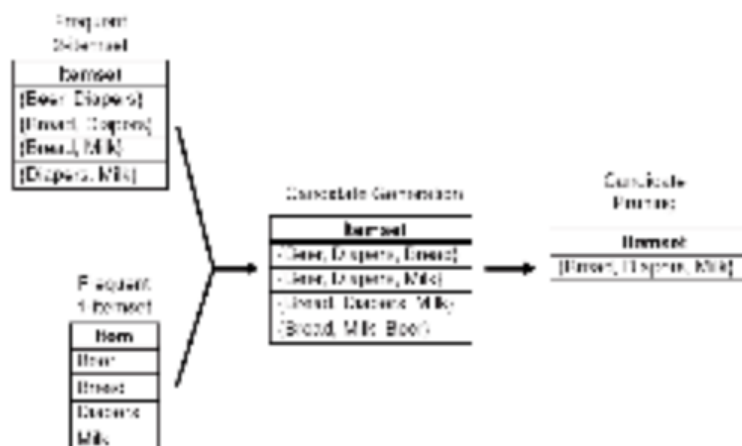


Figure 5.7. Generating and pruning candidate k -itemsets by merging a frequent $(k-1)$ -itemset with a frequent item. Note that some of the candidates are unnecessary because their subsets are infrequent.

Candidate Generation: Fk-1 x Fk-1 Method

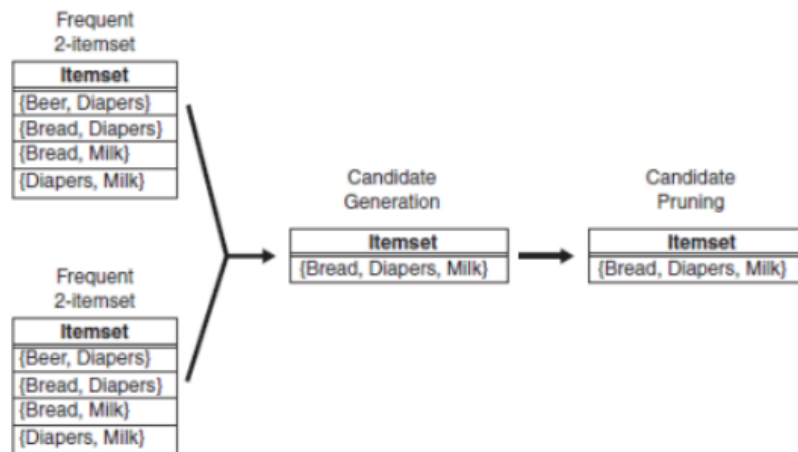


Figure 6.8. Generating and pruning candidate k -itemsets by merging pairs of frequent $(k-1)$ -itemsets.

02/14/2018

Introduction to Data Mining, 2nd Edition

23

02/14/2018 数据挖掘导论，第 2 版 24

候选生成:Fk-1 x Fk-1 方法

- 如果两个频繁(k_1)项目集的第一个(k_2)项目相同，则将其合并
- $F_3 = \{\text{美国广播公司、英国广播公司、英国广播公司、美国广播公司、美国广播公司、英国广播公司、CDE}\}$

合并(ABC, ABD) = ABCD 合并(ABC, ABE) = ABCE 合并(ABD, ABE) = ABDE

不要合并(ABD、ACD)，因为它们只共享长度为 1 的前缀，而不是长度为 2 的前缀

02/14/2018 数据挖掘导论，第 2 版 25

候选修剪

- 让 $F_3 = \{\text{美国广播公司、英国广播公司、英国广播公司、美国广播公司、美国广播公司、英国广播公司、CDE}\}$ 成为 3 个频繁项目集的集合
 - $L_4 = \{\text{ABCD, ABCE, ABDE}\}$ 是生成的一组候选 4 项集(来自上一张幻灯片)
 - 候选修剪
- 修剪 ABCE，因为 ACE 和 BCE 是罕见的
- 候选修剪后: $L_4 = \{\text{ABCD}\}$

02/14/2018 数据挖掘导论，第 2 版 26

替代 Fk-1 x Fk-1 方法

- 如果第一个项目的最后(k_2)项与第二个项目的第一(k_2)项相同，则合并两个频繁(k_1)项目集。
 - $F_3 = \{\text{美国广播公司、英国广播公司、英国广播公司、美国广播公司、美国广播公司、英国广播公司、CDE}\}$
- 合并(美国广播公司、英国广播公司)=美国广播公司 合并(美国广播公司、英国广播公司)=美国广播公司
- 合并(美国广播公司、CDE) = ACDE 合并(CDE 广播公司)= BCDE

02/14/2018 数据挖掘导论，第 2 版 27

备选 Fk-1×Fk-1 方法的候选修剪

- 让 $F3 = \{\text{美国广播公司、英国广播公司、英国广播公司、美国广播公司、美国广播公司、英国广播公司、CDE}\}$ 成为 3 个频繁项目集的集合

- $L4 = \{\text{ABCD, ABDE, ACDE, BCDE}\}$ 是生成的一组候选 4 项集(来自上一张幻灯片)

- 候选修剪

修剪阿卜德，因为阿德不经常修剪 ACDE，因为 ACE 和阿德不经常修剪 BCDE，因为 BCE

- 候选修剪后: $L4 = \{\text{ABCD}\}$

02/14/2018 数据挖掘导论，第 2 版 28

阐释先验原则

项目计数

面包 4

可乐 2

牛奶 4

啤酒 3

尿布 4

鸡蛋 1

项目集计数

$\{\text{面包, 牛奶}\} 3$

$\{\text{面包, 啤酒}\} 2$

$\{\text{面包, 尿布}\} 3$

$\{\text{牛奶, 啤酒}\} 2$

$\{\text{牛奶, 尿布}\} 3$

$\{\text{啤酒, 尿布}\} 3$

项目集计数

$\{\text{面包、尿布、牛奶}\} 2$

项目(1 项目集)

成对(2 个项目集)

(无需生成

涉及可口可乐的候选人

或鸡蛋)

三元组(3 项集)最小支持度 = 3

如果考虑每个子集，

$C1 + C2 + C3 = 6 + 15 + 20 = 41$

通过基于支持的修剪，

$6 + 6 + 1 = 13$ 使用 Fk-1Fk-1 方法生成候选结果

只有一个 3 项目集。这在支撑后被消除

计数步骤。

02/14/2018 数据挖掘导论，第 2 版 29

支持候选项集的计数

- 扫描事务数据库以确定每个候选项目集的支持

必须将每个候选项集与每个事务进行匹配，这是一项昂贵的操作

TID 项目

1 面包, 牛奶
2 啤酒, 面包, 尿布, 鸡蛋 3 啤酒, 可乐, 尿布, 牛奶 4 啤酒, 面包, 尿布, 牛奶 5 面包, 可乐, 尿布, 牛奶
项目集
{啤酒、尿布、牛奶}
{啤酒、面包、尿布}
{面包、尿布、牛奶}
{啤酒、面包、牛奶}

02/14/2018 数据挖掘导论, 第 2 版 30
支持候选项集的计数
●为了减少比较次数, 将候选项集存储在哈希结构中
不要将每个事务与每个候选项进行匹配, 而是将其与哈希桶中包含的候选项进行匹配
TID 项目
1 面包, 牛奶
2 面包, 尿布, 啤酒, 鸡蛋 3 牛奶, 尿布, 啤酒, 可乐 4 面包, 牛奶, 尿布, 啤酒 5 面包, 牛奶, 尿布, 可乐
N
事务哈希结构
k
大量

02/14/2018 数据挖掘导论, 第 2 版 31
支持计数: 一个例子
假设您有 15 个长度为 3 的候选项集:
{1 4 5}、{1 2 4}、{4 5 7}、{1 2 5}、{4 5 8}、{1 5 9}、{1 3 6}、{2 3 4}、{5 6 7}、{3 4 5},
{3 5 6}、{3 5 7}、{6 8 9}、{3 6 7}、{3 6 8}
事务(1, 2, 3, 5, 6)支持多少项集?
1 2 3 5 6
交易, t
1 2 3 5 6 2 3 5 6
1 2 3 5 6 1 3 5 6 1 5 6 2 3 5 6 2 5 6
3 5 6
1 2 3 1 2 5 1 2 6
1 3 5
1 3 6 2 3 6
3 个项目的子集
1 级
二级
三级
3 5 6

02/14/2018 数据挖掘导论, 第 2 版 32
支持使用哈希树进行计数
2 3 4 5 6 7
1 3 6
1 2 4
4 5 7

4 5 8

1 5 9

3 4 5 3 5 6

3 5 7 6 8 9

3 6 7 3 6 8

1, 4, 7

2, 5, 8

3, 6, 9

散列函数

假设您有 15 个长度为 3 的候选项集:

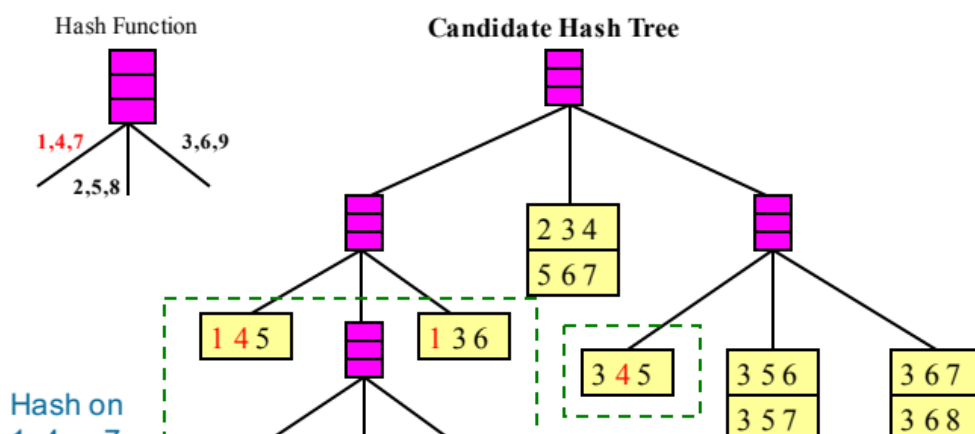
{1 4 5}、{1 2 4}、{4 5 7}、{1 2 5}、{4 5 8}、{1 5 9}、{1 3 6}、{2 3 4}、{5 6 7}、{3 4 5}、
{3 5 6}、{3 5 7}、{6 8 9}、{3 6 7}、{3 6 8}

您需要:

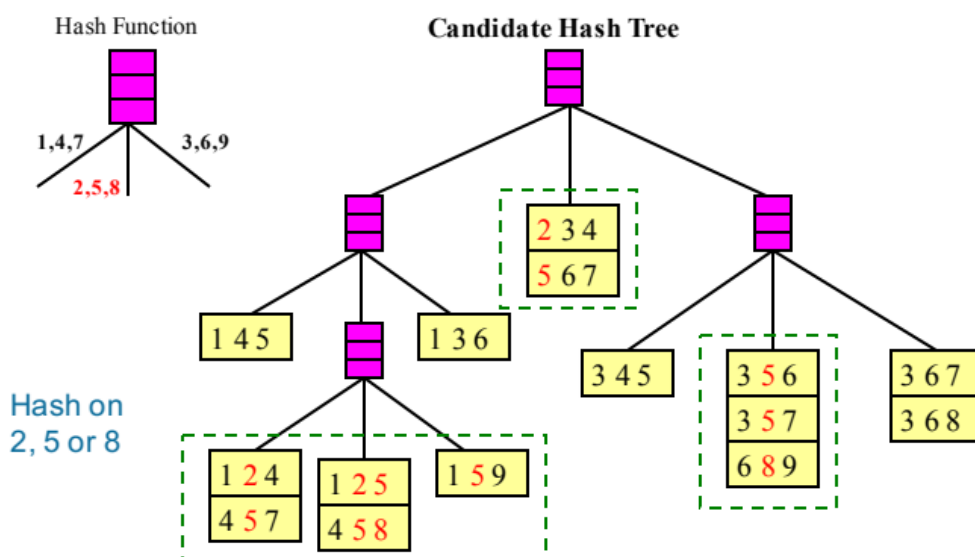
哈希函数

最大叶大小:存储在叶节点中的最大项目集数量(如果数量为
候选项集超过最大叶大小, 分割节点)

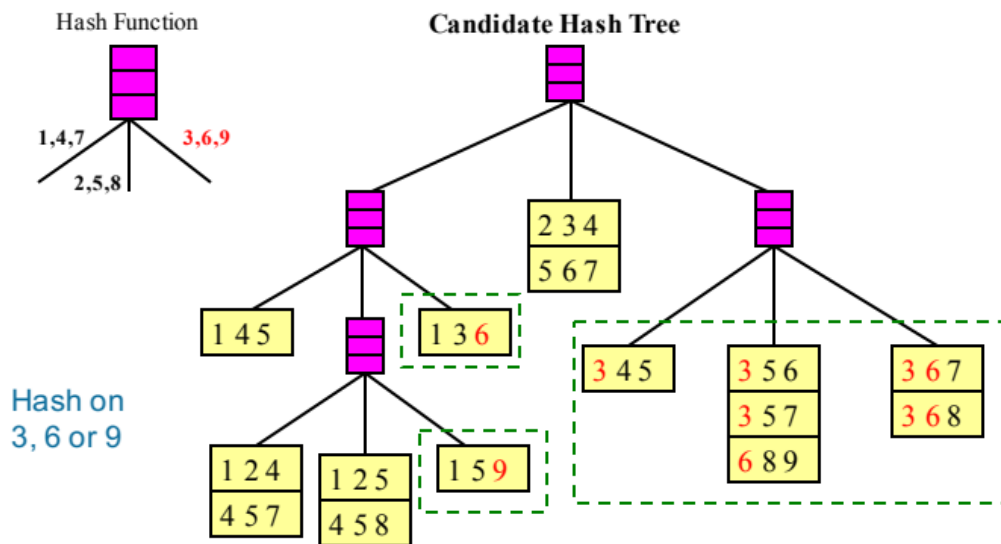
Support Counting Using a Hash Tree



Support Counting Using a Hash Tree



Support Counting Using a Hash Tree

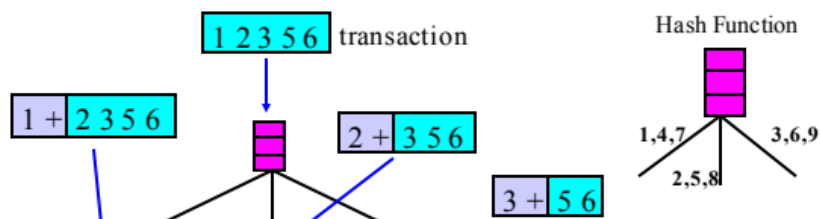


02/14/2018

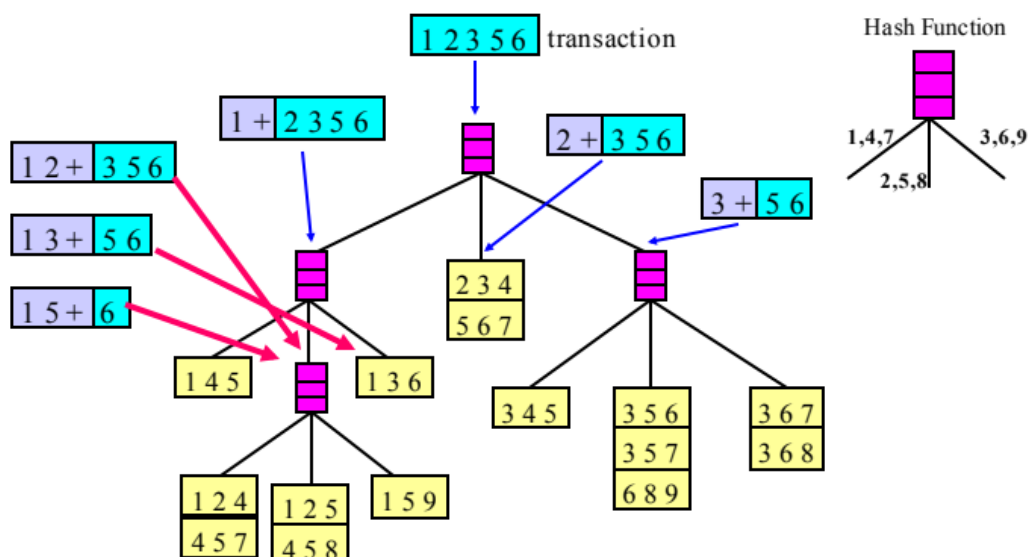
Introduction to Data Mining, 2nd Edition

35

Support Counting Using a Hash Tree



Support Counting Using a Hash Tree

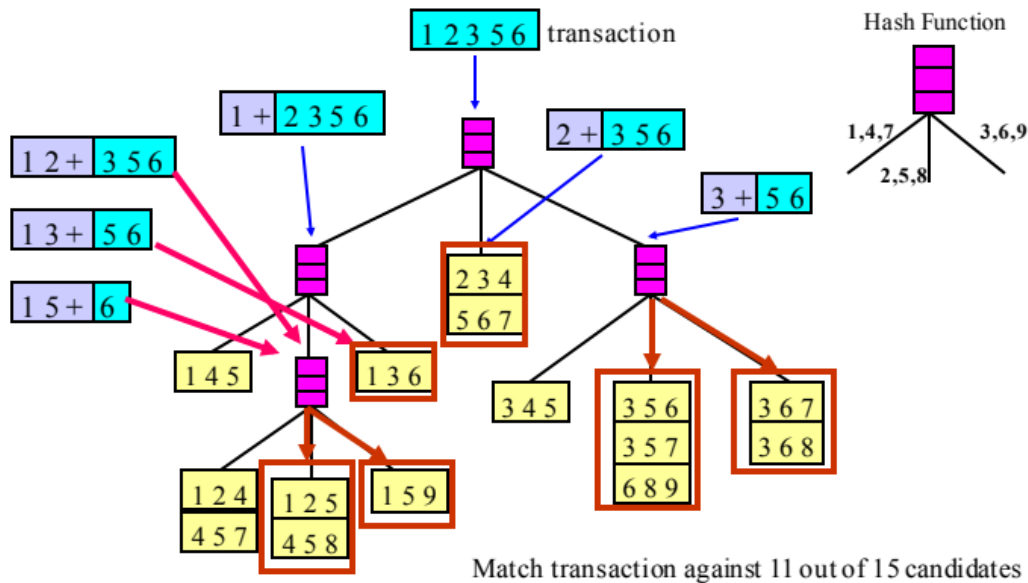


02/14/2018

Introduction to Data Mining, 2nd Edition

37

Support Counting Using a Hash Tree



02/14/2018

Introduction to Data Mining, 2nd Edition

38

02/14/2018 数据挖掘导论，第 2 版 39

规则生成

●给定一个频繁项集 L ，找到 $C \subseteq L$ 的所有非空子集，使得 $f \rightarrow L \setminus f$ 满足最小置信度要求

如果 $\{A, B, C, D\}$ 是频繁项集，候选规则：

$ABC \rightarrow D$, $ABD \rightarrow C$, $ACD \rightarrow B$, $BCD \rightarrow A$,

$A \rightarrow BCD$, $B \rightarrow ACD$, $C \rightarrow ABD$, $D \rightarrow ABC$

$AB \rightarrow CD$, $AC \rightarrow BD$, $AD \rightarrow BC$, $BC \rightarrow AD$,

$BD \rightarrow AC$, $CD \rightarrow AB$,

●如果 $|L| = k$ ，则有 $2^k - 2$ 个候选关联规则(忽略 $L \rightarrow \emptyset$ 和 $\emptyset \rightarrow L$)

02/14/2018 数据挖掘导论，第 2 版 40

规则生成

●一般来说，信心没有反单调的特性

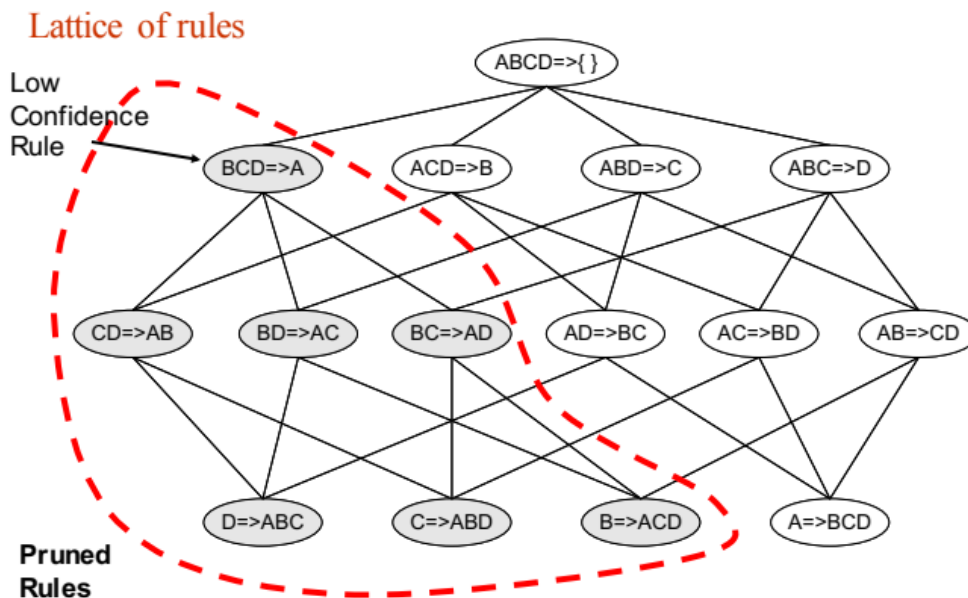
$c(ABC \rightarrow D)$ 可以大于或小于 $c(AB \rightarrow D)$

●但是从同一项目集中生成的规则的可信度具有反光子属性

例如，假设 $\{A, B, C, D\}$ 是一个频繁的 4 项集： $c(ABC \rightarrow D) \geq c(AB \rightarrow CD) \geq c(A \rightarrow BCD)$

置信度是反质子 w.r.t. 规则的 RHS 上的项目数

Rule Generation for Apriori Algorithm



02/14/2018

Introduction to Data Mining, 2nd Edition

41

Association Analysis: Basic Concepts and Algorithms

Algorithms and Complexity

02/14/2018

Introduction to Data Mining, 2nd Edition

42

02/14/2018 数据挖掘导论，第2版 43

影响先验复杂性的因素

- 最低支持阈值的选择

降低支持阈值会产生更多的频繁项集，这可能会增加候选项的数量和频繁项集的最大长度

- 数据集的维数(项目数)

如果频繁项目的数量也增加，计算和输入/输出成本也可能增加

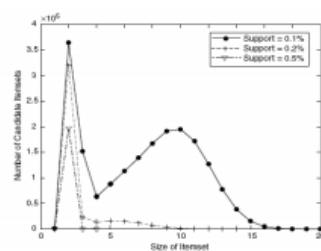
- 数据库的大小

由于 Apriori 进行多次遍历，算法的运行时间可能会随着事务数量的增加而增加

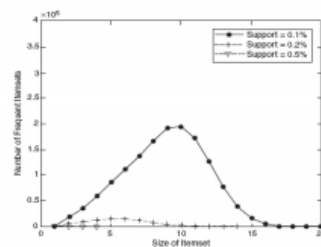
- 平均交易宽度

事务宽度随着更密集的数据集而增加这可能会增加频繁项集和散列树遍历的最大长度(事务中子集的数量随着其宽度)

Factors Affecting Complexity of Apriori

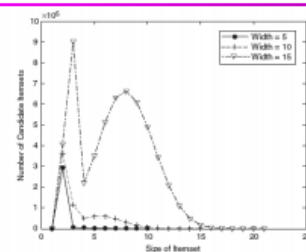


(a) Number of candidate itemsets.

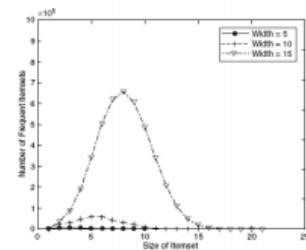


(b) Number of frequent itemsets.

Figure 6.13. Effect of support threshold on the number of candidate and frequent itemsets.



(a) Number of candidate itemsets.



(b) Number of frequent itemsets.

Figure 6.14. Effect of average transaction width on the number of candidate and frequent itemsets.

02/14/2018

Introduction to Data Mining, 2nd Edition

44

02/14/2018 数据挖掘导论，第 2 版 45

频繁项集的紧凑表示

- 有些项目集是多余的，因为它们有与其超集相同的支持

- 频繁项目集的数量●需要一个紧凑的表示

TID A1 A2 A3 A4 A5 A6 A7 A8 A9 A10 B1 B2 B3 B4 B5 B6 B7 B8 B9 B10 C1 C2 C3 C4 C5 C6 C7 C8 C9 C10

```
1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
2 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
3 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
4 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
5 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
6 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0
7 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0
8 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0
9 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0
```

```

100000000000011111111100000000000
110000000000000000000000011111111
120000000000000000000000011111111
130000000000000000000000011111111
140000000000000000000000011111111
150000000000000000000000011111111

```

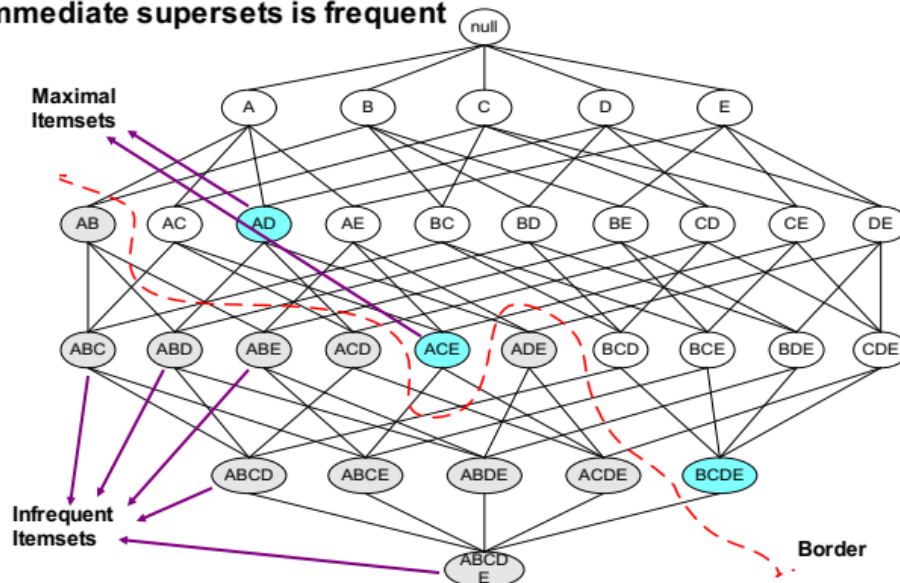
√

=

$| \setminus | \setminus | = \times k k$

Maximal Frequent Itemset

An itemset is maximal frequent if it is frequent and none of its immediate supersets is frequent



02/14/2018

Introduction to Data Mining, 2nd Edition

46

What are the Maximal Frequent Itemsets in this Data?

TID	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1

Minimum support threshold = 5

An illustrative example

	Items									
	A	B	C	D	E	F	G	H	I	J
1										
2										
3										

Support threshold (by count) : 5
Frequent itemsets: ?

An illustrative example

	Items									
	A	B	C	D	E	F	G	H	I	J
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										

Support threshold (by count) : 5
Frequent itemsets: {F}

An illustrative example

Transactions	Items									
	A	B	C	D	E	F	G	H	I	J
	1									
	2									
	3									
	4									
	5									
	6									
	7									
	8									
	9									
	10									

Support threshold (by count) : 5
Frequent itemsets: {F}

Support threshold (by count): 4
Frequent itemsets: ?

An illustrative example

Transactions	Items									
	A	B	C	D	E	F	G	H	I	J
	1									
	2									
	3									
	4									
	5									
	6									
	7									
	8									
	9									
	10									

Support threshold (by count) : 5
Frequent itemsets: {F}

Support threshold (by count): 4
Frequent itemsets: {B, {F}, {E,F}, {J}

An illustrative example

Transactions	Items									
	A	B	C	D	E	F	G	H	I	J
	1									
	2									
	3									
	4									
	5									
	6									
	7									
	8									
	9									
	10									

Support threshold (by count) : 5
Frequent itemsets: {F}

Support threshold (by count): 4
Frequent itemsets: {B, {F}, {E,F}, {J}

Support threshold (by count): 3
Frequent itemsets: ?

An illustrative example

Transactions	Items										
	A	B	C	D	E	F	G	H	I	J	
1											Support threshold (by count) : 5 Frequent itemsets: {F}
2											Support threshold (by count): 4 Frequent itemsets: {B, {F}, {E,F}, {J}
3											Support threshold (by count): 3 Frequent itemsets: All subsets of {C,D,E,F} + {J}
4											
5											
6											
7											
8											
9											
10											

02/14/2018

Introduction to Data Mining, 2nd Edition

53

An illustrative example

Transactions	Items										
	A	B	C	D	E	F	G	H	I	J	
1											Support threshold (by count) : 5 Frequent itemsets: {F} Maximal itemsets: ?
2											Support threshold (by count): 4 Frequent itemsets: {B, {F}, {E,F}, {J} Maximal itemsets: ?
3											

An illustrative example

Transactions	Items										
	A	B	C	D	E	F	G	H	I	J	
1											Support threshold (by count) : 5 Frequent itemsets: {F} Maximal itemsets: {F}
2											Support threshold (by count): 4 Frequent itemsets: {B, {F}, {E,F}, {J} Maximal itemsets: ?
3											Support threshold (by count): 3 Frequent itemsets: All subsets of {C,D,E,F} + {J} Maximal itemsets: ?
4											
5											
6											
7											
8											
9											
10											

02/14/2018

Introduction to Data Mining, 2nd Edition

55

An illustrative example

Transactions	Items										
	A	B	C	D	E	F	G	H	I	J	

Support threshold (by count) : 5

Frequent itemsets: {F}

Maximal itemsets: {F}

Support threshold (by count): 4

Frequent itemsets: {B}, {F}, {E,F}, {J}

Maximal itemsets: {E,F}, {J}

Support threshold (by count): 3

Frequent itemsets:

All subsets of {C,D,E,F} + {J}

Maximal itemsets: ?

An illustrative example

Transactions	Items										
	A	B	C	D	E	F	G	H	I	J	

Support threshold (by count) : 5

Frequent itemsets: {F}

Maximal itemsets: {F}

Support threshold (by count): 4

Frequent itemsets: {B}, {F}, {E,F}, {J}

Maximal itemsets: {E,F}, {J}

Another illustrative example

Transactions	Items										
	A	B	C	D	E	F	G	H	I	J	

Support threshold (by count) : 5

Maximal itemsets: {A}, {B}, {C}

Support threshold (by count): 4

Maximal itemsets: {A,B}, {A,C}, {B,C}

Support threshold (by count): 3

Maximal itemsets: {A,B,C}

封闭项集

- 如果项集 X 的直接超集都没有项集 X 的支持，则项集 X 被关闭
- 如果至少有一个直接超集的支持计数为 X，则不关闭 X

TID 项目

- 1 {甲, 乙}
- 2 {B, C, D}
- 3 {甲、乙、丙、丁}
- 4 {甲、乙、丁}
- 5 {甲、乙、丙、丁}

项目集支持

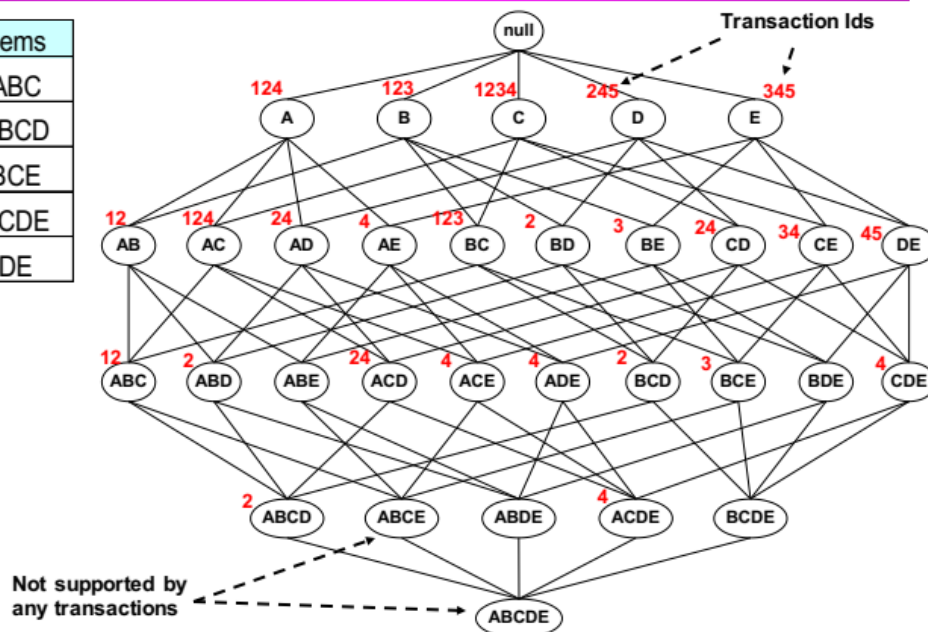
- {A} 4
- {B} 5
- {丙} 3
- {D} 4
- {甲, 乙} 4
- {甲, 丙} 2
- {甲, 丁} 3
- {乙, 丙} 3
- {乙, 丁} 4
- {C, D} 3

项目集支持

- {甲, 乙, 丙} 2
- {甲, 乙, 丁} 3
- {甲、丙、丁} 2
- {乙, 丙, 丁} 2
- {甲, 乙, 丙, 丁} 2

Maximal vs Closed Itemsets

TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE



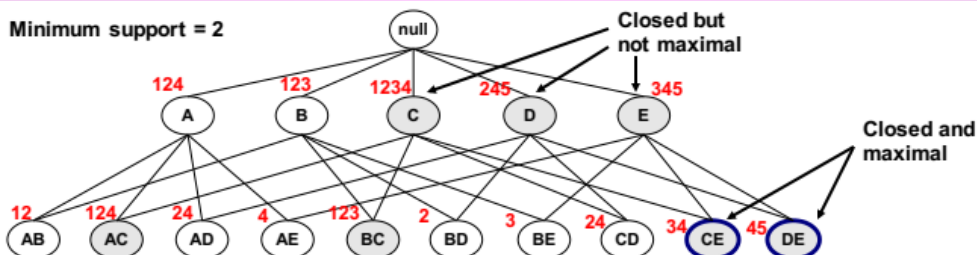
02/14/2018

Introduction to Data Mining, 2nd Edition

60

Maximal vs Closed Frequent Itemsets

Minimum support = 2



What are the Closed Itemsets in this Data?

TID	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1

02/14/2018

Introduction to Data Mining, 2nd Edition

62

Example 1

Transactions	Items										Itemsets	Support (counts)	Closed itemsets
	A	B	C	D	E	F	G	H	I	J			
1													
2											{C}	3	
3											{D}	2	
4											{C,D}	2	
5													
6													
7													
8													
9													
10													

02/14/2018

Introduction to Data Mining, 2nd Edition

63

Example 1

Transactions	Items										Itemsets	Support (counts)	Closed itemsets
	A	B	C	D	E	F	G	H	I	J			
1													
2											{C}	3	✓
3											{D}	2	

Example 2

Transactions	Items										Itemsets	Support (counts)	Closed itemsets
	A	B	C	D	E	F	G	H	I	J			
1													
2											{C}	3	
3											{D}	2	
4											{E}	2	
5											{C,D}	2	
6											{C,E}	2	
7											{D,E}	2	
8											{C,D,E}	2	
9													
10													

02/14/2018

Introduction to Data Mining, 2nd Edition

65

Example 2

Transactions	Items										Itemsets	Support (counts)	Closed itemsets
	A	B	C	D	E	F	G	H	I	J			
1													
2											{C}	3	✓
3											{D}	2	
4											{E}	2	
5											{C,D}	2	
6											{C,E}	2	
7											{D,E}	2	
8											{C,D,E}	2	✓
9													
10													

02/14/2018

Introduction to Data Mining, 2nd Edition

66

Example 3

Transactions	Items										Closed itemsets: {C,D,E,F}, {C,F}
	A	B	C	D	E	F	G	H	I	J	
1											
2											
3											

Example 4

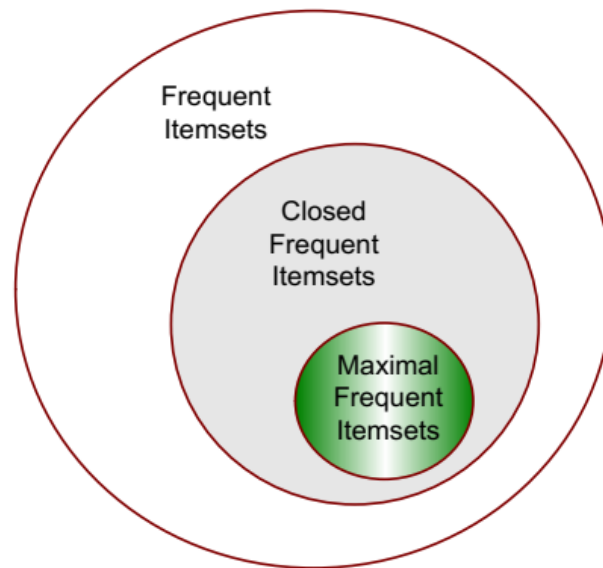
Transactions	Items										Closed itemsets: {C,D,E,F}, {C}, {F}
	A	B	C	D	E	F	G	H	I	J	
1											
2											
3											
4											
5											
6											
7											
8											
9											
10											

02/14/2018

Introduction to Data Mining, 2nd Edition

68

Maximal vs Closed Itemsets



02/14/2018

Introduction to Data Mining, 2nd Edition

69

02/14/2018 数据挖掘导论，第 2 版 70

示例问题

●给定以下交易数据集(黑色单元格表示交易中存在项目)和 20%的支持阈值，回答以下问题

- 每个数据集的频繁项集数量是多少？哪个数据集将产生最大频繁项目集数？
- 哪个数据集将产生最长的频繁项目集？
- 哪个数据集会产生最大支持度最高的频繁项目集？
- 哪个数据集将产生频繁项目集，其中包含具有广泛不同支持的项目级别(即包含混合支持项的项目集，范围从 20%到超过 70%)？
- 每个数据集的最大频繁项目集的数量是多少？哪个数据集将产生最多数量的最大频繁项集？
- 每个数据集的封闭频繁项目集的数量是多少？哪个数据集将产生最大数量的闭频繁项目集？

02/14/2018 数据挖掘导论，第 2 版 71

模式评估

- 关联规则算法可以产生大量规则
 - 兴趣度度量可用于删减/排列模式
- 在最初的表述中，支持和信心是唯一使用的衡量标准

02/14/2018 数据挖掘导论，第 2 版 72

计算兴趣度

- 给定 $X \rightarrow Y$ 或 $\{X, Y\}$ ，计算兴趣度所需的信息可以从列联表中获得

Y Y

X f11 f10 f1+

X f01 f00 f0+

f+1 f+0 N

相依表

f11:支持 X 和 Y f10:支持 X 和 Y f01:支持 X 和 Y f00:支持 X 和 Y

用于定义各种衡量标准◆支持、信心、基尼系数、

熵等。

Drawback of Confidence

Custo mers	Tea	Coffee	...
C1	0	1	...
C2	1	0	...
C3	1	1	...
C4	1	0	...
...			

	Coffee	<u>Coffee</u>	
Tea	15	5	20
<u>Tea</u>	75	5	80
	90	10	100

Association Rule: Tea → Coffee

Confidence $\equiv P(\text{Coffee}|\text{Tea}) = 15/20 = 0.75$

Confidence > 50%, meaning people who drink tea are more likely to drink coffee than not drink coffee

So rule seems reasonable

02/14/2018

Introduction to Data Mining, 2nd Edition

73

02/14/2018 数据挖掘导论，第 2 版 74

自信的缺点

咖啡咖啡

茶 15 5 20

茶 75 5 80

90 10 100

关联规则:茶→咖啡

信心= $P(\text{咖啡}|\text{茶}) = 15/20 = 0.75$

但是 $P(\text{咖啡}) = 0.9$ ，这意味着知道一个人喝酒

茶减少人喝咖啡的可能性！

请注意， $P(\text{咖啡}|\text{茶}) = 75/80 = 0.9375$

02/14/2018 数据挖掘导论，第 2 版 75

关联规则的度量

●那么，我们真正想要什么样的规则？

置信度($X \rightarrow Y$)应该足够高

确保购买 X 的人更有可能购买 Y 而不是不购买 Y

信心($X \rightarrow Y$) > 支持(Y)

否则, 规则将会误导, 因为在同一个交易中拥有 X 项实际上减少了拥有 Y 项的机会
是否有任何措施来捕捉这个约束? 回答:是的。他们有很多。

02/14/2018 数据挖掘导论, 第 2 版 76

统计独立性

●标准

置信度($X \rightarrow Y$) = 支持度(Y)

相当于:

$$P(Y|X) = P(Y)$$

$$P(X, Y) = P(X) \times P(Y)$$

如果 $P(X, Y) > P(X) \times P(Y)$: X 和 Y 正相关

如果 $P(X, Y) < P(X) \times P(Y)$: X 和 Y 是负相关的

Measures that take into account statistical dependence

$$\left. \begin{aligned} Lift &= \frac{P(Y | X)}{P(Y)} \\ Interest &= \frac{P(X, Y)}{P(X)P(Y)} \end{aligned} \right\} \begin{array}{l} \text{lift is used for rules while} \\ \text{interest is used for itemsets} \end{array}$$
$$PS = P(X, Y) - P(X)P(Y)$$
$$\phi - coefficient = \frac{P(X, Y) - P(X)P(Y)}{\sqrt{P(X)[1 - P(X)]P(Y)[1 - P(Y)]}}$$

Example: Lift/Interest

	Coffee	<u>Coffee</u>	
Tea	15	5	20
<u>Tea</u>	75	5	80
	90	10	100

Association Rule: Tea \rightarrow Coffee

Confidence = $P(\text{Coffee} | \text{Tea}) = 0.75$

but $P(\text{Coffee}) = 0.9$

$\Rightarrow \text{Lift} = 0.75/0.9 = 0.8333$ (< 1, therefore is negatively associated)

So, is it enough to use confidence/lift for pruning?

Lift or Interest

	Y	\bar{Y}	
X	10	0	10
\bar{X}	0	90	90
	10	90	100

$$Lift = \frac{0.1}{(0.1)(0.1)} = 10$$

	Y	\bar{Y}	
X	90	0	90
\bar{X}	0	10	10
	90	10	100

$$Lift = \frac{0.9}{(0.9)(0.9)} = 1.11$$

Statistical independence:

If $P(X,Y)=P(X)P(Y) \Rightarrow Lift = 1$

There are lots of measures proposed in the literature

#	Measure	Formula
1	ϕ -coefficient	$\frac{P(A,B) - P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
2	Goodman-Kruskal's (λ)	$\frac{\sum_j \max_k P(A_j, B_k) - \sum_k \max_j P(A_j, B_k) - \max_j P(A_j) - \max_k P(B_k)}{1 - \max_j P(A_j) - \max_k P(B_k)}$
3	Odds ratio (α)	$\frac{P(A,B)P(\bar{A},\bar{B})}{P(A,\bar{B})P(\bar{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(\bar{A},\bar{B}) - P(A,\bar{B})P(\bar{A},B)}{P(A,B)P(\bar{A},\bar{B}) + P(A,\bar{B})P(\bar{A},B)} = \frac{\alpha-1}{\alpha+1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(\bar{A},\bar{B})} - \sqrt{P(A,\bar{B})P(\bar{A},B)}}{\sqrt{P(A,B)P(\bar{A},\bar{B})} + \sqrt{P(A,\bar{B})P(\bar{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$
6	Kappa (κ)	$\frac{P(A,B) + P(\bar{A},\bar{B}) - P(A)P(B) - P(\bar{A})P(\bar{B})}{1 - P(A)P(B) - P(\bar{A})P(\bar{B})}$
7	Mutual Information (M)	$\frac{\sum_i \sum_j P(A_i, B_j) \log \frac{P(A_i, B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i) \log P(A_i), -\sum_j P(B_j) \log P(B_j))}$
8	J-Measure (J)	$\max \left(P(A, B) \log \left(\frac{P(B A)}{P(B)} \right) + P(\bar{A}, \bar{B}) \log \left(\frac{P(\bar{B} \bar{A})}{P(\bar{B})} \right), \right.$ $\left. P(A, B) \log \left(\frac{P(A B)}{P(A)} \right) + P(\bar{A}, \bar{B}) \log \left(\frac{P(\bar{A} \bar{B})}{P(\bar{A})} \right) \right)$
9	Gini index (G)	$\max (P(A)[P(B A)^2 + P(\bar{B} A)^2] + P(\bar{A})[P(B \bar{A})^2 + P(\bar{B} \bar{A})^2])$

Comparing Different Measures

10 examples of contingency tables:

Example	f_{11}	f_{10}	f_{01}	f_{00}
E1	8123	83	424	1370
E2	8330	2	622	1046
E3	9481	94	127	298
E4	3954	3080	5	2961
E5	2886	1363	1320	4431
E6	1500	2000	500	6000
E7	4000	2000	1000	3000
E8	4000	2000	2000	2000
E9	1720	7121	5	1154
E10	61	2483	4	7452

Rankings of contingency tables using various measures:

#	ϕ	λ	α	Q	Y	κ	M	J	G	s	c	L	V	I	IS	PS	F	AV	S	ζ	K
E1	1	1	3	3	3	1	2	2	1	3	5	5	4	6	2	2	4	6	1	2	5
E2	2	2	1	1	1	2	1	3	2	2	1	1	1	8	3	5	1	8	2	3	6
E3	3	3	4	4	4	3	3	8	7	1	4	4	6	10	1	8	6	10	3	1	10
E4	4	7	2	2	2	5	4	1	3	6	2	2	2	4	4	1	2	3	4	5	1
E5	5	4	8	8	8	4	7	5	4	7	9	9	9	3	6	3	9	4	5	6	3
E6	6	6	7	7	7	7	6	4	6	9	8	8	7	2	8	6	7	2	7	8	2
E7	7	5	9	9	9	6	8	6	5	4	7	7	8	5	5	4	8	5	6	4	4
E8	8	9	10	10	10	8	10	10	8	4	10	10	10	9	7	7	10	9	8	7	9
E9	9	9	5	5	5	9	9	7	9	8	3	3	3	7	9	9	3	7	9	9	8
E10	10	8	6	6	6	10	5	9	10	10	6	6	5	1	10	10	5	1	10	10	7

02/14/2018 数据挖掘导论，第 2 版 82

可变置换下的性质

B B

问答

A r s

A A

B p r

B q s

$M(A, B) = M(B, A)$ 吗?

对称度量:

◆支撑、提升、集体力量、余弦、雅克卡等不对称措施:

◆信心、信念、拉普拉斯、测量等

02/14/2018 数据挖掘导论，第 2 版 83

行/列缩放下的属性

女性男性

高 2 3 5

低 1 4 5

3 7 10

女性男性

高 4 30 34

低 2 40 42

6 70 76

性别示例(Mosteller, 1968):

Mosteller:

潜在关联应该独立于样本中男女学生的相对数量

2x 10x

Property under Inversion Operation

	A	B	C	D	E	F
Transaction 1	1	0	0	1	0	0
■	0	0	1	1	1	0
■	0	0	1	1	1	0
■	0	1	1	1	1	0
■	0	0	1	0	1	1
■	0	0	1	1	1	0
■	0	0	1	1	1	0
■	0	0	1	1	1	0
■	0	0	1	1	1	0
Transaction N	1	0	0	1	0	0

(a) (b) (c)

例如: ϕ 系数

• ϕ 系数类似于连续变量的相关系数

Y Y

X 60 10 70

X 10 20 30

70 30 100

Y Y

X 20 10 30

X 10 60 70

30 70 100

0.5238

0.7 0.3 0.7 0.3

0.6 0.7 0.7

=

$\times \times \times$

$\hat{\phi}$

两个表的 ϕ 系数相同

0.5238

0.7 0.3 0.7 0.3

0.2 0.3 0.3

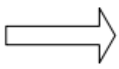
=

$\times \times \times$

$\hat{\phi}$

Property under Null Addition

	B	\bar{B}
A	p	q
\bar{A}	r	s



	B	\bar{B}
A	p	q
\bar{A}	r	s + k

Different Measures have Different Properties

Symbol	Measure	Inversion	Null Addition	Scaling
ϕ	ϕ -coefficient	Yes	No	No
α	odds ratio	Yes	No	Yes
κ	Cohen's	Yes	No	No
I	Interest	No	No	No
IS	Cosine	No	Yes	No
PS	Piatetsky-Shapiro's	Yes	No	No
S	Collective strength	Yes	No	No
ζ	Jaccard	No	Yes	No
h	All-confidence	No	No	No
s	Support	No	No	No

Simpson's Paradox

Buy HDTV	Buy Exercise Machine		
	Yes	No	
Yes	99	81	180
No	54	66	120
	153	147	300

$$c(\{\text{HDTV} = \text{Yes}\} \rightarrow \{\text{Exercise Machine} = \text{Yes}\}) = 99/180 = 55\%$$

$$c(\{\text{HDTV} = \text{No}\} \rightarrow \{\text{Exercise Machine} = \text{Yes}\}) = 54/120 = 45\%$$

=> Customers who buy HDTV are more likely to buy exercise machines

02/14/2018

Introduction to Data Mining, 2nd Edition

88

Simpson's Paradox

Customer Group	Buy HDTV	Buy Exercise Machine		Total
		Yes	No	
College Students	Yes	1	9	10
	No	4	30	34
Working Adult	Yes	98	72	170
	No	50	36	86

College students:

$$c(\{\text{HDTV} = \text{Yes}\} \rightarrow \{\text{Exercise Machine} = \text{Yes}\}) = 1/10 = 10\%$$

$$c(\{\text{HDTV} = \text{No}\} \rightarrow \{\text{Exercise Machine} = \text{Yes}\}) = 4/34 = 11.8\%$$

Working adults:

$$c(\{\text{HDTV} = \text{Yes}\} \rightarrow \{\text{Exercise Machine} = \text{Yes}\}) = 98/170 = 57.7\%$$

$$c(\{\text{HDTV} = \text{No}\} \rightarrow \{\text{Exercise Machine} = \text{Yes}\}) = 50/86 = 58.1\%$$

02/14/2018

Introduction to Data Mining, 2nd Edition

89

02/14/2018 数据挖掘导论，第2版 90

辛普森悖论

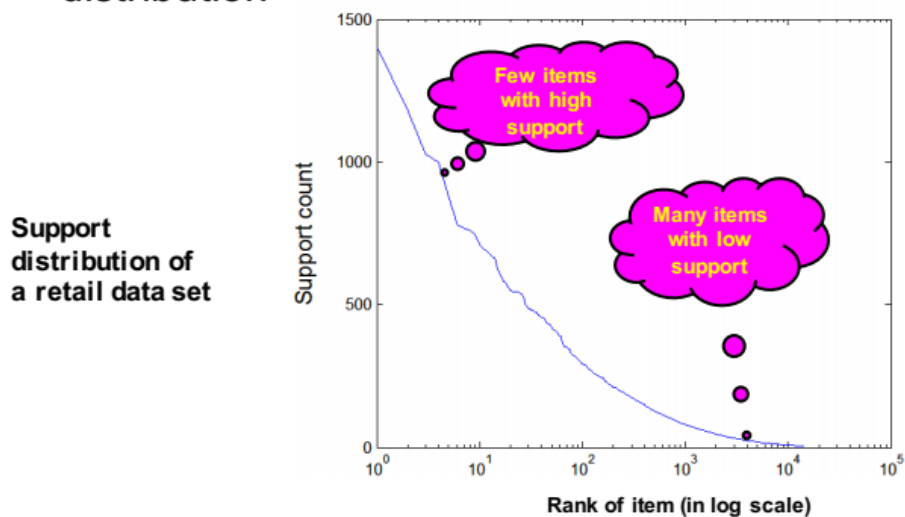
●数据中观察到的关系可能会受到其他混杂因素(隐藏变量)的影响

隐藏的变量可能会导致观察到的关系消失或反向！

- 需要适当的分层，以避免产生虚假模式

Effect of Support Distribution on Association Mining

- Many real data sets have skewed support distribution



02/14/2018

Introduction to Data Mining, 2nd Edition

91

02/14/2018 数据挖掘导论，第2版 92

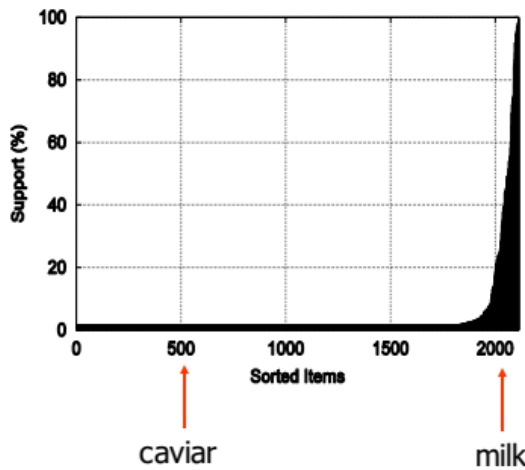
支持分配的效果

- 难以设置适当的最小阈值

如果 minsup 太高，我们可能会错过涉及有趣稀有物品的项目集(例如{鱼子酱，伏特加})

如果 minsup 太低，计算开销很大，并且项目集的数量非常大

Cross-Support Patterns



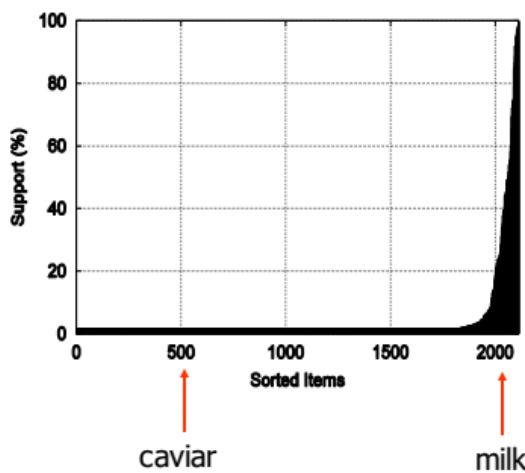
A cross-support pattern involves items with varying degree of support

- Example: {caviar,milk}

How to avoid such patterns?

02/14/2018 数据挖掘导论，第2版 94
交叉支持的度量

Confidence and Cross-Support Patterns



Observation:

$\text{conf}(\text{caviar} \rightarrow \text{milk})$ is very high

but

$\text{conf}(\text{milk} \rightarrow \text{caviar})$ is very low

Therefore,

$\min(\text{conf}(\text{caviar} \rightarrow \text{milk}), \text{conf}(\text{milk} \rightarrow \text{caviar}))$

is also very low

<p>02/14/2018 数据挖掘导论，第 2 版 96</p> <p>身份</p> <ul style="list-style-type: none"> ●为了避免项目具有非常不同的支持的模式，为项目集定义一个新的评估度量 <p>被称为信任或完全信任</p>
<p>02/14/2018 数据挖掘导论，第 2 版 97</p> <p>确认…</p> <p>因此，为了找到最低置信度规则，我们需要找到支持度最高的 $X1 = H(I)$</p> <p>日本 KL 高锰、高钼、…、高锰</p>
<p>02/14/2018 数据挖掘导论，第 2 版 98</p> <p>交叉支持和冲突</p> <ul style="list-style-type: none"> ●通过支持物的锑石性质 ●因此，我们可以推导出项目集的冲突和交叉支持之间的关系 <p>$\leq JIU$ 氢(锰)氢(钼)…氢(锰)</p> <p>日本 KL 高锰、高钼、…、高锰</p>
<p>02/14/2018 数据挖掘导论，第 2 版 99</p> <p>交叉支持和冲突…</p> <ul style="list-style-type: none"> ●请注意 ●任何满足给定验证阈值 h_c 的项目集都被称为超验证●验证可以用来代替支持或与支持结合使用
<p>02/14/2018 数据挖掘导论，第 2 版 100</p> <p>超甘草的性质</p> <ul style="list-style-type: none"> ●超流是项集，但不一定是频繁项集 <p>有利于发现低支持模式</p> <ul style="list-style-type: none"> ●氢是反质子 ●可以根据冲突定义封闭的和最大的超冲突
<p>02/14/2018 数据挖掘导论，第 2 版 101</p> <p>超液体的性质…</p> <ul style="list-style-type: none"> ●超液体具有高亲和性 <p>将单个项目视为稀疏的二进制向量。冲突给我们提供了关于它们成对的信息</p> <p>雅克卡和余弦相似性</p> <p>由雅克卡和余弦测量的非常相似的项目组成的具有高可信度的超液体</p> <ul style="list-style-type: none"> ●在一个超液体中的项目不能有很大不同的支持 <p>允许更有效的修剪</p>
<p>02/14/2018 数据挖掘导论，第 2 版 102</p> <p>超溶液的应用实例</p> <ul style="list-style-type: none"> ●超液体用于寻找强相干的物品组 <p>文档中一起出现的单词</p> <p>蛋白质相互作用网络中的蛋白质</p> <p>在右图中，生物过程的基因本体层次结构显示，在高体液(PRE2, …, SCL1)中识别的蛋白质执行相同</p>

的功能，并参与相同的生物过程