HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

# 582364 Data mining, 4 cu
# Lecture 5:
## Evaluation of Association Patterns

Spring 2010

Lecturer: Juho Rousu

Teaching assistant: Taru Itäpelto

# Evaluation of Association Patterns

- Association rule algorithms potentially generate large quantities of rules
  - Easily thousands to millions of rules depending on the database and the used support and confidence levels
  - All of the patterns cannot be examined manually
- Problem in using the knowledge contained in the rules
  - All of the rules may not be interesting (e.g. Plastic bag -> Bread)
  - Some rules may be redundant (e.g if {A,B,C} → {D} and {A,B} → {D} have same support & confidence)

# Effect of Skewed Support Distribution

- Many real data sets have skewed support distribution
  - Most of items have low to moderate support
  - Small number of items have very high support

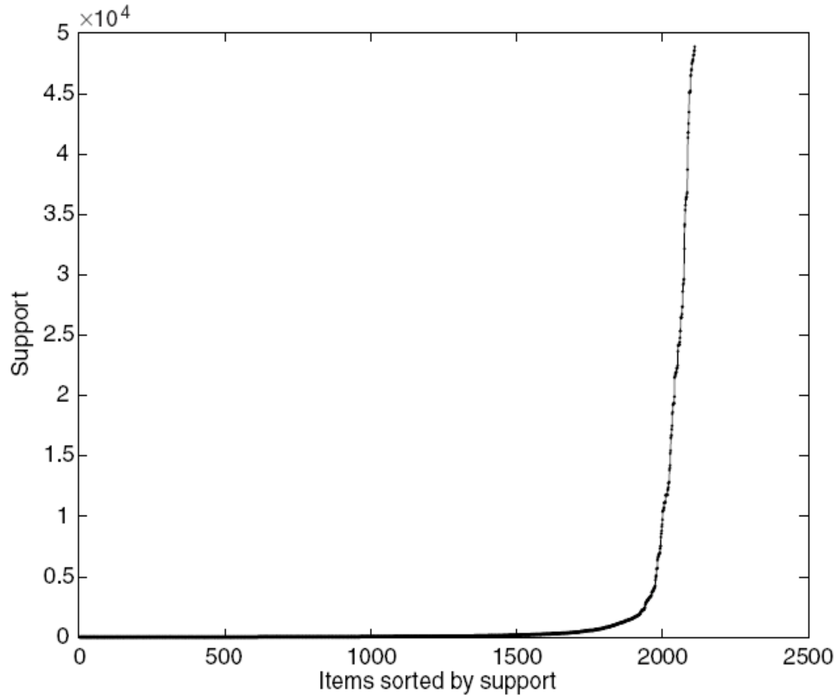


**Figure 6.29.** Support distribution of items in the census data set.

| Group | G1 | G2 | G3 |
|---|---|---|---|
| Support | <1% | 1-90% | > 90% |
| #items | 1735 | 358 | 20 |

# Effect of Skewed Support Distribution

- How to set minsup threshold?
- Too high *minsup* threshold (e.g. 20%) misses interesting items with low support
  - e.g. customers buying expensive jewelry or other high-profit items
- Too low *minsup* threshold
  - generates too many rules
  - easily generates spurious cross-patterns relating a low-frequency item to a high-frequency item: e.g. Caviar → Bread

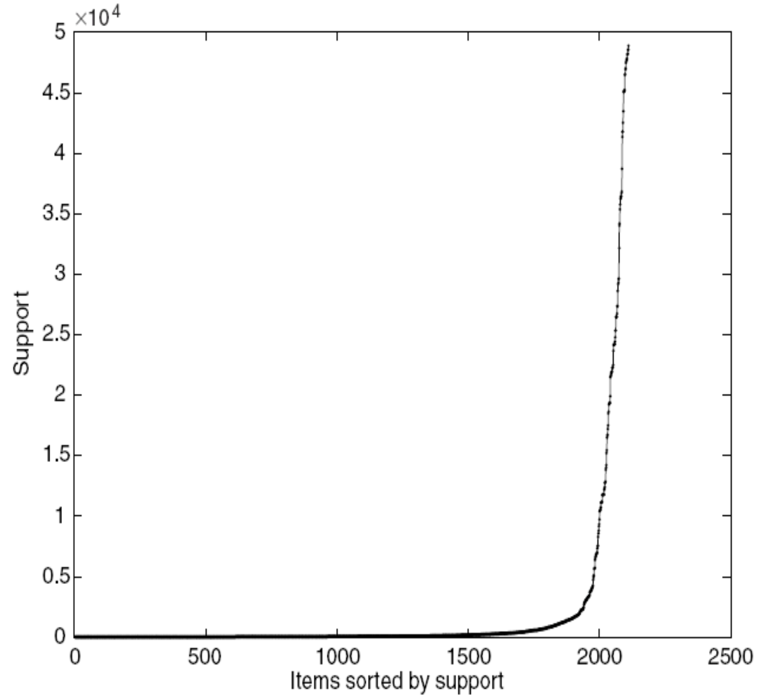


**Figure 6.29.** Support distribution of items in the census data set.

| Group | G1 | G2 | G3 |
|---|---|---|---|
| Support | <1% | 1-90% | > 90% |
| #items | 1735 | 358 | 20 |

# Multiple Minimum Support

- One solution is to apply multiple minimum supports levels
  - MS(i): minimum support for item i
  - e.g.: MS(Milk)=5%, MS(Coke)=3%,
    MS(Broccoli)=0.1%, MS(Salmon)=0.5%
  - MS({Milk, Broccoli})= min(MS(Milk),MS(Broccoli})=0.1%
  - Challenge: Support is no longer anti-monotone
    - Suppose: Support(Milk, Coke) = 1.5% and
      Support(Milk, Coke, Broccoli) = 0.5%
    - {Milk,Coke} is infrequent but {Milk,Coke,Broccoli} is frequent
  - Apriori can be modified to accommodate this change (Liu, 1999)
    - The pruning of candidate itemsets needs to be relaxed

# Cross-support patterns

- Consider the transaction data on the right
- q → p , r → p and {q,r} → p are all high-confidence patterns that look spurious (caused by p being very frequent)
- Eliminating them by tightening the *minsup* requirement also drops the rules r → q and q → r that "look ok"
- A cross-support pattern is an itemset $X=\{X_1,\ldots,X_k\}$ with low ratio

$$r(X) = \frac{\min[s(X_1),\ldots,s(X_k)]}{\max[s(X_1),\ldots,s(X_k)]}$$

| p | q | r |
|---|---|---|
| 0 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 0 | 0 | 0 |
| 0 | 0 | 0 |
| 0 | 0 | 0 |
| 0 | 0 | 0 |

# Eliminating cross-support patterns

- Recall the definition of confidence $c(X \rightarrow Y) = \sigma(X,Y)/\sigma(X)$
and its anti-monotone property (confidence can only decrease when items are moved from left to right-hand side of the rule)
- Given an itemset $X = \{X_1,\ldots,X_k\}$ the lowest confidence rule that can be generated from X is the one with the highest support item on the left-hand side
- The lowest confidence (or *all-confidence*) can be used to measure the potential of the itemset to generate cross-support patterns
- Itemsets with low all-confidence can be filtered out before rule generation

$$allconfidence(X) = \frac{s(X_1,\ldots,X_k)}{\max[s(X_1),\ldots,s(X_k)]}$$

# Evaluation of Association Patterns

- We will look at methods that let us rank or prune the discovered set of rules
  - Called “Interestingness measures” in data mining community
- Objective interestingness measures: statistical methods to measure how exceptional the pattern is with respect to background assumptions
- Subjective interestingness measures:
  - Using domain knowledge, e.g. filtering out obvious patterns or patterns that cannot be acted on
  - In general, requires a “human in the loop”
- To some extent, an art rather than science: “one man’s trash is another man’s treasure”

# Properties of Interestingness Measures

- Interestingness measures can be divided into two main categories based on their use
- Symmetric measures M
  - satisfy M(A→B) = M(B→A)
  - used to evaluate itemsets
  - e.g. support
- Asymmetric measures
  - generally give different values for M(A→B) and M(B→A)
  - used to evaluate association rules
  - e.g. confidence

# Contingency table

- The contingency table for rule X → Y is given by the support of four different combinations of observing X,Y, both or neither of them
- Contingency table contains sufficient information to compute different interestingness measures
- Intuitively: if $f_{11}$ has high support compared to the other cells, the rule is more likely to be interesting than not

Contingency table for X → Y

| | Y | not Y | |
|---|---|---|---|
| X | $f_{11}$ | $f_{10}$ | $f_{11}+f_{10}$ |
| not X | $f_{01}$ | $f_{00}$ | $f_{01}+f_{10}$ |
| | $f_{11}+f_{01}$ | $f_{10}+f_{00}$ | $\|T\|$ |

$f_{11}$: support of X and Y
$f_{10}$: support of X and not Y
$f_{01}$: support of not X and Y
$f_{00}$: support of not X and not Y

# Properties of Objective Measures: Inversion property

- An evaluation measure is invariant under inversion if its value remains the same when
  - flipping the attribute values from 0→1 and 1→0, or equivalently,
  - permuting the contingency table, f00 →f11 and f01→f10
- This property is not desirable for evaluating asymmetric attributes: e.g. items that are not bought by the customer would provide as strong associations as items the customer bought

| | Y | not Y | |
|---|---|---|---|
| X | 60 | 10 | 70 |
| not X | 10 | 20 | 30 |
| | 70 | 30 | 100 |

| | Y | not Y | |
|---|---|---|---|
| X | 20 | 10 | 30 |
| not X | 10 | 60 | 70 |
| | 30 | 70 | 100 |

# Properties of Objective Measures: Null Addition Property

- An evaluation property is invariant under null addition if it does not change its value when the value f00 is inreased in the contingency table
- This is useful property in applications such as market-basket analysis where the non-absense of items is not the focus of the analysis

| | Y | not Y | |
|---|---|---|---|
| X | 60 | 10 | 70 |
| not X | 10 | 20 | 30 |
| | 70 | 30 | 100 |

| | Y | not Y | |
|---|---|---|---|
| X | 60 | 10 | 70 |
| not X | 10 | 920 | 930 |
| | 70 | 930 | 1000 |

# Properties of Objective Measures: Scaling property

- An evaluation measure is invariant under row/column each column and row can be multiplied by a constant without the measure to change its value
- Most evaluation measures do not satisfy this property (odds ratio = f11*f00/(f10*f01) is an exception)
- Below, column ‘not Y’ has been multiplied by 2, row ‘X’ by 3 and row ‘not X’ by 4

| | Y | not Y | |
|---|---|---|---|
| X | 6 | 1 | 7 |
| not X | 1 | 2 | 3 |
| | 7 | 3 | 10 |

| | Y | not Y | |
|---|---|---|---|
| X | 18 | 6 | 24 |
| not X | 4 | 16 | 20 |
| | 22 | 22 | 44 |

# Property under Row/Column Scaling

Grade-Gender Example (Mosteller, 1968):

| | 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 |

↓ 2x ↓ 10x

Mosteller:

Underlying association should be independent of the relative number of male and female students in the samples

# Drawback of Confidence

- Via the use of contingency tables one can illustrate a drawback of the confidence measure
- Consider the rule Tea → Coffee
  - support 15/100 = 15%
  - confidence 15/20 = 75%
- looks ok?

| | Coffee | $\overline{\text{Coffee}}$ | |
|---|---|---|---|
| Tea | 15 | 5 | 20 |
| $\overline{\text{Tea}}$ | 75 | 5 | 80 |
| | 90 | 10 | 100 |

# Drawback of Confidence

- Consider the rule Tea → Coffee
  - support 15/100 = 15%
  - confidence 15/20 = 75%
- But the fraction of people drinking coffee regardsless of whether they drink tea is 90%
- Thus knowing that the person drinks tea actually lowers our expectation that the person drinks coffee
- The rule is misleading!

| | Coffee | $\overline{\text{Coffee}}$ | |
|---|---|---|---|
| Tea | 15 | 5 | 20 |
| $\overline{\text{Tea}}$ | 75 | 5 | 80 |
| | 90 | 10 | 100 |

# Lift and Interest factor

- Confidence c(A→B) = σ(A,B)/σ(A) ignores the support of the itemset on the right-hand side of the rule
- *Lift* is a measure that aims to fix this problem
- For binary variables lift is equal to *interest factor*
- Lift/interest factor is
  - symmetric
  - not invariant under inversion
  - not invariant under null addition
  - not invariant under scaling

$$Lift = \frac{c(A \rightarrow B)}{s(B)}$$

$$I(A,B) = \frac{s(A,B)}{s(A)s(B)}$$

# Lift and interest factor

- Interpretation: compare the support of itemset {A,B} to the expected support under the assumption that A and B are statistically independent:
  - s(A,B) ≈ P(A and B)
  - s(A) ≈ P(A), s(B) ≈ P(B)
  - Statistical independence: P(A and B) = P(A)xP(B)

$$I(A,B) = \frac{s(A,B)}{s(A)s(B)}$$

- Use of interest factor:
  - I(A,B) >1 : A and B occur together more frequently than expected by chance
  - I(A,B) < 1 : A and B occur together less frequently than expected by chance

# Example: Lift/interest factor

- Let us compute the interest factor for our Tea → Coffee rule
- I(Tea,Coffee) = c(Tea → Coffee)/s(Coffee) = 0.75/0.9 = 0.83
- I < 1 denotes the pattern occurs less often than expected from independent events
- Conforms to our everyday intuition!

| | Coffee | $\overline{\text{Coffee}}$ | |
|---|---|---|---|
| Tea | 15 | 5 | 20 |
| $\overline{\text{Tea}}$ | 75 | 5 | 80 |
| | 90 | 10 | 100 |

# Drawback of Lift & Interest

- Lift/Interest loses its sensitivity when support of the itemset is very high
  - in the above contingency table, X and Y look almost statistically independent (I(X,Y) = 1 for independent items)

$$I(X,Y) = \frac{s(X,Y)}{s(X)s(Y)} = \frac{0.9}{0.9 \times 0.9} = 1.11$$

| | Y | not Y | |
|---|---|---|---|
| X | 90 | 0 | 90 |
| not X | 0 | 10 | 10 |
| | 90 | 10 | 100 |

$$I(X,Y) = \frac{s(X,Y)}{s(X)s(Y)} = \frac{0.1}{0.1 \times 0.1} = 10$$

| | Y | not Y | |
|---|---|---|---|
| X | 10 | 0 | 10 |
| not X | 0 | 90 | 90 |
| | 10 | 90 | 100 |

# Correlation analysis: ϕ-coefficient

- For binary variables, correlation can be measured using the ϕ-coefficient:
- In our Tea → Coffee example the ϕ-coefficient amounts to
  ϕ = (15*5-75*5)/√(90x20x10x80)
  =-0.25

$$\phi = \frac{\sigma(XY)\sigma(\bar{X}\bar{Y}) - \sigma(X\bar{Y})\sigma(\bar{X}Y)}{\sqrt{\sigma(X)\sigma(Y)\sigma(\bar{X})\sigma(\bar{Y})}}$$

- ϕ-coefficient is
  - symmetric
  - Invariant under inversion
  - not invariant under null addition
  - not invariant under scaling

| | Coffee | not Coffee | |
|---|---|---|---|
| Tea | 15 | 5 | 20 |
| not Tea | 75 | 5 | 80 |
| | 90 | 10 | 100 |

# Property of $\phi$-Coefficient

- $\phi$-Coefficient considers the co-occurence and co-absense equally important: the two contingency tables evaluate to the same value
- This makes the measure more suitable to symmerical variables

| | Y | Y | |
|---|---|---|---|
| X | 60 | 10 | 70 |
| X | 10 | 20 | 30 |
| | 70 | 30 | 100 |

| | Y | not Y | |
|---|---|---|---|
| X | 20 | 10 | 30 |
| not X | 10 | 60 | 70 |
| | 30 | 70 | 100 |

$$\phi = \frac{60 \times 20 - 10 \times 10}{\sqrt{70 \times 30 \times 70 \times 30}} = 0.5238$$

$$\phi = \frac{20 \times 60 - 10 \times 10}{\sqrt{70 \times 30 \times 70 \times 30}} = 0.5238$$

## IS Measure

- IS Measure is an alternative measure proposed for asymmetric binary variables
- Equivalent to the cosine measure used in information retrieval
- IS Measure is
  - symmetric
  - not invariant under inversion
  - invariant under null addition
  - not invariant under scaling

$$IS(A,B) = \frac{s(A,B)}{\sqrt{s(A)s(B)}}$$

$$= \sqrt{I(A,B) \times s(A,B)}$$

$$\cos ine(x,y) = \sum_t x_t y_t \Big/ \|x\|\|y\|$$

# Testing statistical significance: p-values

- The interestingness measures discussed before are related to the concept of statistical hypothesis testing
- In hypothesis testing, we have two competing hypotheses
  - H0: null hypothesis, assuming that the pattern seen in the data is created by random variation
    - e.g. the value c(X→Y) is a result of random fluctuation
  - H1: hypothesis that the pattern seen in the data represents true phenomenon
- The probability of observing the pattern if the null hypothesis is true is the p-value
  - smaller p-values are more significant

# Randomization

- Randomization is a general family of methods for assessing the statistical validity of data mining results
- Is used as an alternative to statistical tests, when the test statistic is too difficult to determine
- Basic idea:
  - Given dataset D, generate a large collection of datasets $D_1,...,D_N$ where the statistical association of interest has been broken
  - Run the data mining algorithm on all of the generated datasets and record the distribution of the property of interest
  - If the property we observe in the original data falls into top p% of observations, we consider our data mining results significant

# Simple randomization example

- Assume we want to assess the statistical significance of the support $s(X \rightarrow Y)$ and confidence $c(X \rightarrow Y)$ of the association rule $X \rightarrow Y$
  - Y can contain more than one item
- From the dataset D, generate new datasets $D_1,\ldots,D_{1000}$ by generating a random permutation $R_j$ of rows and setting

  $$D_j(i,Y) = D(R_j,Y)$$
- Compute support and confidence of the rule $X \rightarrow Y$ in each version of the data: $s_j(X \rightarrow Y)$, $c_j(X \rightarrow Y)$
- Sort the obtained support and confidence values and record the position from top where the values $s(X \rightarrow Y)$ and $c(X \rightarrow Y)$ fall
- Take the relative positions (fraction from the top) as estimates of statistical significance (p-value)

# Swap randomization

- In the previous example, we could have as well computed analytically the probability of observing such support and confidence values
- The power of randomization comes more evident when the baseline hypothesis is more complicated
- Consider situation where we want to keep both the width of each transaction (how many items per transaction) and the support counts of individual items intact
  - keeping a the size of shopping basket intact
  - as well as the overall demand of items
- We look briefly at the approach described in Hanhijärvi et al.: "Tell me Something I don't know: Randomization Strategies for Iterative Data Mining". Proc. KDD'09.

# Swap randomization

- The row margins (widths of transactions) and column margins (support counts of items) can be preserved by swap randomization
- A randomized version of the dataset is generated via series of swaps
- In each swap,
  - take two rows s,t and two columns x,y such that D(s,x) = D(t,y) = 1 and D(s,y) = D(t,x) = 0
  - swap the contents: $D_j(s,x) = D_j(t,y) = 0$ and $D_j(s,y) = D_j(t,x) = 1$

| | x | | y |
|---|---|---|---|
| s | 1 | | 0 |
| | | | |
| t | 0 | | 1 |

→

| | x | | y |
|---|---|---|---|
| s | 0 | | 1 |
| | | | |
| t | 1 | | 0 |

# Algorithm for creating a swap randomized dataset

Algorithm Swap

Input : Dataset D, num. of swap attempts K

Output : Randomized dataset $\hat{D}$

1 : $\hat{D} = D$

2 : for i = 1 to K do

3 : Pick s,t and x,y such that $\hat{D}(s,x) = 1$, $\hat{D}(t,y) = 1$

4 : if $\hat{D}(s,y) = 0$ and $\hat{D}(t,x) = 0$ then

5 : $\hat{D} =$ swapped version of $\hat{D}$

6 : end if

7 : end for

8 : return $\hat{D}$

# Swap randomization

- After generating the collection of randomized datasets $D_1,...,D_N$, the statistical significance of the quantity of interest (e.g. support, confidence) is extracted
  - Collect the distribution of the quantity of interest from the randomized datasets (e.g. confidence of $c_j(X \rightarrow Y)$ in all of the datasets)
  - Sort the distribution and check how large fraction of the distribution is above the quantity computed from the original dataset
    - This is taken as the statistical significance of the quantity
    - below the confidence value $c(X \rightarrow Y)=0.80$ is in place r, so p-value is $p = r/N$

| **1** | **2** | **3** | **...** | **r** | **r+1** | **...** | **N-1** | **N** |
|---|---|---|---|---|---|---|---|---|
| 0.87 | 0.85 | 0.85 | ... | 0.80 | 0.79 | ... | 0.17 | 0.15 |

# Randomization: summary

- Randomization is powerful an general technique for assessing statistical significance
- It is particularly useful in situations where a traditional statistical testing is too difficult, e.g. when it is not evident what is the statistical distribution and the correct test in the given setting
- The drawback in data mining is its high time-complexity:
  - We need to create large numbers of randomized versions of our data
  - May not be possible with very large datasets