中国科学技术大学
University of Science and Technology of China

数据挖掘与数据仓库

第六讲 模式挖掘（高级）

讲到15页

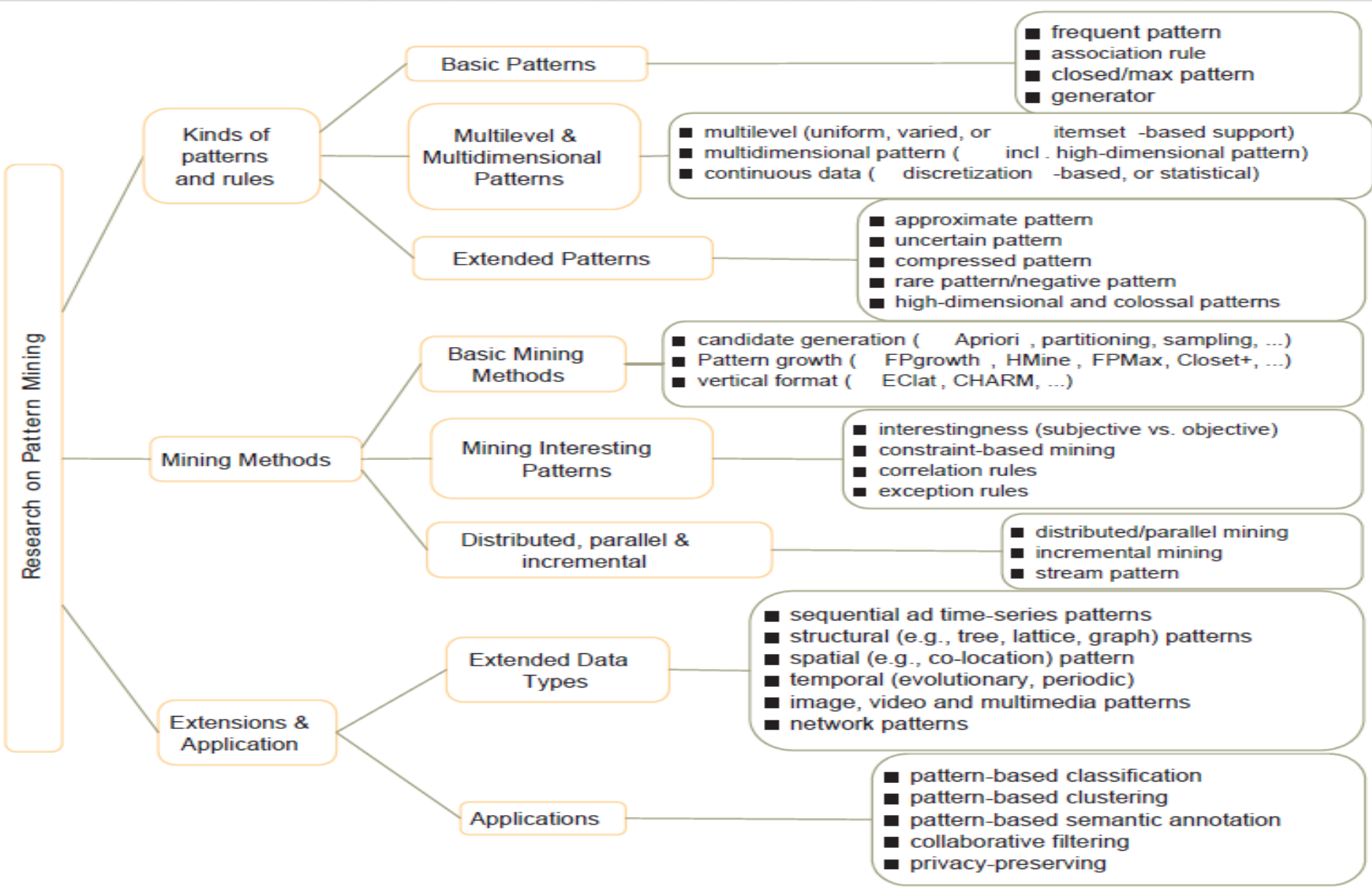
September 25, 2020

Outline

1. 模式挖掘线路图
2. 多层、多维空间中的模式挖掘
3. 基于约束的频繁模式挖掘
4. 挖掘高维数据和巨型模式
5. 挖掘压缩或近似模式

The Road Map

中国科学技术大学
University of Science and Technology of China



多层关联规则

Table 7.1: Task-relevant data, D.

TID	*Items Purchased*
T100	Apple-17"-MacBook-Pro Notebook, HP-Photosmart-Pro-b9180
T200	Microsoft-Office-Professional-2010, Microsoft-Wireless-Optical-Mouse-5000
T300	Logitech-VX-Nano Cordless Laser Mouse, Fellowes-GEL-Wrist-Rest
T400	Dell-Studio-XPS-16-Notebook, Canon-PowerShot-SD1400
T500	Lenovo-ThinkPad-X200 Tablet PC, Symantec-Norton-Antivirus-2010
...	...

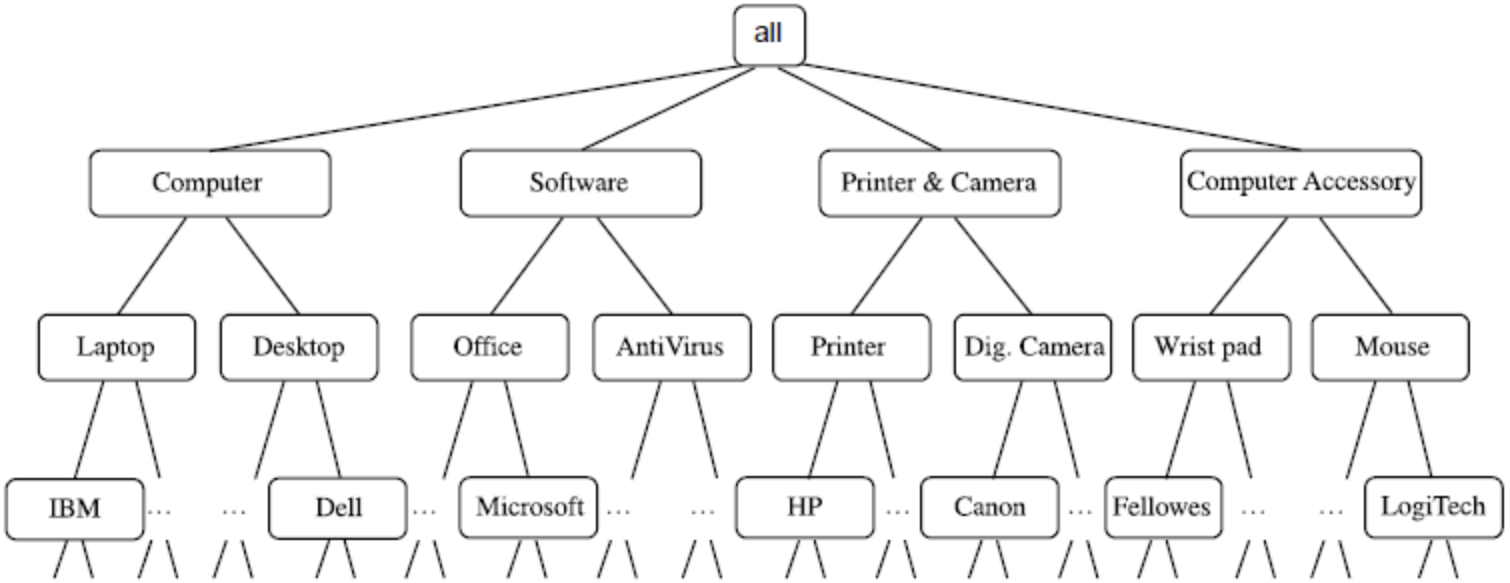


Figure 7.2: A concept hierarchy for *AllElectronics* computer items.

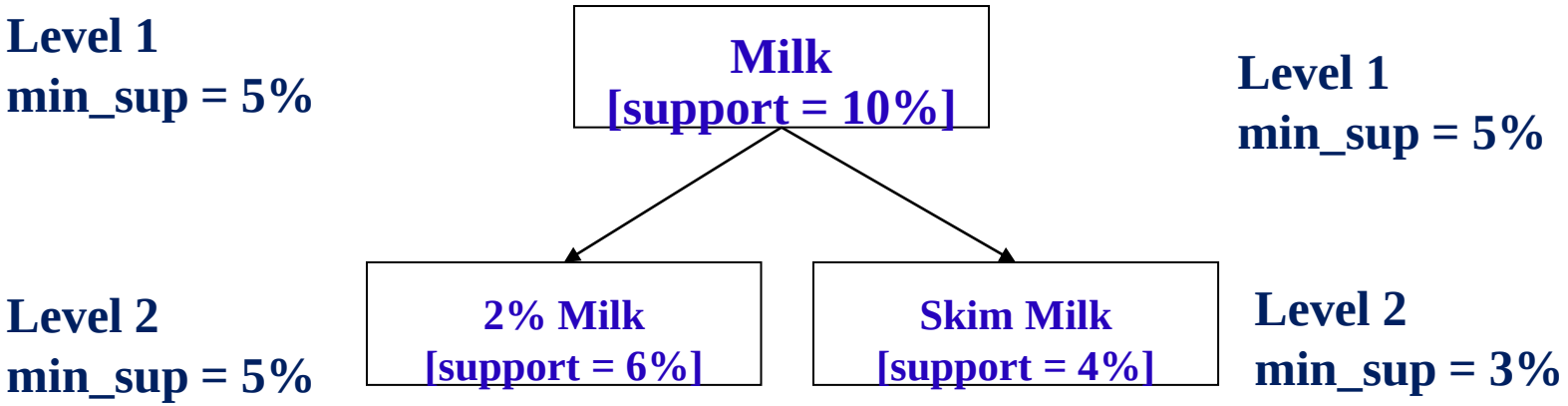
寻找哪种类型(层次)的模式?

多层关联中的支持度

1. 每一层的支持度都相同：低层会有很多模式被遗漏？高层模式是常识？
2. 较低层次采用较小支持度：

uniform support

reduced support



多层关联中的支持度

3. 分组/可变支持度：专家给出/先验知识，
   例如：{diamond, watch, camera}: 0.05%; {bread, milk}: 5%; ...

   非一致的支持度

多层关联规则的冗余

1. 两个不同层次间的关联规则可能存在冗余，其中一个关联规则被另一个'蕴含'，例如：
   - milk ⇒ wheat bread [support = 8%, confidence = 70%]
   - 2% milk ⇒ wheat bread [support = 2%, confidence = 72%]
   - 2% milk是milk的一部分/样本，期望其特点和milk一致，70%和72%很接近!
   - 第二条规则一般性不如前一条，删除第二条规则
   - 更高层的抽象更具一般性，但是也许太一般，变成了常识

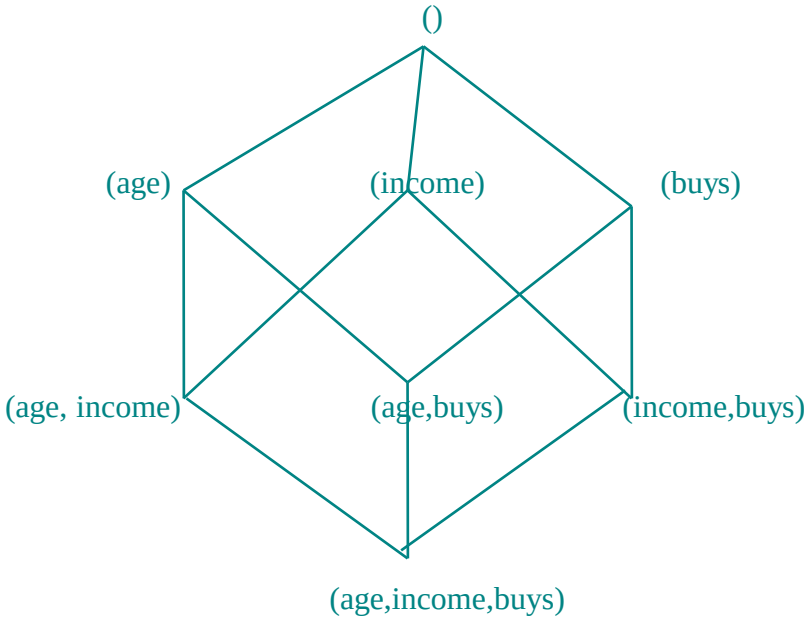
多维关联规则

1. 单维： 购买记录
   - buys(X, “milk”) ⇒ buys(X, “bread”)
2. 多维： 购买记录，顾客信息
   - age(X,”19-25”) ∧ occupation(X,“student”) ⇒ buys(X, “coke”)
   - age(X,”19-25”) ∧ buys(X, “popcorn”) ⇒ buys(X, “coke”)

多维关联：数量属性

1. 关联规则：标称属性
2. 数值型数据呢？多维关联涉及的属性类型可能是多样的
   - 职业/品牌/颜色　VS　年龄/收入/价格
   - 数值型数据种类多，每一种出现次数可能极少

利用数据立方体



Measure/度量:

支持度

存在问题:

计算量/物化

1. 利用数据立方体/概念分层来挖掘量化关联
2. 思考：我们为什么要从不同角度/粒度来观察数据?

多维关联：聚类方法

1. 自顶向下：每个数值属性独立聚类，每个类视为一个标称属性，重建交易数据集；然后使用标准的关联规则挖掘算法
2. 自底向上：在高维空间先聚类，满足最小支持度阈值的类就是频繁模式

多维关联：找异常

1. 发现异常：子集的行为显著区别于其补集
   - (Sex = female) => Wage: mean=$7/hr (overall mean = $9)
   - 需要进行统计检验：z-test
2. 规则的左边是某个集合的子集
   - (Sex = female) ^ (South = yes) => mean wage = $6.3/hr
   - 规则类型：‘标称=>数量’ 规则，‘数量=>数量’规则，例如：Education in [14-18] (yrs) => mean wage = $11.64/hr
3. 待研究问题：关联规则左边多个维度有效的生成方法，即子集的有效构造方法

稀有模式

1. 稀有模式：支持度极低，但是'有趣'

- 买劳力士钻石手表，稀少事件，但是令人感兴趣
- 分组/特殊物品的价值定义不同的支持度阈值
- 例如：将单位价值超过3000元的商品，其支持度设置为0.005%

负模式

1. 负相关：独自频繁的项，却几乎不同时出现
   - 福特征服者（SUV）和丰田普锐斯（混动车）不会被一个人买走
   - 经典可口可乐和无糖可乐基本不会被人同时买
   - sup(X U Y) < sup (X) * sup(Y)，X和Y负相关
2. 问题：缝纫店销售针A和B，只有一个交易同时销售了A和B各100
   - 若共有200个交易记录，则s(A U B) = 0.005, s(A) * s(B) = 0.25, s(A U B) < s(A) * s(B)
   - 若有10^5个交易记录，则s(A U B) = $1/10^5$, s(A) * s(B) = $1/10^3$ * $1/10^3$, s(A U B) > s(A) * s(B)
   - 零事务/ Null transactions，零不变性

负模式：Kulczynski

1. 零不变度量
   - (P(X|Y) + P(Y|X))/2 < ϵ, 其中 ϵ 负模式阈值, X和Y负相关
2. 问题：缝纫店销售针A和B
   - 不管其它事务的数量是多少，我们有
   - ϵ = 0.05，(P(A|B) + P(B|A))/2 = (0.01 + 0.01)/2 < ϵ
   - 课本例子7.4~7.6：包括A和B的事务只有一条，只出现了一次，交易中的数量100视为100次出现, 其它交易事务中没有A或B的出现

基于约束的频繁模式

1. 通常挖掘出来的规则太多，大多数是“无趣的”，能否预先定义一些条件，去掉这些“无趣的”，既得到更好的结果又有更快的挖掘速度?
   - 交互式的过程，用户人工来指导挖掘
   - 基于约束的频繁模式挖掘：用户提供约束，找到满足约束的所有频繁模式

约束是什么?

约束是什么？

1. 知识型约束
   - 待挖掘的知识类型，如关联、相关、分类或聚类
2. 数据约束
   - 指定任务相关的数据集，如在京东今年的销售记录中寻找同时被购买的产品对/组
3. 维/层次约束
   - 指定任务使用的数据维或概念分层

约束是什么？

4. 规则约束

- 指定要挖掘的规则形式或条件
- 如：规则左边必须是两个维
- 如：一种廉价商品的销售会促进另一种昂贵商品的销售

5. 兴趣度约束

- 指定最小支持度，执行度和相关性的计算方法或值
- 如：强规则满足 $min\_support \geq 3\%$, $min\_confidence \geq 60\%$

元规则指导的挖掘

1. 找顾客的两个特点，这些顾客会购买ipad
   - P1(X, Y) ^ P2(X, W) => buys(X, “iPad”)
   - X表示顾客，Y和W表示某两个顾客的属性的取值
   - 元规则就是描述上述规则的“指定形式”
   - age(X, “15-25”) ^ profession(X, “student”) => buys(X, “iPad”)
2. 元规则的一般形
   - $P_1 \wedge P_2 \wedge \ldots \wedge P_l \Rightarrow Q_1 \wedge Q_2 \wedge \ldots \wedge Q_r$
3. 挖掘方法/过程
   - 首先找到频繁项集$L_{(l+r)}$
   - Push约束到挖掘过程中
   - 利用兴趣度，相关性，置信度等

模式空间剪枝

TDB (min_sup=2)

TID	Transaction
10	a, b, c, d, f
20	b, c, d, f, g, h
30	a, c, d, e, f
40	c, e, f, g

Item	Profit	Price
a	40	15
b	0	23
c	-20	41
d	10	22
e	-30	10
f	30	3
g	20	1
h	-10	20

1. 约束：反单调性/ *anti-monotonicity*

- sum(s.price)<=30
- range(s.profit)<=15
- 支持度
- 若当前项集s违背了约束，则可以将当前项集删除，不再扩展；若s没有违背约束，则考察事务/元组中其它频繁项/集的价格与s的价格和，超过则可以删除元组
- 约束具有反单调性是指：约束被某个项集满足，则这个项集的任何子集也满足该约束

模式空间剪枝

TDB (min_sup=2)

TID	Transaction
10	a, b, c, d, f
20	b, c, d, f, g, h
30	a, c, d, e, f
40	c, e, f, g

Item	Profit	Price
a	40	15
b	0	23
c	-20	41
d	10	22
e	-30	10
f	30	3
g	20	1
h	-10	20

2. 约束：单调性/*monotonicity*

- sum(s.price)>=30
- ab价格和超过30，所以任何ab的超集都满足上述条件，因此不需要再ab的超集上进行该约束的进一步检查
- 约束具有单调性是指：约束被某个项集满足，则这个项集的任何超集也满足该约束
- 单调性约束对于剪枝作用有限

模式空间剪枝

TDB (min_sup=2)

TID	Transaction
10	a, b, c, d, f
20	b, c, d, f, g, h
30	a, c, d, e, f
40	c, e, f, g

Item	Profit	Price
a	40	5
b	0	12
c	-20	5
d	10	2
e	-30	10
f	30	3
g	20	1
h	-10	20

3. 约束：数据反单调性/ *data anti-monotonicity*

- sum(s.price)>=30
- cf的的价格和是8，T30中的其它频繁项的价格和相加不可能超过30，故T30可以删除
- 约束具有数据反单调性是指：一个模式的某个约束不能被一个交易事务满足，那么模式的超集也不能被满足，可以删掉该事务数据

模式空间剪枝

TDB (min_sup=2)

TID	Transaction
10	a, b, c, d, f
20	b, c, d, f, g, h
30	a, c, d, e, f
40	c, e, f, g

Item	Profit	Price
a	40	15
b	0	23
c	-20	41
d	10	22
e	-30	10
f	30	3
g	20	1
h	-10	20

4. 约束：简洁性

- min(i.price)>=30
- 将项的集合约简，只包括价格大于等于30的项
- 约束的简洁性是指可以在统计支持度之前就约简项集

模式空间剪枝

TDB (min_sup=2)

TID	Transaction
10	a, b, c, d, f
20	b, c, d, f, g, h
30	a, c, d, e, f
40	c, e, f, g

Item	Profit	Price
a	40	15
b	0	23
c	-20	41
d	10	22
e	-30	10
f	30	3
g	20	1
h	-10	20

5. 约束：可转变的

- avg(s.price)<=20 是不是单调的，也不是不反单调的
- 若把事务中的项按price增序排序，添加到项集中去时，按该序添加，则上述约束被转变为了反单调的
- 约束是可转变的，是指通过某种对项/事务的操作，可以把约束变成单调的或反单调的

可转变的约束

Constraint	Convertible anti-monotone	Convertible monotone	Strongly convertible
avg(S) ≤ , ≥ v	Yes	Yes	Yes
median(S) ≤ , ≥ v	Yes	Yes	Yes
sum(S) ≤ v (items could be of any value, v ≥ 0)	Yes	No	No
sum(S) ≤ v (items could be of any value, v ≤ 0)	No	Yes	No
sum(S) ≥ v (items could be of any value, v ≥ 0)	No	Yes	No
sum(S) ≥ v (items could be of any value, v ≤ 0)	Yes	No	No
……			

总结：约束

Constraint	Anti-monotone	Monotone	Succinct
$v \in S$	no	yes	yes
$S \supseteq V$	no	yes	yes
$S \subseteq V$	yes	no	yes
$\min(S) \leq v$	no	yes	yes
$\min(S) \geq v$	yes	no	yes
$\max(S) \leq v$	yes	no	yes
$\max(S) \geq v$	no	yes	yes
$\text{count}(S) \leq v$	yes	no	weakly
$\text{count}(S) \geq v$	no	yes	weakly
$\text{sum}(S) \leq v$ ($a \in S, a \geq 0$)	yes	no	no
$\text{sum}(S) \geq v$ ($a \in S, a \geq 0$)	no	yes	no
$\text{range}(S) \leq v$	yes	no	no
$\text{range}(S) \geq v$	no	yes	no
$\text{avg}(S) \theta v, \theta \in \{=, \leq, \geq\}$	convertible	convertible	no
$\text{support}(S) \geq \xi$	yes	no	no
$\text{support}(S) \leq \xi$	no	yes	no

高维数据/巨型模式

1. 属性很多，如何挖掘？

- 垂直数据格式：列多，行少时，如基因表达分析
- 模式融合

2. 挑战

- 如果一个频繁模式长度为n则其任何子集都是频繁模式，不管是Apriori还是FP-tree方法，模式都是增长的，在找到该频繁的长度为n的模式前，要探索指数个小一些的模式！
- 模式融合的想法：先构建完全的长度不超过k（=3)的模式集合，然后直接每次多个小模式合并为大模式，而不是逐步合并越来越大的模式
- 模式融合的问题：不具备完备性，模式的质量等不可靠！

压缩/近似模式

1. 找到模式太多，如何控制？
   - 增大最小支持度：可能获得常识
   - 兴趣度
   - 基于约束的频繁模式
   - Top-k频繁闭模式
2. 压缩表示
   - 一组频繁模式，用一个频繁模式来代表/表示
   - 闭频繁项集/极大频繁项集
   - 聚类的方法，距离计算：

$$D(P_1, P_2) = 1 - \frac{|T(P_1) \cap T(P_2)|}{|T(P_1) \cup T(P_2)|}$$