Outline

- Introduction
- Basic concepts
- Frequent Itemsets Mining (FIM) – Apriori
- Association Rules Mining

Association Rules Mining

- (Recall the) 2-step method to extract the association rules:
  - Determine the frequent itemsets w.r.t. min support s ⟵ FIM problem (Apriori)
  - Generate the association rules w.r.t. min confidence *c.*

- Regarding step 2, the following method is followed:
  - For every frequent itemset X
  - for every subset *Y* of *X: Y≠Ø, Y ≠X,* the rule $Y \rightarrow (X - Y)$ is formed
  - Remove rules that violate min confidence c

$$confidence(Y \rightarrow (X - Y)) = \frac{support\_count(X)}{support\_count(Y)}$$

  - Store the frequent itemsets and their supports in a hash table
    - no database access!

Let *X*={1,2,3} be frequent
There are 6 candidate rules that can be generated from *X*:

- {1,2}→3
- {1,3}→2
- {2,3}→1
- {1}→{2,3}
- {2}→{1,3}
- {3}→{1,2}

To identify strong rules, we can use the support counts (already computed during the FIM step)

Pseudocode

Input:
- D //**Database of transactions**
- I //**Items**
- L //**Large itemsets** (= set of Frequent Itemsets)
- s //**Support** (= minSupport)
- α //**Confidence** (= minConfidence)

Output:
- R //**Association Rules satisfying** s *and* α

ARGen Algorithm:

$R = \emptyset$;
for each $l \in L$ **do**
  for each $x \subset l$ **such that** $x \neq \emptyset$ **and** $x \neq l$ **do**
    if $\frac{support(l)}{support(x)} \geq \alpha$ **then**
      $R = R \cup \{x \Rightarrow (l - x)\}$;

Confidence-based pruning

- How to efficiently generate rules from frequent itemsets?
- Confidence does not follow the monotonicity property
  - i.e., confidence (X→Y) can be >,<,= to confidence (X'→Y'), X' ⊆X, Y' ⊆Y
  - e.g., confidence(ABC→D) can be larger or smaller than confidence(AB →D)

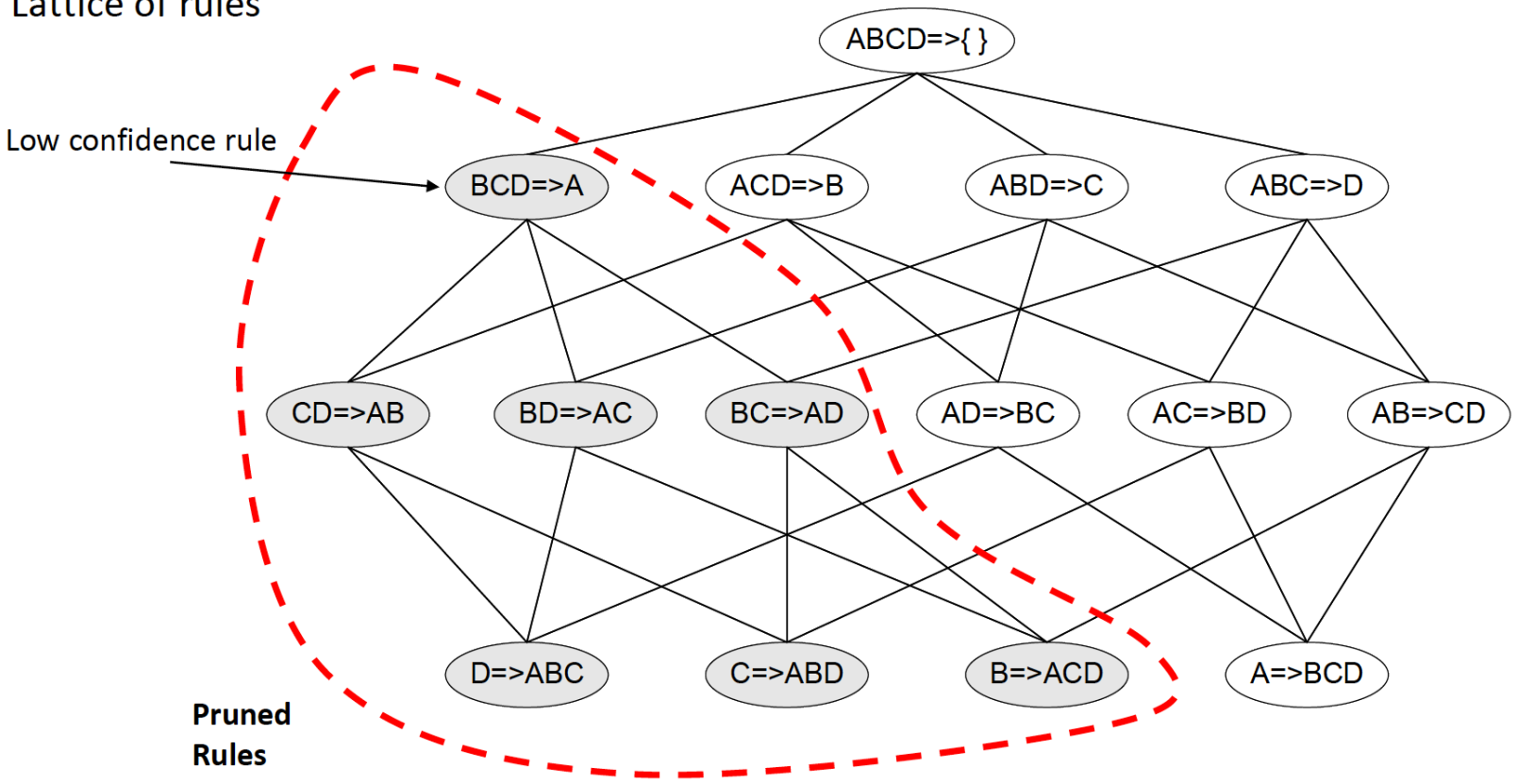
- But the confidence of rules generated from the same itemset does

If rule X→Y-X does not satisfy the minConfidence threshold, then any rule X'→Y-X', where X' ⊆X, must not satisfy the minConfidence threshold as well.

  - For example, for X={ABCD}, then
    - confidence(ABC → D) ≥ confidence(AB → CD) ≥ confidence(A → BCD)

Confidence-based pruning

- Lattice of rules



Example

tid	X_T
1	{Bier, Chips, Wine}
2	{Bier, Chips}
3	{Pizza, Wine}
4	{Chips, Pizza}

Transaction database

I = {Bier, Chips, Pizza, Wine}

Itemset	Cover	Sup.	Freq.
{}	{1,2,3,4}	4	100 %
{Bier}	{1,2}	2	50 %
{Chips}	{1,2,4}	3	75 %
{Pizza}	{3,4}	2	50 %
{Wine}	{1,3}	2	50 %
{Bier, Chips}	{1,2}	2	50 %
{Bier, Wine}	{1}	1	25 %
{Chips, Pizza}	{4}	1	25 %
{Chips, Wine}	{1}	1	25 %
{Pizza, Wine}	{3}	1	25 %
{Bier, Chips, Wine}	{1}	1	25 %

Rule	Sup.	Freq.	Conf.
{Bier} ⇒ {Chips}	2	50 %	100 %
{Bier} ⇒ {Wine}	1	25 %	50 %
{Chips} ⇒ {Bier}	2	50 %	66 %
{Pizza} ⇒ {Chips}	1	25 %	50 %
{Pizza} ⇒ {Wine}	1	25 %	50 %
{Wine} ⇒ {Bier}	1	25 %	50 %
{Wine} ⇒ {Chips}	1	25 %	50 %
{Wine} ⇒ {Pizza}	1	25 %	50 %
{Bier, Chips} ⇒ {Wine}	1	25 %	50 %
{Bier, Wine} ⇒ {Chips}	1	25 %	100 %
{Chips, Wine} ⇒ {Bier}	1	25 %	100 %
{Bier} ⇒ {Chips, Wine}	1	25 %	50 %
{Wine} ⇒ {Bier, Chips}	1	25 %	50 %

Evaluating Association Rules 1/2

Interesting and misleading association rules

Example:

- Database on the behavior of students in a school with 5.000 students
- Itemsets:
  - 60% of the students play Soccer,
  - 75% of the students eat chocolate bars
  - 40% of the students play Soccer and eat chocolate bars
- Association rules: *{"Play Soccer"}* → *{"Eat chocolate bars"}*, confidence = 40%/60%= 67%
  - The rule has a high confidence, however:

    {"Eat chocolate bars"}, support= 75%, regardless of whether they play soccer.

  - Thus, knowing that one is playing soccer decreases his/her probability of eating chocolate (from 75%→67%)
  - Therefore, the rule *{"Play Soccer"}* → *{"Eat chocolate bars"}* is misleading despite its high confidence

Evaluating Association Rules 2/2

Task: Filter out misleading rules

Let $\{A\} \rightarrow \{B\}$

- Measure of "interestingness"-score of a rule:

$$interest = \frac{support(A \cup B)}{support(A)} - support(B)$$

  - the higher the value the more interesting the rule is

- Measure of dependent/correlated events:

$$lift = \frac{support(A \cup B)}{support(A)support(B)}$$

  - the ratio of the *observed* support to that *expected* if X and Y were independent.
  - Lift > 1 means that the rule is interesting, lift < 1 means that the presence of one item has negative effect on presence of other item and vice versa.

Measuring Quality of Association Rules

For a rule A→ B

- Support $support(A \cup B)$ $P(E_A \cap E_B)$ E_X := Event that itemset X appears in a transaction
  - e.g. support(milk, bread, butter)=20%, i.e. 20% of the transactions contain these
- Confidence $\frac{support(A \cup B)}{support(A)}$ $\frac{P(E_A \cap E_B)}{P(E_A)}$
  - e.g. confidence(milk, bread→ butter)=50%, i.e. 50% of the times a customer buys milk and bread, butter is bought as well.
- Lift $\frac{support(A \cup B)}{support(A)support(B)}$ $\frac{P(EA \cap E_B)}{P(E_A)P(EB)}$
  - e.g. lift(milk, bread→ butter)=20%/(40%*40%)=1.25. the observed support is 20%, the expected (if they were independent) is 16%.