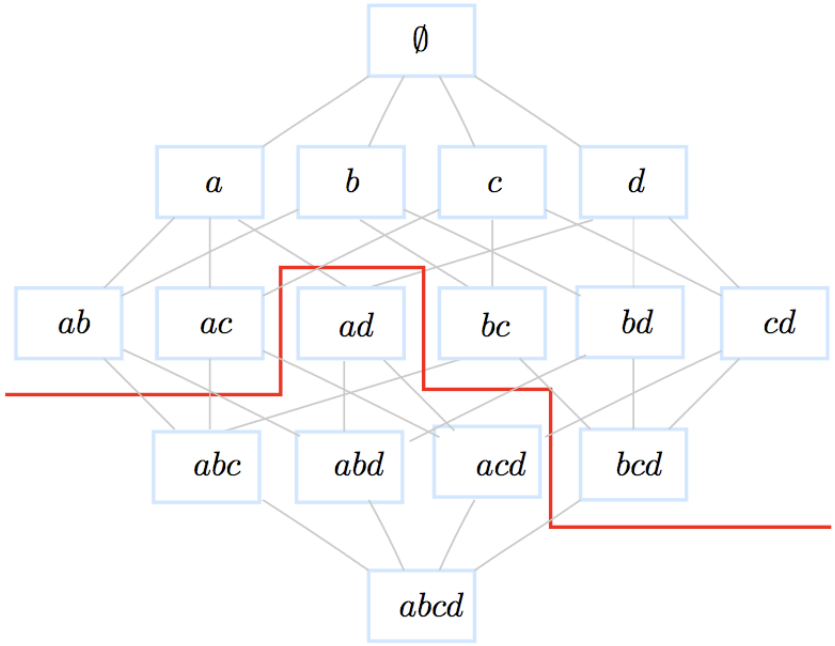
Artificial Intelligence

Search Algorithms in Association Rules



Applications

- Market Basket Analysis: cross-selling (ex. Amazon), product placement, affinity promotion, customer behavior analysis
- Collaborative filtering
- Web organization
- Symptoms-diseases associations
- Supervised classification

Example

item	name
a	coffee
b	milk
c	butter
d	bread

$\mathcal{D}$	
tid	**transaction**
1	$a\ b$
2	$a\ c$
3	$c\ d$
4	$b\ c\ d$
5	$a\ b\ c\ d$

$\mathcal{I}=$

$\mathcal{T}=$

$\mathcal{D}=$

Example

item	name
a	coffee
b	milk
c	butter
d	bread

$\mathcal{D}$	
tid	**transaction**
1	$a\ b$
2	$a\ c$
3	$c\ d$
4	$b\ c\ d$
5	$a\ b\ c\ d$

$$\mathcal{I} = \{a,\ b,\ c,\ d\}$$

$$\mathcal{T} = \{1,\ 2,\ 3,\ 4,\ 5\}$$

$$\mathcal{D} = \{(1,\ ab), (2,\ ac), (3,\ cd), (4,\ bcd), (5,\ abcd)\}$$

E.g., $\{b, c\}$ is a 2-itemset, for writing simplification we will give up the braces and write bc. $\{3, 4, 5\}$ is a tidset similarly let's abandon the braces here too and write 345.

A two-step process

Given a transaction dataset $\mathcal{D}$

1. Mining **frequent** patterns in $\mathcal{D}$

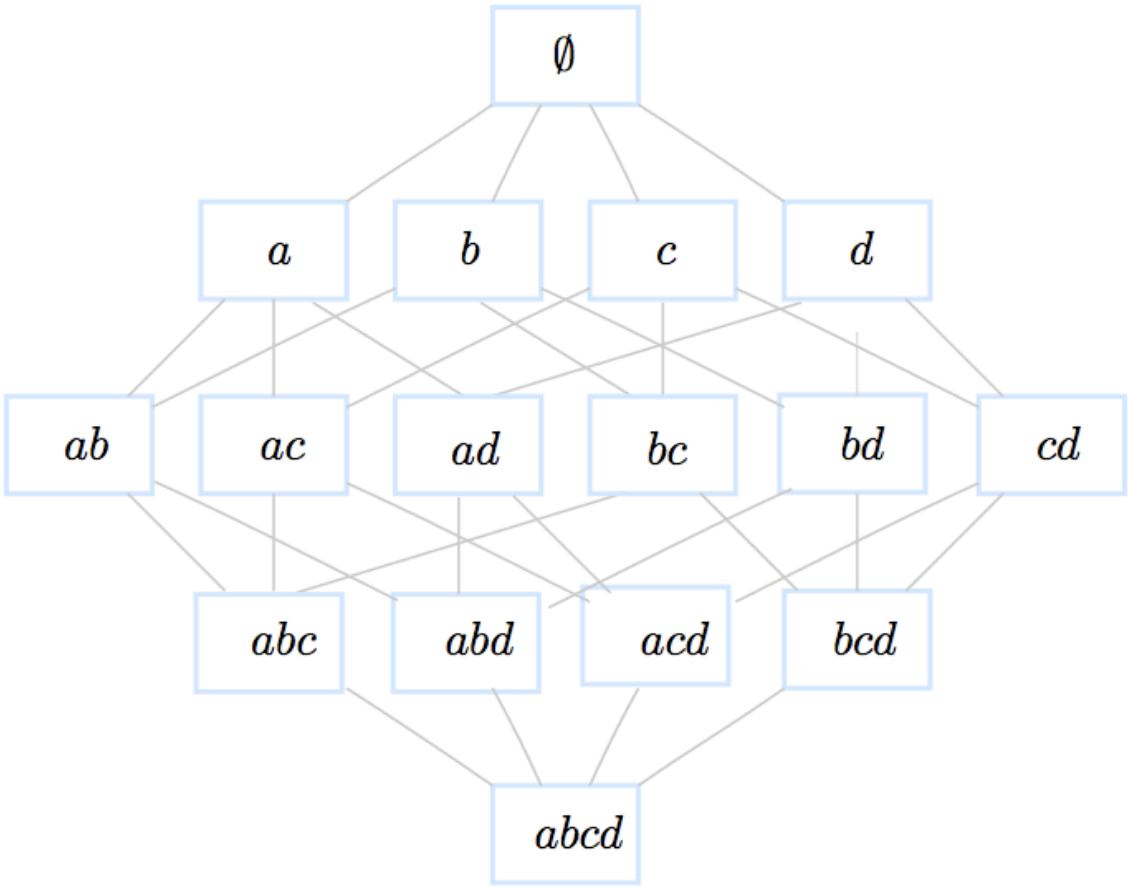
2. Generation of **strong** association rules

Example:

$\{Bread,\ Butter\}$ is a frequent pattern (itemset)

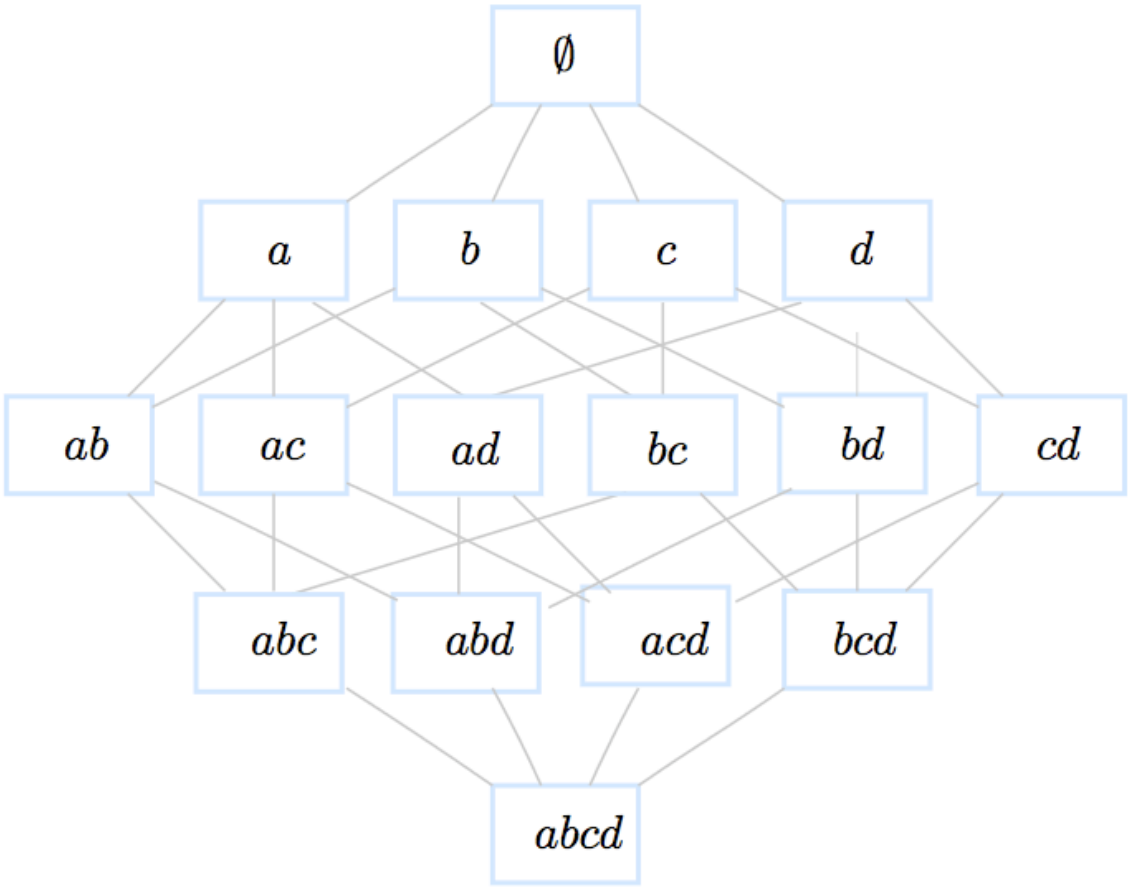
$Bread \rightarrow Butter$ is a strong rule

Example



Lattice of itemsets of size ...

Example



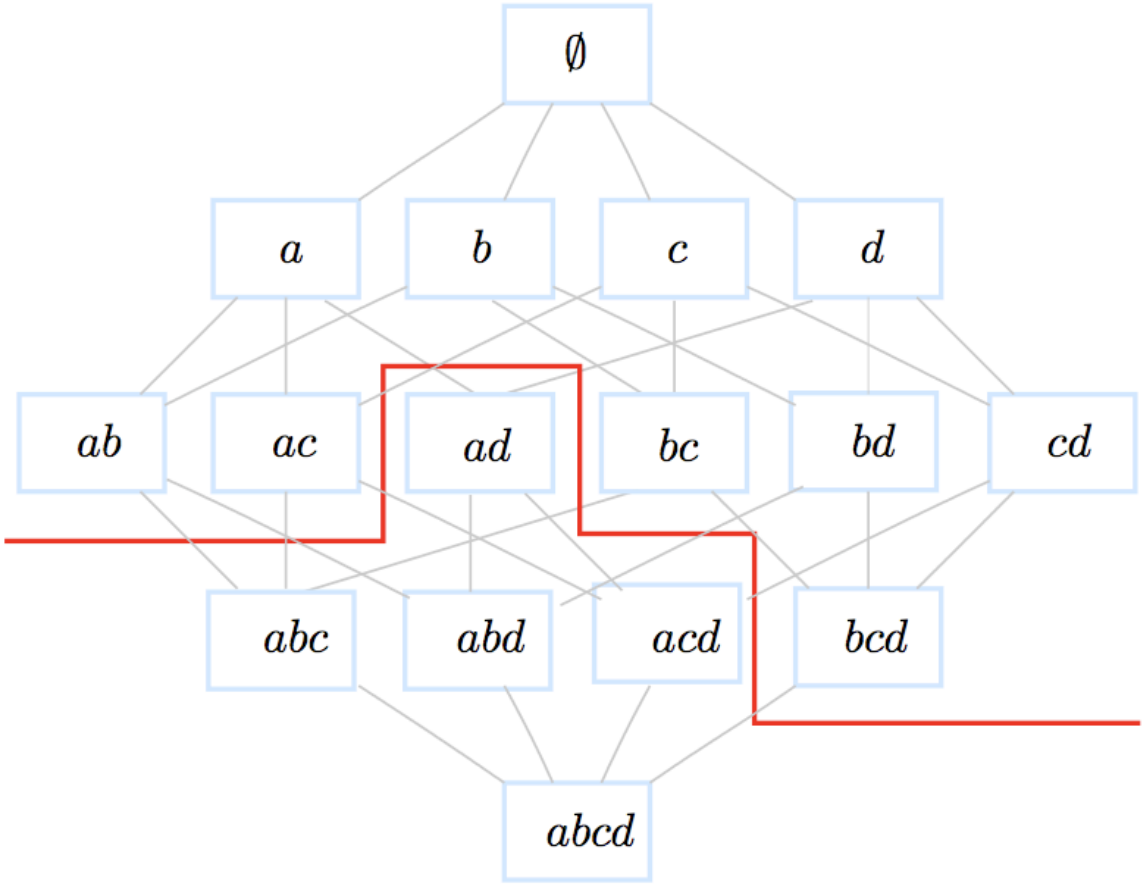
Lattice of itemsets of size $2^{|\mathcal{I}|} = 16$.

Definitions cont'd

- **Frequency**: $freq(X) = |\{(tid, X_{tid}) \in \mathcal{D}/X \subseteq X_{tid}\}| = |t(X)|$
- **Support**: $supp(X) = \frac{|t(X)|}{|\mathcal{D}|}$
- **Frequent itemset**: X is frequent iff supp(X) $\geq$ MinSupp
- **Property (Support downward closure)** : if an itemset is frequent then all its subsets also are frequent.
- **Mining Frequent Itemsets**:

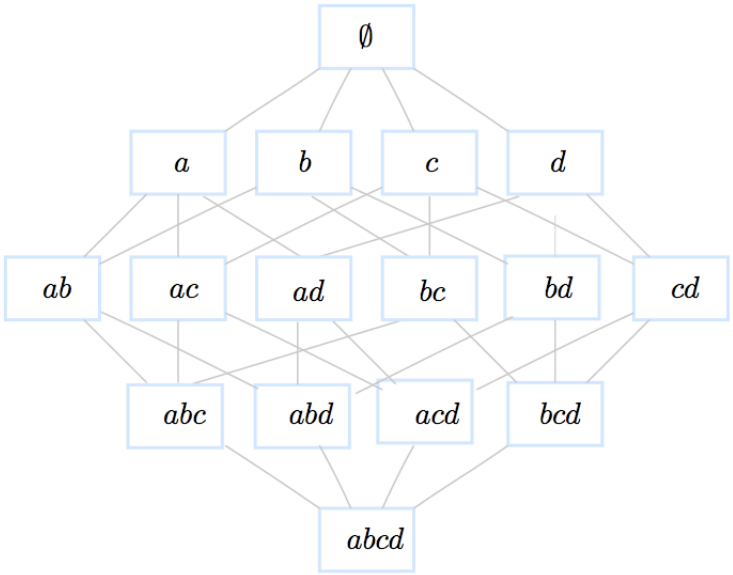
$$\mathcal{F} = \{ X \subseteq \mathcal{I} | \text{ supp(X)} \geq MinSupp\}$$

Example

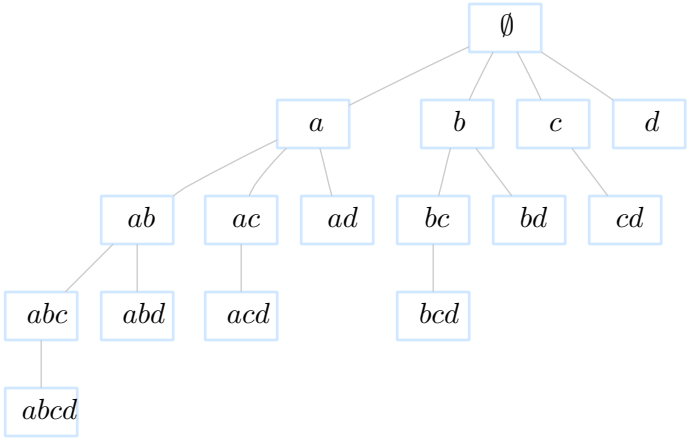


MinSupp=40%

BFS and DFS



Breadth First Search



Depth First Search

Example

Minsupp=2/5 (40%)

$\mathcal{D}$

tid	transaction
1	$a\ b$
2	$a\ c$
3	$c\ d$
4	$b\ c\ d$
5	$a\ b\ c\ d$

$\mathcal{C}_1$

Itemset
a
b
c
d

$\xrightarrow[of\ \mathcal{D}]{Scan}$

$\mathcal{C}_1$

Itemset	Support
a	3/5
b	3/5
c	4/5
d	3/5

$\rightarrow$

$\mathcal{F}_1$

Itemset	Support
a	3/5
b	3/5
c	4/5
d	3/5

$\mathcal{C}_2$

Itemset
ab
ac
ad
bc
bd
cd

$\xrightarrow[of\ \mathcal{D}]{Scan}$

$\mathcal{C}_2$

Itemset	Support
ab	2/5
ac	2/5
ad	1/5
bc	2/5
bd	2/5
cd	3/5

$\rightarrow$

$\mathcal{F}_2$

Itemset	Support
ab	2/5
ac	2/5
bc	2/5
bd	2/5
cd	3/5

$\mathcal{C}_3$

Itemset
abc
bcd

$\xrightarrow[of\ \mathcal{D}]{Scan}$

$\mathcal{C}_3$

Itemset	Support
abc	1/5
bcd	2/5

$\rightarrow$

$\mathcal{F}_3$

Itemset	Support
bcd	2/5

Apriori bottleneck

Characteristics of real-life datasets:

1. Billions of transactions,
2. Tens of thousands of items,
3. Tera-bytes of data.

This leads to:

1. Multiple scans of the dataset residing in the disk (costly I/O operations)
2. A HUGE number of candidates sets.

Example

Minconf=60%

Itemset	Rule#	Rule	Confidence	Strong?
ab	1	$a \rightarrow b$	$2/3 = 66.66\%$	yes
	2	$b \rightarrow a$	$2/3 = 66.66\%$	yes
ac	3	$a \rightarrow c$	$2/3 = 66.66\%$	yes
	4	$c \rightarrow a$	$2/4 = 50.00\%$	no
bc	5	$b \rightarrow c$	$2/3 = 66.66\%$	yes
	6	$c \rightarrow b$	$2/4 = 50.00\%$	no
bd	7	$b \rightarrow d$	$2/3 = 66.66\%$	yes
	8	$d \rightarrow b$	$2/3 = 66.66\%$	yes
cd	9	$c \rightarrow d$	$3/4 = 75.00\%$	yes
	10	$d \rightarrow c$	$3/3 = 100.00\%$	yes

Example

Minconf=60%

Itemset	Rule#	Rule	Confidence	Strong?
bcd	11	$cd \rightarrow b$	$2/3 = 66.66\%$	yes
	12	$bd \rightarrow c$	$2/2 = 100.00\%$	yes
	13	$bc \rightarrow d$	$2/2 = 100.00\%$	yes

Itemset	Rule#	Rule	Confidence	Strong?
bcd	14	$d \rightarrow bc$	$2/3 = 66.66\%$	yes
	15	$c \rightarrow bd$	$2/4 = 50.00\%$	no
	16	$b \rightarrow cd$	$2/3 = 66.66\%$	yes

Probabilistic Interpretation

Brin et al. 97

$$R: \quad A \longrightarrow C$$

- R measures the distribution of A and C in the finite space $\mathcal{D}$.
- The sets A and C are 2 events
- $P(A)$ and $P(C)$ the probabilities that events A and C happen resp. estimated by the the frequency of A and C resp. in $\mathcal{D}$

$$supp(A \rightarrow C) = supp(A \cup C) = P(A \wedge C)$$

$$conf(A \rightarrow C) = P(C|A) = \frac{P(A \wedge C)}{P(A)}$$

Support-Confidence: cons

- Example (Brin et al. 97)

	$coffee$	$\overline{coffee}$	$\sum rows$
tea	20	5	25
$\overline{tea}$	70	5	75
$\sum columns$	90	10	100

$$tea \rightarrow coffee \quad (supp = 20\%, conf = 80\%)$$

Strong rule?

Support-Confidence: cons

- Example (Brin et al. 97)

	$coffee$	$\overline{coffee}$	$\sum rows$
tea	20	5	25
$\overline{tea}$	70	5	75
$\sum columns$	90	10	100

$$tea \rightarrow coffee \quad (supp = 20\%, conf = 80\%)$$

Strong rule? Yes but a misleading one!

$Support(coffee) = 90\%$ is a bias that the confidence cannot detect because it ignores support(coffee).

Other evaluation Measures

- Interest (Piatetsky-Shapiro 91) or Lift (Bayardo et al. 99)

$$Interest(A \to C) = \frac{P(A \wedge C)}{P(A) \times P(C)} = \frac{supp(A \cup C)}{supp(A) \times supp(C)}$$

Interest is between 0 and $+\infty$:

1. If $Interest(\mathcal{R}) = 1$ then A and C are independent;
2. If $Interest(\mathcal{R}) > 1$ then A and C are positively dependent;
3. If $Interest(\mathcal{R}) < 1$ then A and C are negatively dependent.

$$Interest(A \to C) = \frac{conf(A \to C)}{supp(C)} = \frac{conf(C \to A)}{supp(A)}$$

Other evaluation Measures

	$coffee$	$\overline{coffee}$	$\sum rows$
tea	20	5	25
$\overline{tea}$	70	5	75
$\sum columns$	90	10	100

$$Interest(tea \rightarrow coffee) = \frac{P(tea \wedge coffee)}{P(tea) \times P(coffee)} = \frac{0.2}{0.25 * 0.9} = 0.89 < 1$$

	$coffee$	$\overline{coffee}$
tea	0.89	2
$\overline{tea}$	1.03	0.66

Multi-dimensional rules

- One-dimensional rules:

$$buy(x, \text{“}Bread''") \longrightarrow buy(x, \text{“}Butter''")$$

- Multi-dimensional rules:

$$buy(x, \text{“}Pizza''") \wedge age(x, \text{“}Young''") \longrightarrow buy(x, \text{“}Coke''")$$

- Construct k-predicatesets instead of k-itemsets

- How about numerical features?

$$buy(x, \text{“}Pizza''") \wedge age(x, \text{“}18 - 22''") \longrightarrow buy(x, \text{“}Coke''")$$

FP algorithms

According to the strategy to traverse the search space:

- Breadth First Search (ex: Apriori, AprioriTid, Partition, DIC)
- Depth First Search (ex: Eclat, Clique, Depth project)
- Hybrid (ex: AprioriHybrid, Hybrid, Viper, Kdci)
- Pattern growth, i.e. no candidate generation (ex: Fpgrowth, HMine, Cofi)

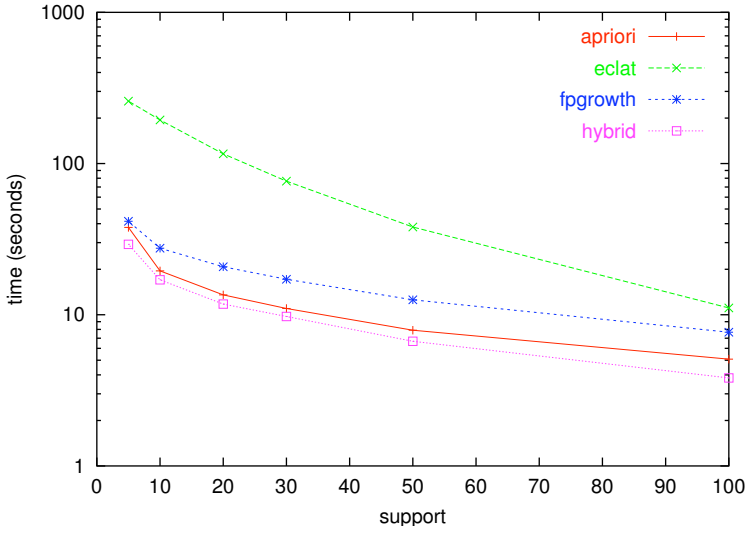
Performance

(Goethals 2004)

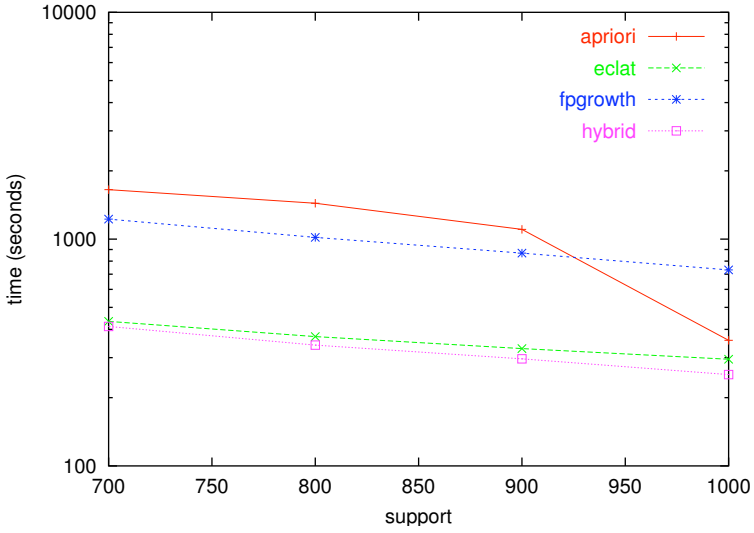
Data set	#Items	#Transactions	$min\|T\|$	$max\|T\|$	$avg\|T\|$
T40I10D100K	942	100 000	4	77	39
mushroom	119	8 124	23	23	23
BMS-Webview-1	497	59 602	1	267	2
basket	13 103	41 373	1	52	9

Table 5: Data Set Characteristics.

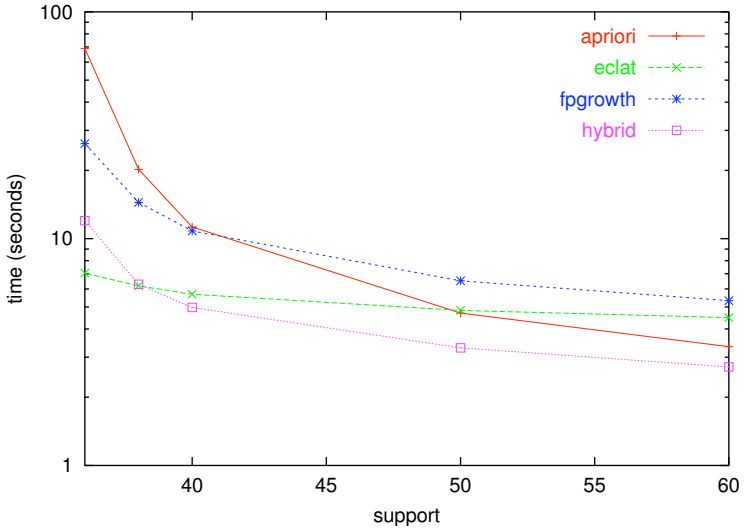
Performance



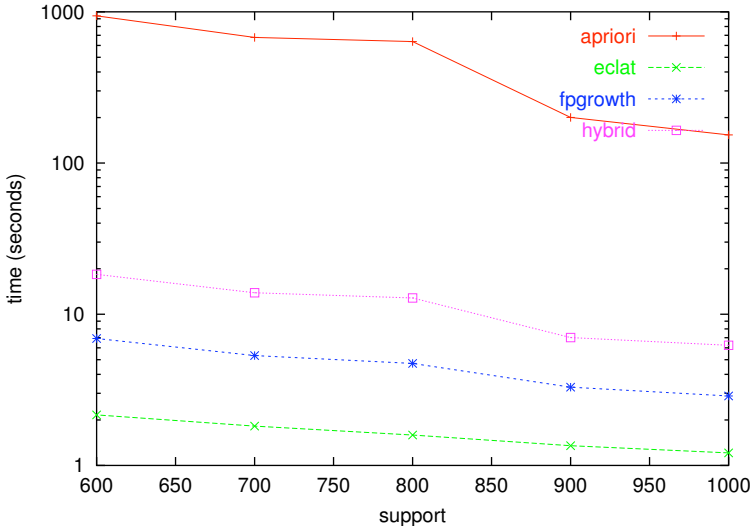
(a) basket



(c) T40I10D100K



(b) BMS-Webview-1



(d) mushroom

Example

$\mathcal{D}$: people			
id	**age**	**married?**	**#cars**
1	23	no	1
2	25	yes	1
3	29	no	0
4	34	yes	2
5	38	yes	2

Examples of frequent itemsets	
itemset	**support**
(age, 20..29)	3
(age, 30..39)	2
(married?, yes)	3
(married?, no)	2
(#cars, 1)	2
(#cars, 2)	2
(age, 30..39),(married?, yes)	2

Examples of rules		
rule	**support**	**confidence**
(age, 30..39) et (married?, yes) $\longrightarrow$ (#cars, 2)	40%	100%
(age, 20..29) $\longrightarrow$ (#cars, 1)	60%	66.6%

Quantitative AR

Question: Mining Quantitative AR is not a simple extension of mining categorical AR. why?

- **Infinite search space:** In Boolean AR, the Ariori property allows to prune the search space efficiently, but we do explore the whole space of hypothesis (lattice of itemsets), which is IMPOSSIBLE for Quantitative AR.

- **The support-confidence tradeoff:** Choosing intervals is quite sensitive to support and confidence.

  - intervals too small, not enough support;
  - intervals too large, not enough confidence.

- What is the difference between supervised and **unsupervised discretization**?

Approaches to mine QARs

- Discretization-based approaches
- Distribution-based approaches
- Optimization-based approaches

Approaches to mine QARs

Discretization-based approaches

- A pre-processing step
- Use equi-depth, equi-width, domain-knowledge
- Lent et al., 1997; Miller and Yang, 1997; Srikant and Agrawal, 1996; Wang et al., 1998
- Discretization combined with clustering or interval merging.
- Problems: univariate, sensitive to outliers, loss of information.

Approaches to mine QARs

Distribution-based approaches

$$Sex = female \rightarrow Height : mean = 168 \wedge Weight : mean = 68$$

- Aumann and Lindell, 1999, Webb 2001.

- Restricted form of rules:
  1. A set of categorical attributes on the left-hand side and several distributions on the right-hand side,

  2. A single discretized numeric attribute on the left-hand side and a single distribution on the right-hand side.

Approaches to mine QARs

Optimization-based approaches

- Numerical attributes are optimized during the mining process
- Fukuda et al., 96, Rastogi and Shim 99, Brin et al. 2003. Techniques inspired from image segmentation. Form of the rules restricted to 1 or 2 numerical attributes.
- Mata et al. 2002 Use genetic algorithms to optimize the support of itemsets with non instantiated intervals.

$$\text{Fitness} = \text{cov} - (\psi * \text{ampl}) - (\omega * \text{mark}) + (\mu * \text{nAtr})$$

Apriori-like algorithm to mine association rules.

Approaches to mine QARs

Optimization-based approaches

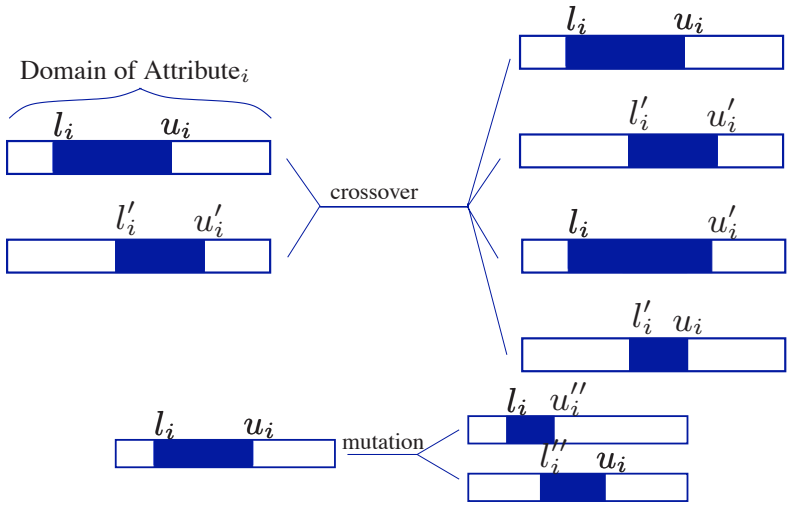
- Ruckert et al. 2004 use half-spaces to mine such rules like:

$$x_1 > 20 \rightarrow 0.5x_3 + 2.3x_6 \geq 100$$

  Cannot handle categorical attributes.

- Salleb et al 2007: QuantMiner Optimize the *Gain* of rules templates using a genetic algorithm.

$$Gain(A \rightarrow B) = Supp(AB) - MinConf * Supp(A)$$



Approaches to mine QARs

Optimization-based approaches: QuantMiner cont'd.

Example UCI Iris dataset:

rectcartouche

$$\texttt{Species=value} \Rightarrow \left\{ \begin{array}{l} \texttt{PW} \in [l_1, u_1] \ \texttt{SW} \in [l_2, u_2] \\ \texttt{PL} \in [l_3, u_3] \ \texttt{SL} \in [l_4, u_4] \end{array} \right\} \begin{array}{l} \texttt{supp\%} \\ \texttt{conf\%} \end{array}$$

$$\texttt{Species=setosa} \Rightarrow \left\{ \begin{array}{l} \texttt{PW} \in [1, 6] \ \texttt{SW} \in [31, 39] \\ \texttt{PL} \in [10, 19] \ \texttt{SL} \in [46, 54] \end{array} \right\} \begin{array}{l} 23\% \\ 70\% \end{array}$$

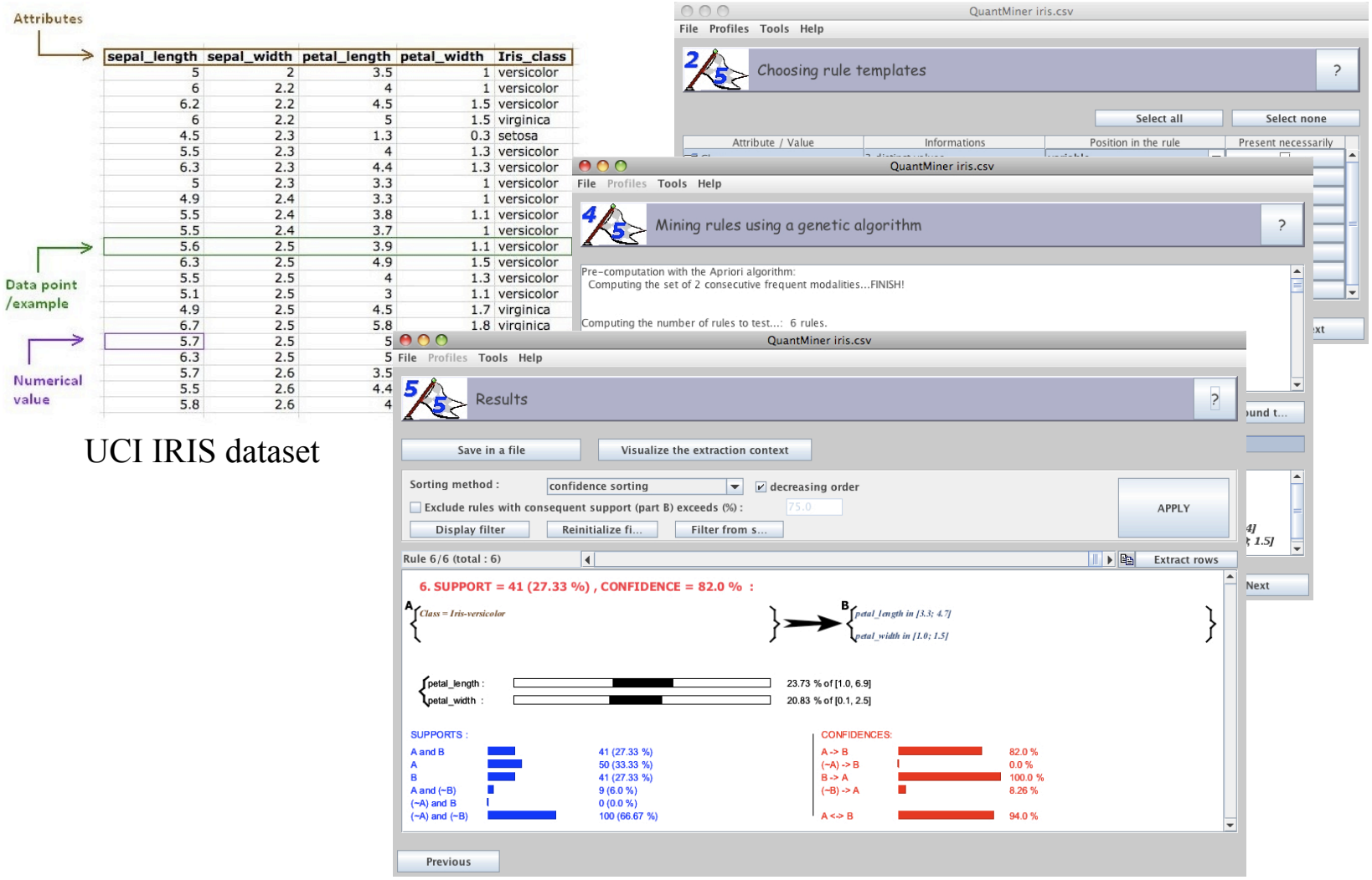
$$\texttt{Species=versicolor} \Rightarrow \left\{ \begin{array}{l} \texttt{PW} \in [10, 15] \ \texttt{SW} \in [22, 30] \\ \texttt{PL} \in [35, 47] \ \texttt{SL} \in [55, 66] \end{array} \right\} \begin{array}{l} 21\% \\ 64\% \end{array}$$

$$\texttt{Species=virginica} \Rightarrow \left\{ \begin{array}{l} \texttt{PW} \in [18, 25] \ \texttt{SW} \in [27, 33] \\ \texttt{PL} \in [48, 60] \ \texttt{SL} \in [58, 72] \end{array} \right\} \begin{array}{l} 20\% \\ 60\% \end{array}$$

QuantMiner

http://quantminer.github.io/QuantMiner/

QuantMiner



UCI IRIS dataset

References

- R. Agrawal, T. Imielinski and A.N. Swami "Mining Association Rules between sets of items in large databases". SIGMOD 1993.
- R. Agrawal, R. Srikant "Fast algorithms for mining association rules" VLDB 1994.
- B. Goethals "Survey on Frequent Pattern Mining" Technical report, Helsinki Institute for Information Technology, 2003.
- S. Brin et al. "Beyond Market Baskets: Generalizing Association Rules to Correlations". SIGMOD 1997.
- R. Agrawal et al. "Mining association rules with item constraints". KDD 1997.
- A. Salleb et al. "QuantMiner: A Genetic Algorithm for Mining Quantitative Association Rules", IJCAI 2007.
- U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. "From Data Mining to Knowledge Discovery: An Overview". In Advances in Knowledge Discovery and Data Mining, 1996.