Generating association rules

Lecture 14

Mining Association Rules

- Two-step approach:
  1. Frequent Itemset Generation
     - Generate all itemsets whose support $\geq$ minsup (these itemsets are called *frequent itemset*)
  2. Rule Generation
     - Generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset (these rules are called *strong rules*)

We focus on **rule generation from frequent itemsets**.

Rule Generation

- An association rule can be extracted by partitioning a frequent itemset *Y* into two nonempty subsets, X and $Y - X$, such that

$$X \rightarrow Y - X$$

satisfies the confidence threshold.

- Each frequent k-itemset, Y, can produce up to $2^k - 2$ association rules
  - ignoring rules that have empty antecedents or consequents.

Rule Generation

Example

Let Y = {1, 2, 3} be a frequent itemset.

Six candidate association rules can be generated from Y:

{1, 2} → {3},
{1, 3} → {2},
{2, 3} → {1},
{1} → {2, 3},
{2} → {1, 3},
{3} → {1, 2}.

Computing the confidence of an association rule does not require additional scans of the database.

Consider {1, 2} → {3}.

The confidence is **σ ({1, 2, 3}) / σ ({1, 2})**

Because {1, 2, 3} is frequent, the antimonotone property of support ensures that {1, 2} must be frequent, too, and we store the supports of frequent itemsets.

Confidence, unlike support is not anti-monotone:

Knowing that $c(X \text{ -> } Y) < minConfidence$, we cannot tell whether $c(X' \text{ -> } Y') < minConfidence$ or $c(X' \text{ -> } Y') > minConfidence$, for $X' \subseteq X$ and $Y' \subseteq Y$

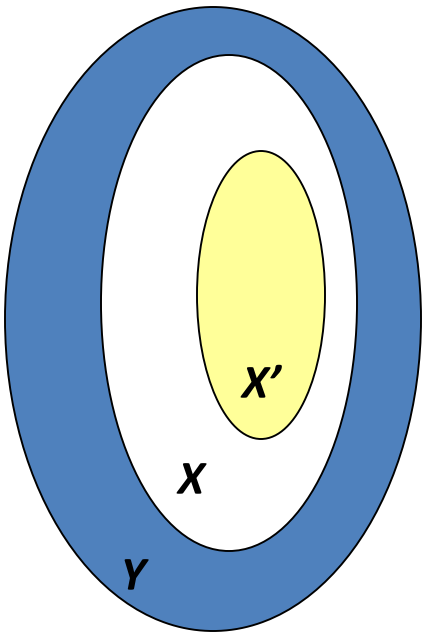
Do we need to compute confidence for all possible rules for each frequent itemset Y?

Confidence-based rule pruning

Theorem.

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, cannot satisfy the confidence threshold as well.



Confidence-based rule pruning

Proof.

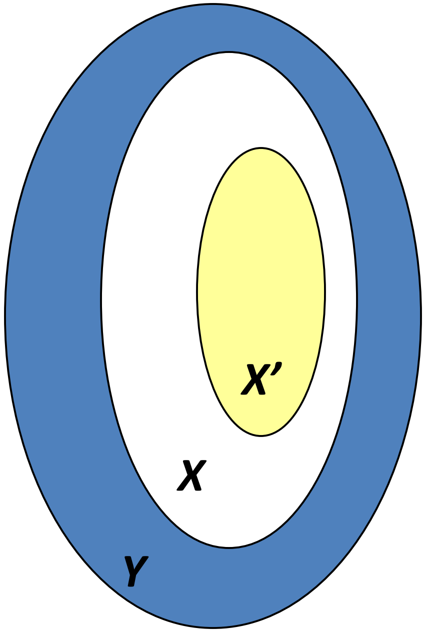
Consider the following two rules:

$X' \rightarrow Y - X'$ and $X \rightarrow Y - X$, where $X' \subseteq X$.

The confidence of the rules are $\sigma(Y) / \sigma(X')$ and $\sigma(Y) / \sigma(X)$, respectively.

Since X' is a subset of X, $\sigma(X') \geq \sigma(X)$.

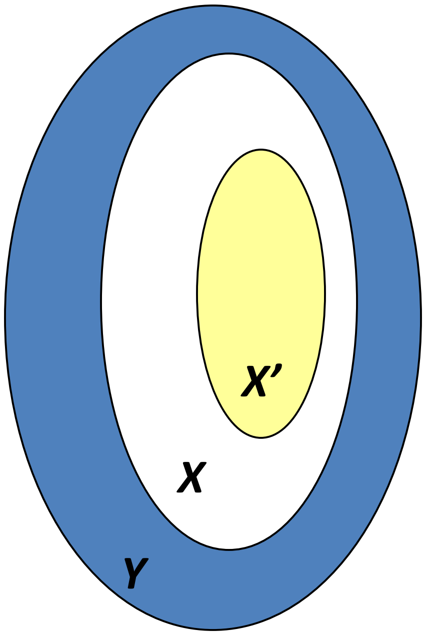
Therefore, the former rule cannot have a higher confidence than the latter rule.

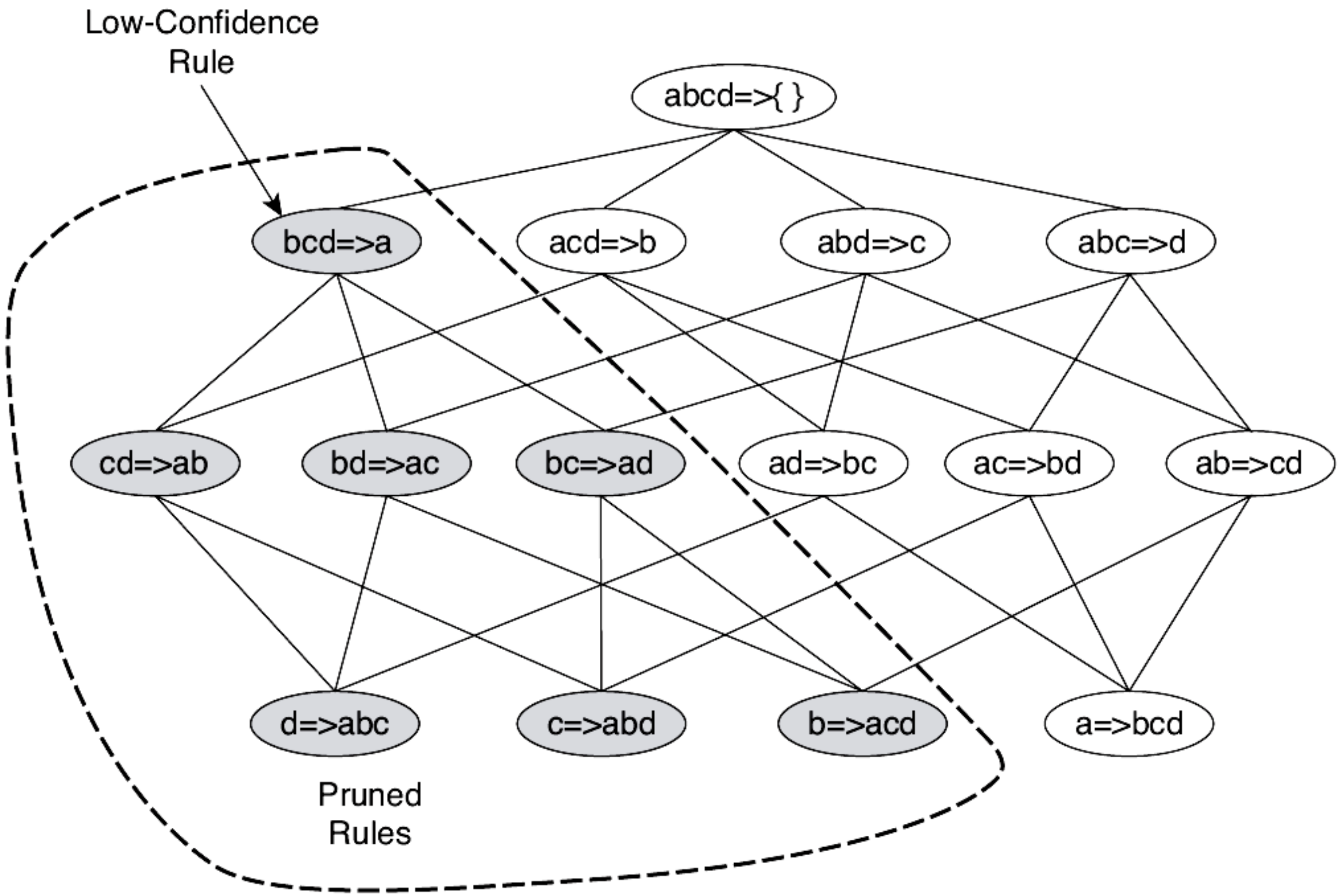


Confidence-Based Pruning

- Observe that:

  $X' \subseteq X$ implies that $Y - X' \supseteq Y - X$





Algorithm for rule generation

- Initially, all the highconfidence rules that have only one item in the rule consequent are extracted.

- These rules are then used to generate new candidate rules.

- For example, if
  - {acd} → {b} and {abd} → {c} are highconfidence rules, then the candidate rule {ad} → {bc} is generated by merging the consequents of both rules.

Example

Item	Count
Bread	**4**
Coke	**2**
Milk	**4**
Beer	**3**
Diaper	**4**
Eggs	**1**

Items (1-itemsets)

Itemset	Count
{Bread,Milk}	**3**
{Bread,Beer}	**2**
{Bread,Diaper}	**3**
{Milk,Beer}	**2**
{Milk,Diaper}	**3**
{Beer,Diaper}	**3**

Pairs (2-itemsets)

Triplets (3-itemsets)

Itemset	Count
{Bread,Milk,Diaper}	**3**

High-confidence rules with 1 item in consequent

{Bread,Milk}→{Diaper} (confidence = 3/3) threshold=50%

{Bread,Diaper}→{Milk} (confidence = 3/3)

{Diaper,Milk}→{Bread} (confidence = 3/3)

Example

Merge:

{Bread,Milk}→{Diaper}

{Bread,Diaper}→{Milk}

{Bread}→{Diaper,Milk} (confidence = 3/4)

…