

# Data Mining

## Association Analysis Basic Concepts and Algorithms

Slides by Tan, Steinbach, Kumar adapted by Pimprapai Thainiam

# Topics

- ▶ **Rule Generation (Subtask 2)**
- ▶ Evaluation of Association Rules
- ▶ Handling Categorical and Continuous Attributes

# Rule Generation

- **Rule generation** is the method to **extract association rules from a given frequent itemsets** obtained from frequent itemsets generation.
- Given a frequent  $k$ -itemset,  $Y$ , an association rule can be extracted by partitioning the itemset  $Y$  into two non-empty subsets:

$X$

and

$Y - X$

$$X \cap Y - X = \emptyset$$

such that rule  $X \rightarrow Y - X$  satisfies the confidence threshold.

- If a frequent itemset,  $Y$ , contains  $k$  items, then  $Y$  can produce up to  $2^k - 2$  association rules (ignoring  $Y \rightarrow \emptyset$  and  $\emptyset \rightarrow Y$ ).

Example: If  $\{A,B,C\}$  is a frequent itemset, it can generate  $2^3 - 2 = 6$  candidate association rules:

$$\{A,B\} \rightarrow \{C\}$$

$$\{A,C\} \rightarrow \{B\}$$

$$\{B,C\} \rightarrow \{A\}$$

$$\{A\} \rightarrow \{B,C\}$$

$$\{B\} \rightarrow \{A,C\}$$

$$\{C\} \rightarrow \{A,B\}$$

# Rule Generation

- Since  $Y$  is frequent, the anti-monotone property of support ensures that  $X$  ( $X \subseteq Y$ ) must also be frequent; this indicates that there is no need to read the entire data set again to find the support count of  $X$ ,  $\sigma(X)$ .

$$c(X \rightarrow Y - X) = \frac{\sigma(Y)}{\sigma(X)}$$

*frequent itemsets*

Example: If  $\{1,2,3\}$  is frequent, then all of its subsets must also be frequent.

$$c(\{1,2\} \rightarrow \{3\}) = \frac{\sigma(\{1,2,3\})}{\sigma(\{1,2\})}$$

# Rule Generation

- Confidence does not have an anti-monotone property, but confidences of rules generated from the same frequent itemset,  $Y$ , possess an **anti-monotone property**.

*If a rule  $X \rightarrow Y - X$  does not satisfy the confidence threshold, then any rule  $X' \rightarrow Y - X'$ , where  $X'$  is a subset of  $X$ , must not satisfy the confidence threshold as well.*

Given  $X' \subset X \subset Y$  *→ frequent itemset generated from subtask 1*

$$\Rightarrow \sigma(X') \geq \sigma(X)$$

$$\Rightarrow c(X' \rightarrow Y - X') \leq c(X \rightarrow Y - X)$$

$$c(X \rightarrow Y - X) = \frac{\sigma(Y)}{\sigma(X)}$$

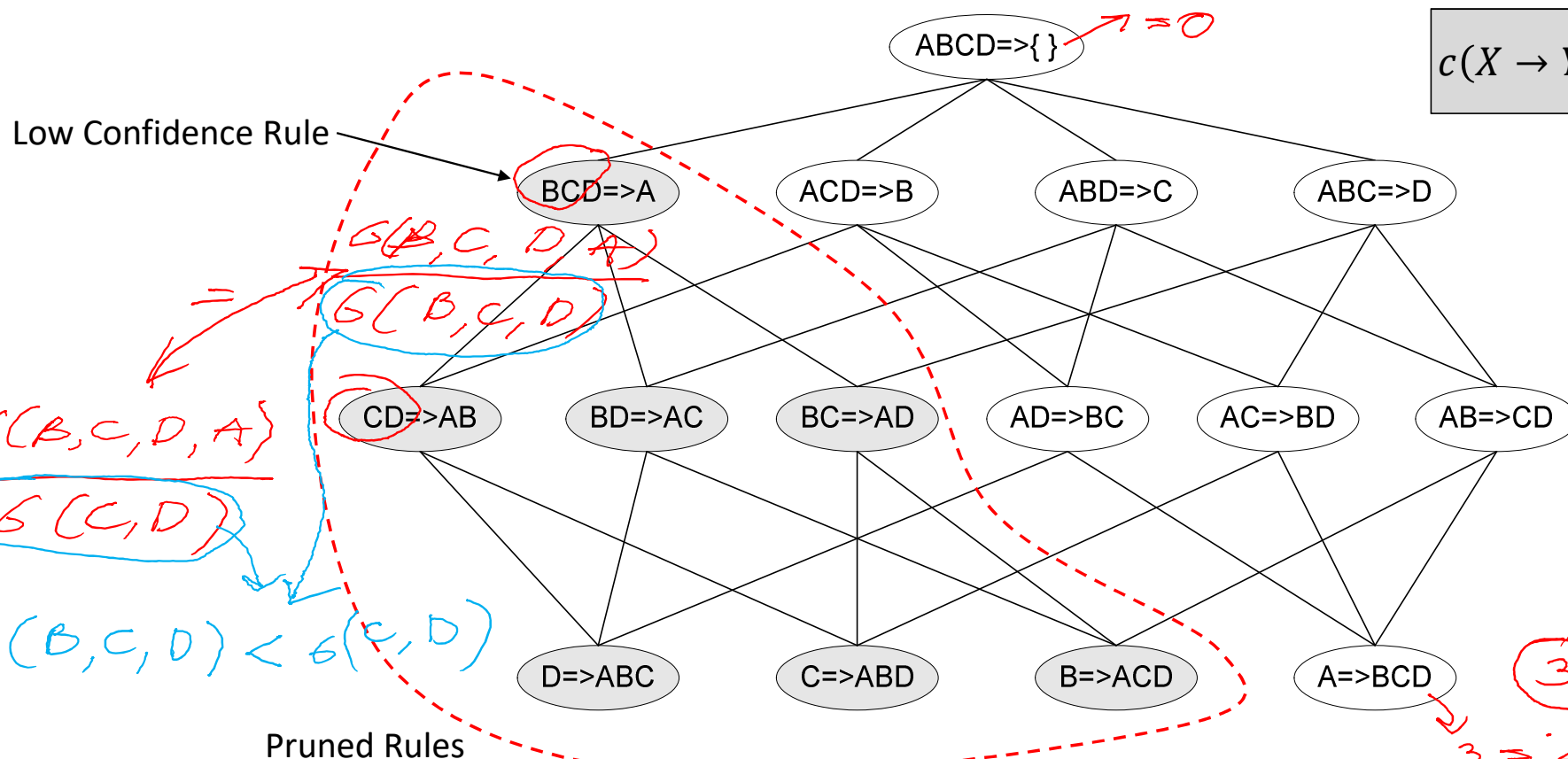
*→ same*

- This can be concluded that **confidence is anti-monotone with respect to number of items on the right-hand side of the rule**.
- The *Apriori* algorithm uses a **level-wise approach** for generating association rules, where each level corresponds to the number of items that belong to the rule consequent.

# Rule Generation

Confidence-based pruning prunes association rules using the confidence measure.

$$c(X \rightarrow Y - X) = \frac{\sigma(Y)}{\sigma(X)}$$



# Rule Generation in *Apriori* Algorithm

$minconf = 0.8$

- Rule generation steps in *Apriori* algorithm:

$\{A, B\} \rightarrow \{A\}, \{B\}$

Let  $m = 0$  and  $k = 2$  ( $m$  = Size of rule consequence and  $k$  = Size of frequent itemset)

Step 1: Generate a set of rule consequences of size  $m + 1$  from a given frequent itemset of size  $k$

$\{B\} \rightarrow \{A\}$  ;  $\{A\} \rightarrow \{B\}$  ;  $C(\{B\} \rightarrow \{A\}) = 0.7$   
 $C(\{A\} \rightarrow \{B\}) = 0.9$  ✓

Step 2: Generate all the high-confidence rules ( $conf \geq minconf$ ) from the rule consequence set obtained in step 1 and its related frequent itemset  $\{A, B\}$

Step 3: Generate a set of rule consequences of size  $m + 1$  from all rule consequences of size  $m$  found in high-confidence rules obtained in step 2.

see next slide

Step 4: Generate all the high-confidence rules ( $conf \geq minconf$ ) from the rule consequence set obtained in step 3 and its related frequent itemset

Step 5: Repeat step 3-4 until  $m = k - 1$

Step 6: Set  $m = 0$

Step 7: Repeat step 1-6 for all frequent  $k$ -itemset

Step 8: Set  $k = k + 1$

Step 9: Repeat step 1-8 for all sizes of  $k$

# Rule Generation in *Apriori* Algorithm

- Step 3 in rule generation can be done by

Step 1: Merge two consequences that share the same prefix (same first  $m - 1$  items)

Example: (1) Rule consequences are  $\{B,C\}$  and  $\{B,D\}$   $\Leftarrow$  sorted in lexicographic order

Merging result:  $\{B,C,D\}$

(2) Rule consequences are  $\{B,D\}$  and  $\{C,D\}$   $\Leftarrow$  sorted in lexicographic order

Merging result: Cannot be merged

Step 2: Prune the rule obtained from consequence generated in step 1 and its subset rules if its confidence is less than *minconf*.

Example: Frequent itemset =  $\{A,B,C,D\}$

Rule consequence =  $\{A,B\}$

Low-confidence rule =  $\{C,D\} \rightarrow \{A,B\}$   $\Leftarrow$

Its subsets =  $\{C\} \rightarrow \{A,B,D\}$   
 $\{D\} \rightarrow \{A,B,C\}$  } pruned



# Topics

- ▶ Rule Generation (Subtask 2)
- ▶ **Evaluation of Association Rules**
- ▶ Handling Categorical and Continuous Attributes

# Rules Evaluation

It is important to establish a set of well-accepted criteria for evaluating the quality of association patterns.

- 1. Statistical arguments criteria:** Patterns that involve a set of mutually independent items or cover very few transactions are considered **uninteresting** because they may capture spurious relationships in the data which can be eliminated by applying an **objective interestingness measure** that uses statistics derived from data to determine whether a pattern is interesting, such as support and confidence.
- 2. Subjective arguments criteria:** A pattern is considered subjectively uninteresting unless it reveals unexpected information about the data or provides useful knowledge that can lead to profitable actions. For example,

Not interesting pattern:  $\{Butter\} \rightarrow \{Bread\}$     Obvious!

Interesting pattern:  $\{Diapers\} \rightarrow \{Beer\}$     Unexpected relationship!

This relationship may suggest a new cross-selling opportunity for retailers. It is difficult to incorporate subjective knowledge into pattern evaluation.

# Objective Measures of Interestingness

## Contingency Table for $X \rightarrow Y$

Information needed for determining rule interestingness:

$X$  and  $Y$  indicate the **presence** of  $X$  and  $Y$ .

$\bar{X}$  and  $\bar{Y}$  indicate the **absence** of  $X$  and  $Y$ .

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

$f_{11}$  is the number of transactions that contain both  $X$  and  $Y$

$f_{01}$  is the number of transactions that contain  $Y$  but not  $X$

$f_{10}$  is the number of transactions that contain  $X$  but not  $Y$

$f_{00}$  is the number of transactions that do not contain both  $X$  and  $Y$

$f_{1+}$  is the support count for  $X$ ,  $\sigma(X)$

$f_{+1}$  is the support count for  $Y$ ,  $\sigma(Y)$

total number  
of transaction

This table is used to define various measures support, confidence, lift, Gini, J-measure, etc.

# Support and Confidence

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

Support of  $X \rightarrow Y$  :  $s(X \rightarrow Y) = \frac{f_{11}}{N}$

$$\rightarrow \frac{G(X \cup Y)}{N}$$

Confidence of  $X \rightarrow Y$  :  $c(X \rightarrow Y) = \frac{f_{11}}{f_{1+}}$

$$\rightarrow \frac{G(X \cup Y)}{G(X)}$$

# Limitation of the Support-Confidence Framework

	<i>Coffee</i>	$\overline{Coffee}$	
<i>Tea</i>	150	50	200
$\overline{Tea}$	650	150	800
	800	200	1,000

Association Rule:  $\{Tea\} \rightarrow \{Coffee\}$

$$\text{Rule's Support} = \frac{f_{11}}{N} = \frac{150}{1,000} = 0.15$$

$$\text{Rule's Confidence} = \frac{f_{11}}{f_{1+}} = \frac{150}{200} = 0.75$$

Handwritten notes: "doesn't include S(Coffee)" and "f<sub>1+</sub>" with arrows pointing to the denominator 200.

$$\text{Support of Coffee} = \frac{800}{1,000} = 0.8$$

- The fraction of people who drink coffee, regardless of whether they drink tea, is 80%, while the fraction of tea drinkers who drink coffee is only 75%.
- Thus, knowing that a person is a tea drinker actually decreases her probability of being a coffee drinker from 80% to 75%!
- The rule  $\{Tea\} \rightarrow \{Coffee\}$  is misleading despite its high confidence value.
- The pitfall of confidence is that the confidence ignores the support of the itemset in the rule consequent.

# Limitation of the Support-Confidence Framework

- Because of the limitations in the support-confidence framework, various objective measures have been used to evaluate the quality of association patterns, such as
  - Interest Factor
  - Correlation Analysis
  - IS Measure
  - etc.

# Interest Factor

- The high-confidence rules can sometimes be misleading because the confidence measure ignores the support of the itemset appearing in the rule consequent.
- **Lift** is the ratio between the rule's confidence and the support of the itemset in the rule consequent.

$$Lift = \frac{c(X \rightarrow Y)}{s(Y)}$$

→ additional part

- For binary variable, lift is equivalent to interest factor.

$$Interest\ Factor = I(X, Y) = \frac{s(X \cup Y)}{s(X) \times s(Y)} = \frac{N f_{11}}{f_{1+} f_{+1}}$$

- Interpretation of interest factor:

$$I(X, Y) \begin{cases} = 1, & \text{if } X \text{ and } Y \text{ are independent;} \\ > 1, & \text{if } X \text{ and } Y \text{ are positively correlated;} \\ < 1, & \text{if } X \text{ and } Y \text{ are negatively correlated;} \end{cases}$$

# Interest Factor

Example:

	<i>Coffee</i>	$\overline{Coffee}$	
<i>Tea</i>	150	50	200
$\overline{Tea}$	650	150	800
	800	200	1,000

Association Rule:  $\{Tea\} \rightarrow \{Coffee\}$

$$I = \frac{Nf_{11}}{f_{1+}f_{+1}} = \frac{1,000 \times 150}{200 \times 800} = 0.9375$$

This suggests a slight negative correlation between tea drinkers and coffee drinkers.



# Limitation of Interest Factor

**Example:** Contingency tables for the word pairs  $\{p, q\}$  and  $\{r, s\}$

	$p$	$\bar{p}$	
$q$	880	50	930
$\bar{q}$	50	20	70
	930	70	1,000

$$s(p \cup q) = \frac{880}{1,000} = 0.88$$

$$I(p, q) = \frac{(1,000)(880)}{(930)(930)} = 1.017$$

	$r$	$\bar{r}$	
$s$	20	50	70
$\bar{s}$	50	880	930
	70	930	1,000

$$s(r \cup s) = \frac{20}{1,000} = 0.02$$

$$I(r, s) = \frac{(1,000)(20)}{(70)(70)} = 4.082$$

- $p$  and  $q$  appear together in 88% of the documents, their interest factor is close to 1 which **indicates independency of  $p$  and  $q$** .
- $r$  and  $s$  rarely appear together in the same document, but they are positively correlated.
- Confidence is perhaps the better choice in this situation. [ $c(p, q) = 94.6\% > c(r, s) = 28.6\%$ ]

# Correlation Analysis

- **Correlation Analysis** is a statistical-based technique for analyzing relationships between a pair of variables which can be measured using the  $\phi$ -coefficient.

$$\phi = \frac{f_{11}f_{00} - f_{01}f_{10}}{\sqrt{f_{1+}f_{+1}f_{0+}f_{+0}}}$$

- It ranges from -1 (perfect negative correlation) to +1 (perfect positive correlation); and  $\phi = 0$  means the variables are statistically independent.
- **Example:**  $\{Tea\} \rightarrow \{Coffee\}$

$$\phi = \frac{(150)(150) - (650)(50)}{\sqrt{(200)(800)(800)(200)}} = -0.0625$$

# Limitation of Correlation Analysis

**Example:** Contingency tables for the word pairs  $\{p, q\}$  and  $\{r, s\}$

	$p$	$\bar{p}$	
$q$	880	50	930
$\bar{q}$	50	20	70
	930	70	1,000

	$r$	$\bar{r}$	
$s$	20	50	70
$\bar{s}$	50	880	930
	70	930	1,000

$$\phi(p, q) = \frac{(880)(20) - (50)(50)}{\sqrt{(930)(930)(70)(70)}} = 0.232$$

$$\phi(r, s) = \frac{(20)(880) - (50)(50)}{\sqrt{(70)(70)(930)(930)}} = 0.232$$

- The words  $p$  and  $q$  appear together more often than  $r$  and  $s$ , but their  $\phi$ -coefficients are identical.
- This is because the  $\phi$ -coefficient gives equal importance to both co-presence and co-absence of items in a transaction.

# IS Measure

- **IS measure** is an alternative measure that has been proposed for handling asymmetric binary variable.

$$IS(X, Y) = \frac{s(X \cup Y)}{\sqrt{s(X) \times s(Y)}} = \frac{f_{11}}{\sqrt{f_{1+} f_{+1}}}$$

- IS is large when the interest factor and support of the pattern are large.
- **Example:**

$$IS(p, q) = \frac{880}{\sqrt{(930)(930)}} = 0.946$$

$$IS(r, s) = \frac{20}{\sqrt{(70)(70)}} = 0.286$$

The IS measure suggests that the association between  $\{p, q\}$  is stronger than  $\{r, s\}$ .

# IS Measure

- The IS value of an itemset  $\{p, q\}$  is low whenever one of its rules,  $p \rightarrow q$  or  $q \rightarrow p$ , has low confidence.

$$\begin{aligned} IS(X, Y) &= \frac{s(X \cup Y)}{\sqrt{s(X) s(Y)}} \\ &= \sqrt{\frac{s(X \cup Y)}{s(X)} \times \frac{s(X \cup Y)}{s(Y)}} \\ &= \sqrt{\underbrace{c(X \rightarrow Y)} \times \underbrace{c(Y \rightarrow X)}} \end{aligned}$$

- Since the IS value depends on  $c(X \rightarrow Y)$  and  $c(Y \rightarrow X)$ , thus IS also has a similar problem as the confidence measure.

# Limitation of IS Measure

frequent itemset  $X \& Y$

- The IS value for a pair of independent itemsets,  $X$  and  $Y$ , is

$X \rightarrow Y$

$X \cap Y$  must be  $\phi$

$$IS_{indep}(X, Y) = \frac{s(X \cup Y)}{\sqrt{s(X) \times s(Y)}} = \frac{s(X) \times s(Y)}{\sqrt{s(X) \times s(Y)}} = \sqrt{s(X) \times s(Y)}$$

- Since the value depends on  $s(X)$  and  $s(Y)$ , IS has a similar problem as the confidence measure where the value of the measure can be quite large, even for uncorrelated and negatively correlated pattern.

- Example:**

original formula

$$IS(p, q) = \frac{800}{\sqrt{(900)(900)}} = 0.889$$

in theory, these 2 numbers must be the same.

	$q$	$\bar{q}$	
$p$	800	100	900
$\bar{p}$	100	0	100
	900	100	1,000

$$IS_{indep}(p, q) = \sqrt{(0.9)(0.9)} = 0.9$$

# Alternative Objective Interestingness Measures

- The interestingness measures can be divided into **two categories**:

**1. Symmetric measures:** A measure  $M$  is symmetric when

$$M(X \rightarrow Y) = M(Y \rightarrow X)$$

support measure

**2. Asymmetric measures:** A measure  $M$  is asymmetric when

$$M(X \rightarrow Y) \neq M(Y \rightarrow X)$$

confidence measure  
frequent

- Symmetric measures** are generally used for **evaluating itemsets**, while **asymmetric measures** are more suitable for **analyzing association rules**.

## Symmetric objective measures

Measure (Symbol)	Definition
Correlation ( $\phi$ )	$\frac{Nf_{11} - f_{1+}f_{+1}}{\sqrt{f_{1+}f_{+1}f_{0+}f_{+0}}}$
Odds ratio ( $\alpha$ )	$(f_{11}f_{00})/(f_{10}f_{01})$
Kappa ( $\kappa$ )	$\frac{Nf_{11} + Nf_{00} - f_{1+}f_{+1} - f_{0+}f_{+0}}{N^2 - f_{1+}f_{+1} - f_{0+}f_{+0}}$
Interest ( $I$ )	$(Nf_{11})/(f_{1+}f_{+1})$
Cosine ( $IS$ )	$(f_{11})/(\sqrt{f_{1+}f_{+1}})$
Piatetsky-Shapiro ( $PS$ )	$\frac{f_{11}}{N} - \frac{f_{1+}f_{+1}}{N^2}$
Collective strength ( $S$ )	$\frac{f_{11} + f_{00}}{f_{1+}f_{+1} + f_{0+}f_{+0}} \times \frac{N - f_{1+}f_{+1} - f_{0+}f_{+0}}{N - f_{11} - f_{00}}$
Jaccard ( $\varsigma$ )	$f_{11}/(f_{1+} + f_{+1} - f_{11})$
All-confidence ( $h$ )	$\min \left[ \frac{f_{11}}{f_{1+}}, \frac{f_{11}}{f_{+1}} \right]$




## Asymmetric objective measures

Measure (Symbol)	Definition
Goodman-Kruskal ( $\lambda$ )	$\left( \sum_j \max_k f_{jk} - \max_k f_{+k} \right) / \left( N - \max_k f_{+k} \right)$
Mutual Information ( $M$ )	$\left( \sum_i \sum_j \frac{f_{ij}}{N} \log \frac{N f_{ij}}{f_{i+} f_{+j}} \right) / \left( - \sum_i \frac{f_{i+}}{N} \log \frac{f_{i+}}{N} \right)$
J-Measure ( $J$ )	$\frac{f_{11}}{N} \log \frac{N f_{11}}{f_{1+} f_{+1}} + \frac{f_{10}}{N} \log \frac{N f_{10}}{f_{1+} f_{+0}}$
GINI index ( $G$ )	$\frac{f_{1+}}{N} \left[ \left( \frac{f_{11}}{f_{1+}} \right)^2 + \left( \frac{f_{10}}{f_{1+}} \right)^2 \right] - \left( \frac{f_{+1}}{N} \right)^2 + \frac{f_{0+}}{N} \left[ \left( \frac{f_{01}}{f_{0+}} \right)^2 + \left( \frac{f_{00}}{f_{0+}} \right)^2 \right] - \left( \frac{f_{+0}}{N} \right)^2$
Laplace ( $L$ )	$(f_{11} + 1) / (f_{1+} + 2)$
Conviction ( $V$ )	$(f_{1+} f_{+0}) / (N f_{10})$
Certainty factor ( $F$ )	$\left( \frac{f_{11}}{f_{1+}} - \frac{f_{+1}}{N} \right) / \left( 1 - \frac{f_{+1}}{N} \right)$
Add Value ( $AV$ )	$\frac{f_{11}}{f_{1+}} - \frac{f_{+1}}{N}$

# Comparing Different Measure

10 examples of contingency tables:



Example	$f_{11}$	$f_{10}$	$f_{01}$	$f_{00}$
$E_1$	8,123	83	424	1,370
$E_2$	8,330	2	622	1,046
$E_3$	9,481	94	127	298
$E_4$	3,954	3,080	5	2,961
$E_5$	2,886	1,363	1,320	4,431
$E_6$	1,500	2,000	500	6,000
$E_7$	4,000	2,000	1,000	3,000
$E_8$	4,000	2,000	2,000	2,000
$E_9$	1,720	7,121	5	1,154
$E_{10}$	61	2,483	4	7,452

# Comparing Different Measure

## Rankings of contingency tables using various measures:

- Ranks of contingency tables provided by a measure are consistent with ranks obtained from some other measures.

Example, the  $\phi$ -coefficient agree with Kappa and Collective strength.

- A contingency table such as  $E_{10}$  is ranked lowest according to the  $\phi$ -coefficient, but highest according to interest factor.

- Ranks of contingency tables provided by a measure are consistent with ranks obtained from some other measures.

Example, the  $\phi$ -coefficient agree with Kappa and Collective strength.

- A contingency table such as  $E_{10}$  is ranked lowest according to the  $\phi$ -coefficient, but highest according to interest factor.

#	Measure
1	$\phi$ -coefficient
2	Goodman-Kruskal's ( $\lambda$ )
3	Odds ratio ( $\alpha$ )
4	Yule's $Q$
5	Yule's $Y$
6	Kappa ( $\kappa$ )
7	Mutual Information ( $M$ )
8	J-Measure ( $J$ )
9	Gini index ( $G$ )
10	Support ( $s$ )
11	Confidence ( $c$ )
12	Laplace ( $L$ )
13	Conviction ( $V$ )
14	Interest ( $I$ )
15	cosine ( $IS$ )
16	Piatetsky-Shapiro's ( $PS$ )
17	Certainty factor ( $F$ )
18	Added Value ( $AV$ )
19	Collective strength ( $S$ )
20	Jaccard ( $\zeta$ )
21	Klogsen ( $K$ )

# Topics

- ▶ Rule Generation (Subtask 2)
- ▶ Evaluation of Association Rules
- ▶ **Handling Categorical and Continuous Attributes**

# Continuous and Categorical Attributes

- This topic extends the formulation to data sets with **categorical**, and **continuous attributes**.
- The overall structure of association analysis algorithms remains unchanged, but certain aspects of the algorithms must be modified to handle the non-traditional entities.

TID	items
1	milk, bread
2	beer, ...
...	...

Session Id	Country	Session Length (sec)	Number of Web Pages viewed	Gender	Browser Type	Buy
1	USA	982	8	Male	IE	No
2	China	811	10	Female	Netscape	No
3	USA	2125	45	Female	Mozilla	Yes
4	Germany	596	4	Male	IE	Yes
5	Australia	123	9	Male	Mozilla	No
...	...	...	...	...	...	...

## Example of Association Rule:

$\{\text{Number of Pages} \in [5, 10), (\text{Browser} = \text{Mozilla})\} \rightarrow \{\text{Buy} = \text{No}\}$

# Handling Categorical Attributes

- Many data sets contain binary attribute, nominal attribute, and ordinal attribute.
- **Example:** Internet survey data with categorical attributes

Gender	Level of Education	State	Computer at Home	Chat Online	Shop Online	Privacy Concerns
Female	Graduate	Illinois	Yes	Yes	Yes	No
Male	College	California	No	No	No	Yes
Male	Graduate	Michigan	Yes	No	Yes	Yes
Female	High School	Virginia	No	Yes	Yes	No
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Binary attributes  $\Rightarrow$  Gender, Computer at Home, Chat Online, Shop Online, and Privacy Concerns

Nominal attribute  $\Rightarrow$  State

Ordinal attribute  $\Rightarrow$  Level of Education

# Handling Categorical Attributes

- To extract association rules, **the categorical and binary attributes need to be transformed into “items” first**, so that the association rule mining algorithms can be applied.
- It can be done by creating a new “item” for each **distinct attribute-value pair**.
- **Example:** Level of Education attribute can be replaced by 3 items.

- ① Education = College
- ② Education = Graduate
- ③ Education = High School

{ Graduate, College, High School }

# Handling Categorical Attributes

**Example:** Internet survey data after binarizing categorical and symmetric binary attributes

Gender = Male	Gender = Female	Education = College	Education = Graduate	Education = High School	...	Privacy = No
0	1	0	1	0	...	1
1	0	1	0	0	...	0
1	0	0	1	0	...	0
0	1	0	0	1	...	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮

$$\text{Number of newly generated items} = \sum_{i=1}^k N_i$$

where  $k$  denotes number of attributes

$N_i$  denotes number of values of attribute  $i$

Number of newly generated items =  $2 + 3 + 50 + 2 + 2 + 2 + 2 = 63$  items (columns)



# Handling Categorical Attributes

- **Potential issues** to consider when applying association analysis to the binarized data:
  - 1. Some attribute values may not be frequent enough to be part of frequent itemset.**
    - Example: attribute country has more than 200 possible values where many of the attribute values may have very low support
    - Potential solutions:
      - (1) Aggregate the less frequent attribute values into a single category called “Other”.
      - (2) Group related attribute values into a small number of categories.
  - 2. Some attribute values may have considerably higher frequencies than others.**
    - Example: {Computer at home = Yes} = 85%  
This item always be included in several rules making them redundant.
    - Potential solution: Remove such items before applying standard association analysis algorithms.

# Handling Categorical Attributes

**3. Computation time may increase especially when many of the newly created items become frequent.**

- Potential solution: Avoid generating candidate itemsets that contain more than one item from the same attribute
- Example: We do not have to generate a candidate itemset such as  
    {State = Texas, State = Ohio, ...}  
because the support count of the itemset is 0.

# Handling Continuous Attributes

- **Discretization** is the most common approach for handling continuous attributes.
- It groups the adjacent values of a continuous attribute into a finite number of intervals (See Data Preprocessing slides).
- **Discretization techniques:**
  1. Equal width discretization
  2. Equal frequency discretization
  3. k-means discretization
  4. Entropy-based discretization