Data Mining:

2. Assoziationsanalyse

B) Rule Generation

Rule Generation

- Given a frequent itemset L, find all non-empty subsets $f \subset L$ such that $f \rightarrow L - f$ satisfies the minimum confidence requirement
  - If {A,B,C,D} is a frequent itemset, candidate rules are:

$ABC \rightarrow D$,	$ABD \rightarrow C$,	$ACD \rightarrow B$,	$BCD \rightarrow A$,
$A \rightarrow BCD$,	$B \rightarrow ACD$,	$C \rightarrow ABD$,	$D \rightarrow ABC$
$AB \rightarrow CD$,	$AC \rightarrow BD$,	$AD \rightarrow BC$,	$BC \rightarrow AD$,
$BD \rightarrow AC$,	$CD \rightarrow AB$,		

- If $|L| = k$, then there are $2^k - 2$ candidate association rules (ignoring $L \rightarrow \emptyset$ and $\emptyset \rightarrow L$)

Rule Generation

- How to efficiently generate rules from frequent itemsets?
  - In general, confidence does not have an anti-monotone property

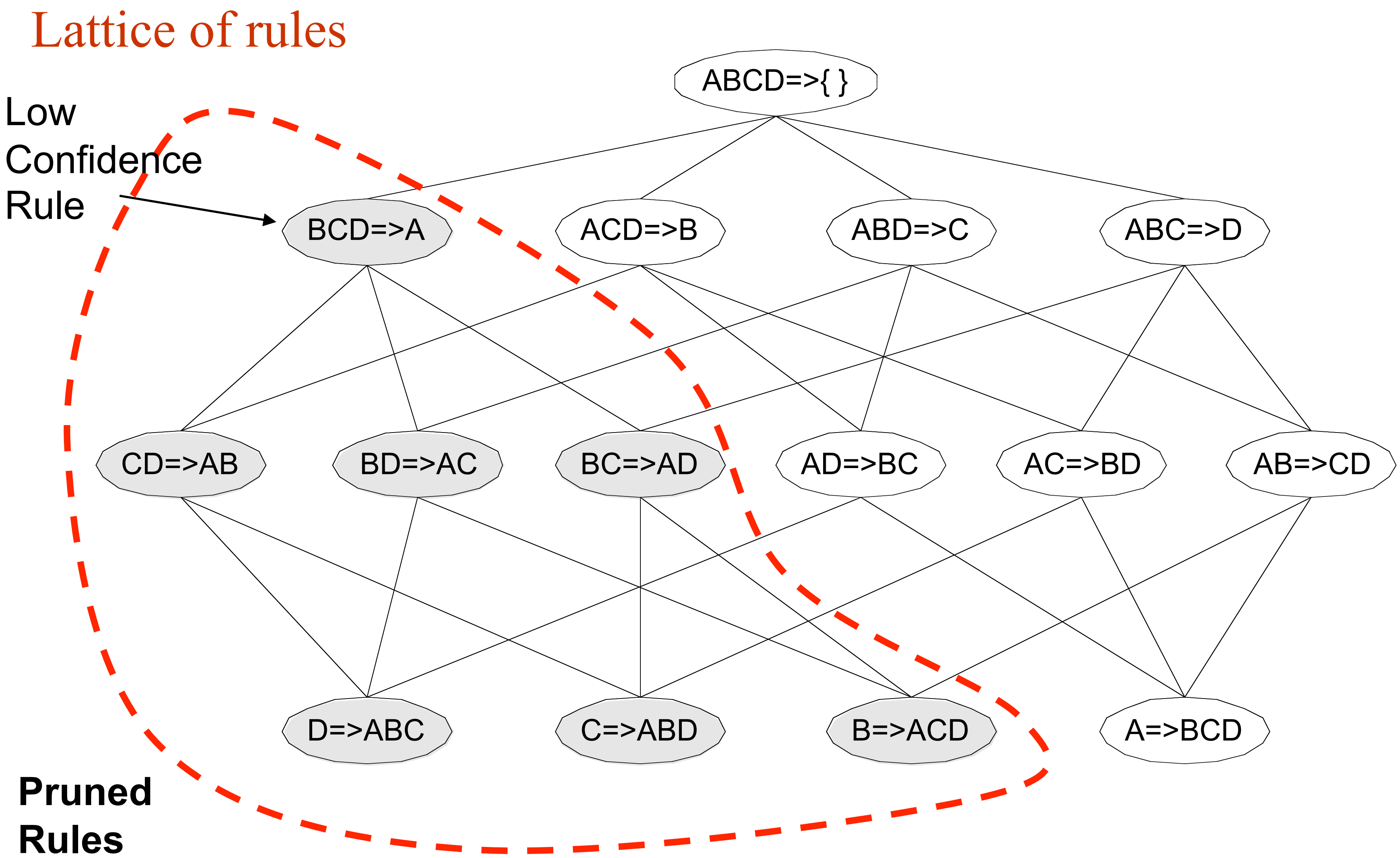
    $c(\text{ABC} \rightarrow \text{D})$ can be larger or smaller than $c(\text{AB} \rightarrow \text{D})$

  - But confidence of rules generated from the same itemset has an anti-monotone property

    e.g., L = {A,B,C,D}:

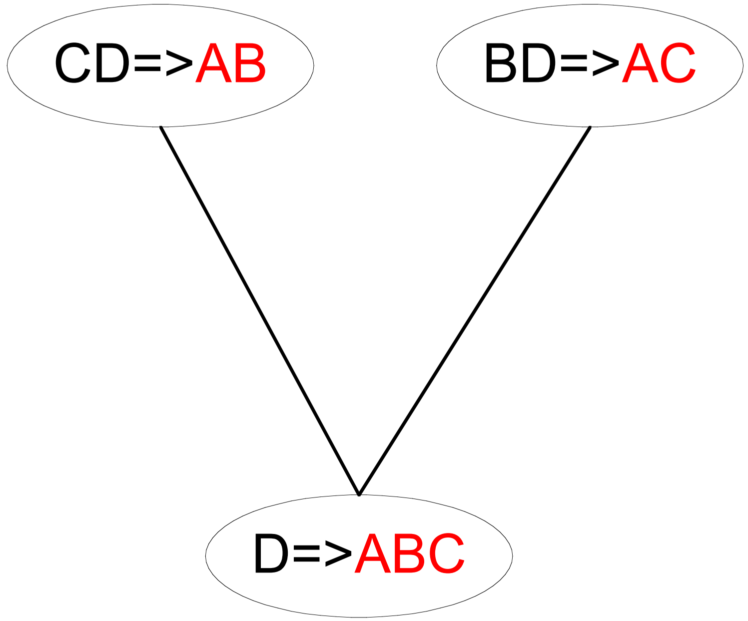
    $$c(\text{ABC} \rightarrow \text{D}) = s(\text{ABCD})/s(\text{ABC})$$
    $$\geq c(\text{AB} \rightarrow \text{CD}) = s(\text{ABCD})/s(\text{AB})$$
    $$\geq c(\text{A} \rightarrow \text{BCD}) = s(\text{ABCD})/s(\text{A})$$

Rule Generation for Apriori Algorithm



Rule Generation for Apriori Algorithm

- From each frequent itemset, here ABCD:
- Candidate rules with consequent length (m+1) are generated by merging two rules [of high confidence] with consequents of length m that share the same (m-1)-prefix
- Join (CD=>AB, BD=>AC) would produce the candidate rule (ABCD-ABC)=D => ABC
- Prune rule D=>ABC if its confidence is not high:* c(D=>ABC) = s(ABCD)/s(D)
- If BCD=>A has no high confid., e.g. CD=>AB is not generated.



*)Support counts can be re-used from itemset generation.

Rule Generation for Apriori Algorithm

Algorithm Rule generation of the *Apriori* algorithm.

for each frequent k-itemset f_k, $k \geq 2$ **do**
  call `ap-genrules`$(f_k, \{\})$
end for

Algorithm Procedure `ap-genrules`$(f_k,\ H_m)$.

$k = \{$size of frequent itemset $f_k.\}$
$m = \{$consequent size of rules in $H_m.\}$
if $k > m + 1$ **then**
  if $m = 0$ **then** $H_1 = \{i \mid i \in f_k\}$ **else** $H_{m+1} = \text{apriori-gen}(H_m)$.
  for each $h \in H_{m+1}$ **do**
    $conf = \sigma(f_k)/\sigma(f_k - h)$.
    if $conf \geq minconf$ **then**
      output the rule $(f_k - h) \longrightarrow h$.
    else
      delete h from H_{m+1}.
    end if
  end for
  call ap-genrules$(f_k, H_{m+1}.)$
end if

Example: Rules from Congress Voting Records

Data is obtained from the 1984 Congressional Voting Records Database, which is available at the UCI machine learning data repository. Each transaction contains information about the party affiliation for a representative along with his or her voting record on 16 key issues. There are 435 transactions and 34 items in the data set.

The *Apriori* algorithm is then applied to the data set with $minsup = 30\%$ and $minconf = 90\%$. Some of the high-confidence rules extracted:

Association Rule	Confidence
{budget resolution = no, MX-missile=no, aid to El Salvador = yes } $\longrightarrow$ {Republican}	91.0%
{budget resolution = yes, MX-missile=yes, aid to El Salvador = no } $\longrightarrow$ {Democrat}	97.5%
{crime = yes, right-to-sue = yes, physician fee freeze = yes} $\longrightarrow$ {Republican}	93.5%
{crime = no, right-to-sue = no, physician fee freeze = no} $\longrightarrow$ {Democrat}	100%

Evaluation of Association Patterns

- Association rule algorithms tend to produce too many rules
  - Many of them are uninteresting or redundant
  - Redundant if {A,B,C} $\rightarrow$ {D} and {A,B} $\rightarrow$ {D} have same support & confidence

- Interestingness measures can be used to prune/rank the derived patterns
  - May be subjective, based on domain knowledge
  - Should be objective, domain-independent, based on statistics

- In the original formulation of association rules, support and confidence are the only objective interestingness measures used

Computing Interestingness Measure

- Given a rule $X \rightarrow Y$, information needed to compute rule interestingness can be obtained from a contingency table

Contingency table for $X \rightarrow Y$

	Y	$\overline{Y}$	
X	f_{11}	f_{10}	f_{1+}
$\overline{X}$	f_{01}	f_{00}	f_{o+}
	f_{+1}	f_{+0}	N

f_{11}: support count of X and Y
f_{10}: support count of X and $\overline{Y}$
f_{01}: support count of $\overline{X}$ and Y
f_{00}: support count of $\overline{X}$ and $\overline{Y}$

Such tables are typically used for binary variables/attributes X,Y. $\overline{X}$ denotes absence of X from a transaction.

Association analysis focuses on asymmetric binary variables (only presence/1 is important).

Used to define various measures

- support = f_{11}/N, confid. = f_{11}/f_{1+}, lift, Gini, J-measure, etc.

Caution with Confidence

	Coffee	$\overline{\text{Coffee}}$	
Tea	150	50	200
$\overline{\text{Tea}}$	650	150	800
	800	200	1000

Consider the association rule: Tea $\rightarrow$ Coffee

Support: $P(\text{Coffee} \wedge \text{Tea}) = f_{11}/N = 150/1000 = 0.15$

Confidence: $P(\text{Coffee}|\text{Tea}) = f_{11}/f_{1+} = 150/200 = 0.75$

but $P(\text{Coffee}) = f_{+1}/N = 0.80$

$\Rightarrow$ Although confidence is high, rule is misleading

Reason: Statistical Independence/Correlation

- Example: Consider a Population of 1000 students
  - 600 students know how to swim (S)
  - 700 students know how to bike (B)
  - 420 students know how to swim and bike (S,B)
  - $P(S \wedge B) = 420/1000 = 0.42$
  - $P(S) x P(B) = 0.6 \times 0.7 = 0.42$
  - $P(S \wedge B) = P(S) \times P(B)$ iff S and B are statistically independent
  - $P(S \wedge B) > P(S) \times P(B)$ iff S and B are positively correlated
  - $P(S \wedge B) < P(S) \times P(B)$ iff S and B are negatively correlated
- Remember: Confidence($S \rightarrow B$) is $P(B|S)=P(S \wedge B)/P(S)$
  - $\ldots = P(S) \times P(B) / P(S) = P(B)$ iff stat. Independent
  - … > / < P(B) iff positively/negatively correlated

Alternative interestingness measures

- **Lift**$(X \rightarrow Y) = \dfrac{c(X \rightarrow Y)}{s(Y)} = \dfrac{P(Y|X)}{P(Y)} = \dfrac{f_{11} N}{f_{1+} f_{+1}}$

- **Interest Factor I**$(X,Y) = \dfrac{s(XY)}{s(X)s(Y)} = \dfrac{P(XY)}{P(X)P(Y)} = \dfrac{f_{11} N}{f_{1+} f_{+1}}$

  – coincide for binary variables
  – =1, if X and Y are statistically independent (P(X,Y)=P(X)P(Y))
  – >1, if X and Y are positively correlated
  – <1, if X and Y are negatively correlated

Confidence vs. Lift/Interest

	Coffee	$\overline{\text{Coffee}}$	
Tea	150	50	200
$\overline{\text{Tea}}$	650	150	800
	800	200	1000

Consider the association rule: Tea → Coffee

Confidence: P(Coffee|Tea) = f_{11}/f_{1+} = 0.75

but P(Coffee) = f_{+1} = 0.80

Lift/Interest: 0.75/0.8 = 0.15/(0.2x0.8)= 0.9375

(< 1, explicitly indicates a negative correlation)

Confidence vs. Lift/Interest (contd.)

	p	$\bar{p}$	
q	880	50	930
$\bar{q}$	50	20	70
	930	70	1000

	r	$\bar{r}$	
s	20	50	70
$\bar{s}$	50	880	930
	70	930	1000

(Think of words occurring together in documents.)

Lift(p,q) = (880*1000)/(930*930) = 1,02

Lift(r,s) = (20*1000)/(70*70) = 4,08

Although p and q occur together in much more cases, the lift measure gets worse.

Here better choice: Confidence

$c(p \rightarrow q)$= 880/930= 94.6% vs. $c(r \rightarrow s)$ = 20/70 = 28.6%

Another interestingness measure

- **Φ-coefficient** (correlation of binary variables):

$$\Phi(X,Y) = \frac{P(X,Y) - P(X)P(Y)}{\sqrt{P(X)[1-P(X)]P(Y)[1-P(Y)]}} = \frac{f_{11}N - f_{1+}f_{+1}}{\mathrm{sqrt}(f_{1+}f_{0+}f_{+1}f_{+0})}$$

  – Ranges from -1 to +1
  – -1 = perfect negative correlation, +1= perfect positive correlation
  – 0 = statistical independence
  – For the coffee/tea-example: -0.0625
  – Not helpful for p/q- vs. r/s-example: $\Phi(p,q) = \Phi(r,s) = 0.232$
  – More suitable for symmetric variables

- Alternative measure for asymmetric variables: IS

There are lots of measures proposed in the literature.

Some measures are good for certain applications, but not for others.

What criteria should we use to determine whether a measure is good or bad?

#	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)	$\frac{\sum_j \max_k P(A_j,B_k)+\sum_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-\max_k P(B_k)}$
3	Odds ratio (α)	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(A,\overline{B})P(\overline{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(\overline{AB})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{AB})+P(A,\overline{B})P(\overline{A},B)} = \frac{\alpha-1}{\alpha+1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(\overline{AB})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{AB})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$
6	Kappa (κ)	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
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)	$\max\Big(P(A,B)\log(\frac{P(B\|A)}{P(B)})+P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}),$ $P(A,B)\log(\frac{P(A\|B)}{P(A)})+P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})\Big)$
9	Gini index (G)	$\max\Big(P(A)[P(B\|A)^2+P(\overline{B}\|A)^2]+P(\overline{A})[P(B\|\overline{A})^2+P(\overline{B}\|\overline{A})^2]$ $-P(B)^2-P(\overline{B})^2,$ $P(B)[P(A\|B)^2+P(\overline{A}\|B)^2]+P(\overline{B})[P(A\|\overline{B})^2+P(\overline{A}\|\overline{B})^2]$ $-P(A)^2-P(\overline{A})^2\Big)$
10	Support (s)	$P(A,B)$
11	Confidence (c)	$\max(P(B\|A),P(A\|B))$
12	Laplace (L)	$\max\Big(\frac{NP(A,B)+1}{NP(A)+2},\frac{NP(A,B)+1}{NP(B)+2}\Big)$
13	Conviction (V)	$\max\Big(\frac{P(A)P(\overline{B})}{P(A\overline{B})},\frac{P(B)P(\overline{A})}{P(B\overline{A})}\Big)$
14	Interest (I)	$\frac{P(A,B)}{P(A)P(B)}$
15	cosine (IS)	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
16	Piatetsky-Shapiro's (PS)	$P(A,B)-P(A)P(B)$
17	Certainty factor (F)	$\max\Big(\frac{P(B\|A)-P(B)}{1-P(B)},\frac{P(A\|B)-P(A)}{1-P(A)}\Big)$
18	Added Value (AV)	$\max(P(B\|A)-P(B),P(A\|B)-P(A))$
19	Collective strength (S)	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})}\times\frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
20	Jaccard (ζ)	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
21	Klosgen (K)	$\sqrt{P(A,B)}\max(P(B\|A)-P(B),P(A\|B)-P(A))$

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

Properties of Measures

- Invariance under Variable Permutation (Symmetry)? X ↔ Y
- Invariance under Row/Column Scaling? E.g.:

	Male	Female	
High	2	3	5
Low	1	4	5
	3	7	10

↔

	Male	Female	
High	4	30	34
Low	2	40	42
	6	70	76

- Invariance under Inversion? X ↔ $\bar{X}$
- Invariance under Null Addition?
  - Adding data unrelated to X/Y, i.e. increasing f_{00} and N, should not affect the measure.

Different Measures have Different Properties

Symbol	Measure	Range	Symm	Scal	Invers	Null
Φ	Correlation	-1 … 0 … 1	Yes	No	Yes	No
λ	Lambda	0 … 1	Yes	No	Yes	No
α	Odds ratio	0 … 1 … ×	Yes	Yes	Yes	No
Q	Yule's Q	-1 … 0 … 1	Yes	Yes	Yes	No
Y	Yule's Y	-1 … 0 … 1	Yes	Yes	Yes	No
κ	Cohen's	-1 … 0 … 1	Yes	No	Yes	No
M	Mutual Information	0 … 1	Yes	No	Yes	No
J	J-Measure	0 … 1	No	No	No	No
G	Gini Index	0 … 1	No	No	Yes	No
s	Support	0 … 1	Yes	No	No	No
c	Confidence	0 … 1	Yes	No	No	Yes
L	Laplace	0 … 1	Yes	No	No	No
V	Conviction	0.5 … 1 … ×	No	No	Yes	No
I	Interest	0 … 1 … ×	Yes	No	No	No
IS	IS (cosine)	0 .. 1	Yes	No	No	Yes
PS	Piatetsky-Shapiro's	-0.25 … 0 … 0.25	Yes	No	Yes	No
F	Certainty factor	-1 … 0 … 1	No	No	Yes	No
AV	Added value	0.5 … 1 … 1	No	No	No	No
S	Collective strength	0 … 1 … ×	Yes	No	Yes	No
ζ	Jaccard	0 .. 1	Yes	No	No	Yes
K	Klosgen's	$\left(\sqrt{\frac{2}{\sqrt{3}}-1}\right)\left(2-\sqrt{3}-\frac{1}{\sqrt{3}}\right)\ldots 0 \ldots \frac{2}{3\sqrt{3}}$	No	No	No	No

A Paradox beyond Pairs of Binary Variables

Table 6.19. A two-way contingency table between the sale of high-definition television and exercise machine.

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

Table 6.20. Example of a three-way contingency table.

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

- HDTV=Yes → Exercise machine=Yes *has* confidence 99/180 = 55%
  thus has higher confidence than
- HDTV=No → Exercise machine=Yes *has* confidence 54/120=45%

but...

A Paradox (contd.)

For college students:

$$\begin{aligned} c(\{\text{HDTV=Yes}\} \longrightarrow \{\text{Exercise machine=Yes}\}) &= 1/10 = 10\%, \\ c(\{\text{HDTV=No}\} \longrightarrow \{\text{Exercise machine=Yes}\}) &= 4/34 = 11.8\%, \end{aligned}$$

while for working adults:

$$\begin{aligned} c(\{\text{HDTV=Yes}\} \longrightarrow \{\text{Exercise machine=Yes}\}) &= 98/170 = 57.7\%, \\ c(\{\text{HDTV=No}\} \longrightarrow \{\text{Exercise machine=Yes}\}) &= 50/86 = 58.1\%. \end{aligned}$$

- Hidden variables may cause observed relationships to disappear or to be reversed (Simpson‘s paradox).
- Data should be stratified “properly“ before analysis.

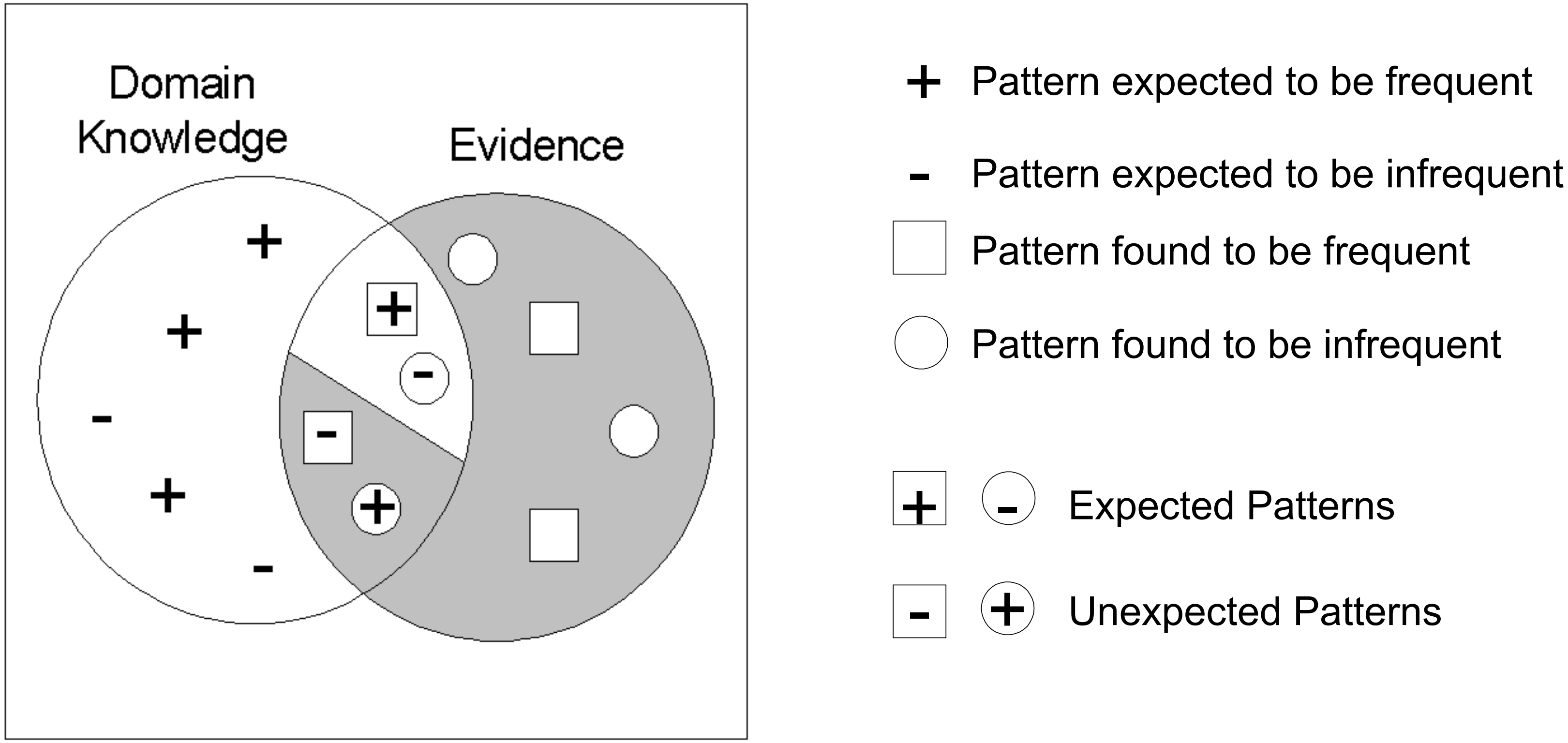
Subjective Interestingness Measure

- Objective measure:
  - Rank patterns based on statistics computed from data
  - e.g., 21 measures from above

- Subjective measure:
  - Rank patterns according to user's interpretation
    - A pattern may be subjectively interesting if it contradicts the expectation of a user
    - A pattern is subjectively interesting if it is actionable

Interestingness via Unexpectedness

- Need to model expectation of users (domain knowledge)



- Need to combine expectation of users with evidence from data (i.e., extracted patterns)

Effect of Support Distribution

- Many real data sets have skewed support distribution

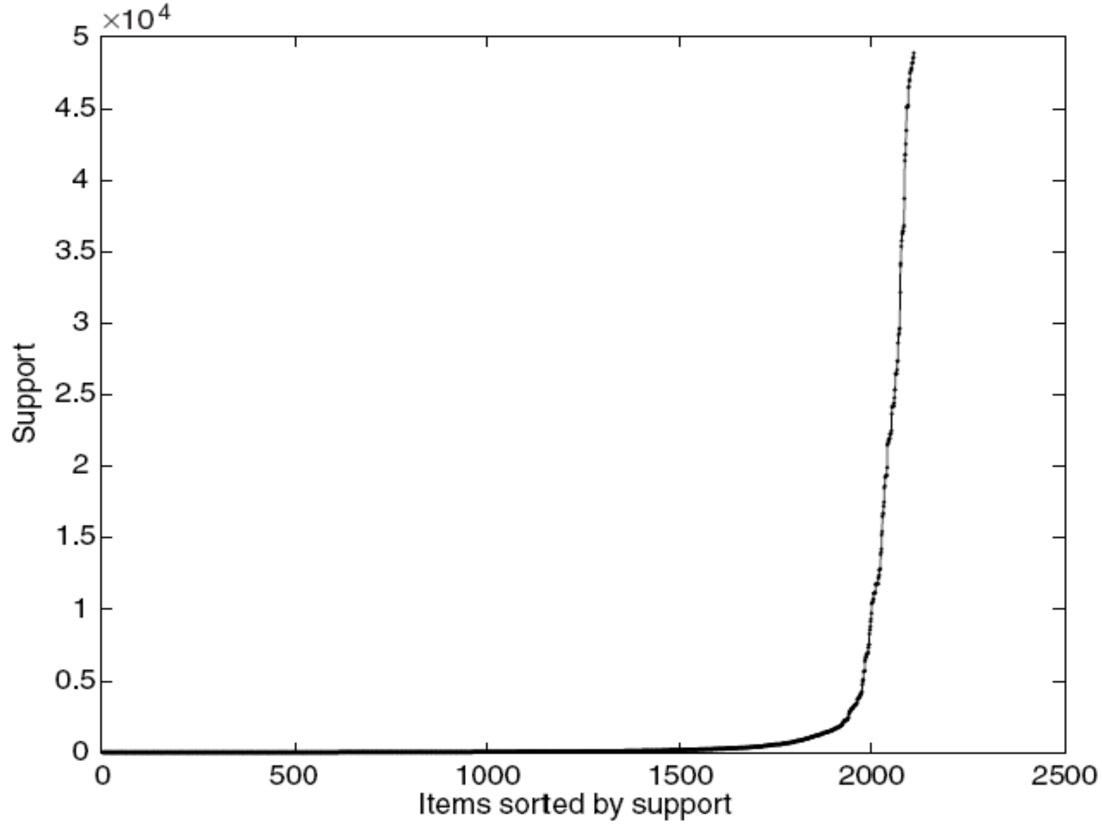


Figure 6.29. Support distribution of items in the census data set.

Effect of Support Distribution (contd.)

- How to set the appropriate *minsup* threshold?
  - If *minsup* is set too high (e.g. 20%), we could miss itemsets involving interesting rare items (e.g., expensive products)
  - If *minsup* is set too low,
    - it is computationally expensive
    - the number of extracted patterns is very large
    - there may be many spurious patterns that relate a high-frequency item to a low frequency item , e.g. in the rule “caviar → milk”, but weak correlation

Cross-support patterns

- Using a single minimum support threshold may not be effective => try multiple supports

Multiple Minimum Support

- How to apply multiple minimum supports?
  - Let MIS(i)=**minimum item support** for item i be given.
    - MIS(Milk)=5%, MIS(Coke) = 3%, MIS(Broccoli)=0.1%, MIS(Salmon)=0.5%
  - Define **minimum support** of an itemset

    $$MS(X):=\min\{ MIS(i) \mid i \in X\}$$

    - MS({Milk, Broccoli}) := min(MIS(Milk), MIS(Broccoli)) = 0.1%
  - Challenge: Frequentness is no longer anti-monotone
    - Suppose: Support(Milk, Coke) = 1.5%
      Support(Milk, Coke, Broccoli) = 0.5%
    - {Milk,Coke} is infrequent (Support<MS=3%) , but
      {Milk,Coke,Broccoli} is frequent (Support ≥ MS=0.1%)
  - But minimum support MS is.

Multiple Minimum Support Algorithm

- Modify Apriori.
- Order the items according to their minimum support (in ascending order)
  - e.g. acc. to MIS above: Broccoli, Salmon, Coke, Milk
  - thus $MS(X) = MIS(i_1)$, if $X=(i_1,...,i_k)$:
- Generate longer frequent itemsets by again joining
  - If (Broccoli,Coke,Milk) is frequent, it must be generatable from (Broccoli,Coke) & (Broccoli,Milk), which must be frequent as well

    support({Broccoli,Coke}) ≥ support({Broccoli,Coke,Milk}) ≥ min.support({Broccoli,Coke,Milk}) = MIS(Broccoli)
- But pruning may only check subsets with same MS
  - A candidate is pruned if it contains any infrequent subsets of size k <u>with same first item</u>.
  - Candidate (Broccoli,Coke,Milk) is not pruned although (Coke, Milk) is infrequent.

Eliminating Cross-Support Patterns

- Cross-support patterns often induce rules of very high confidence, e.g. caviar → milk, or of very low confidence, e.g milk → caviar. They should be eliminated.
- A reasonable criterion for a **cross-support pattern** is: an itemset $X=\{i_1,\ldots,i_k\}$, whose support ratio

  $$r(X) = \min[s(i_1),\ldots,s(i_k)] \;/\; \max[s(i_1),\ldots,s(i_k)]$$

  is below a user-specified threshold h_{cross} .
- The **all-confidence** - measure of an itemset $X=\{i_1,\ldots,i_k\}$:

  $$ac(X) = s(\{i_1,\ldots,i_k\}) \;/\; \max\,[s(i_1),\ldots,s(i_k)] \leq r(X)$$

  is a lower bound for the lowest confidence attainable from X. If it stays above h_{cross}, the support ratio r does as well.
- The latter measure is even anti-monotone, i.e. its control can be incorporated into the generation algorithm.