Association Rule

Jun Du

The University of Western Ontario

jdu43@uwo.ca

Outline

- Association Rule Overview
- Association Rule Algorithm
  - Frequent Itemset Generation
  - Rule Generation
- Association Rule Evaluation
- Summary

Association Rule Mining

- Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction

Market-Basket Transactions

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example of Association Rules

{Diaper} → {Beer},
{Milk, Bread} → {Eggs,Coke},
{Beer, Bread} → {Milk},

Implication means co-occurrence, not causality!

Definition: Frequent Itemset

- Itemset
  - A collection of one or more items
    - Example: {Milk, Bread, Diaper}
  - k-itemset
    - An itemset that contains k items
- Support count (σ)
  - Frequency of occurrence of an itemset
  - E.g. σ({Milk, Bread, Diaper}) = 2
- Support
  - Fraction of transactions that contain an itemset
  - E.g. s({Milk, Bread, Diaper}) = 2/5
- Frequent Itemset
  - An itemset whose support is greater than or equal to a *minsup* threshold

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Definition: Association Rule

- Association Rule
  - An implication expression of the form $X \rightarrow Y$, where X and Y are itemsets
  - Example: {Milk, Diaper} $\rightarrow$ {Beer}
- Rule Evaluation Metrics
  - Support (s)
    - Fraction of transactions that contain both X and Y
    - Coverage of the rule
  - Confidence (c)
    - Measures how often items in Y appear in transactions that contain X
    - Accuracy of the rule

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example:

$$\{\text{Milk, Diaper}\} \Rightarrow \text{Beer}$$

$$s = \frac{\sigma(\text{Milk, Diaper, Beer})}{|\text{T}|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk, Diaper, Beer})}{\sigma(\text{Milk, Diaper})} = \frac{2}{3} = 0.67$$

Association Rule Mining Task

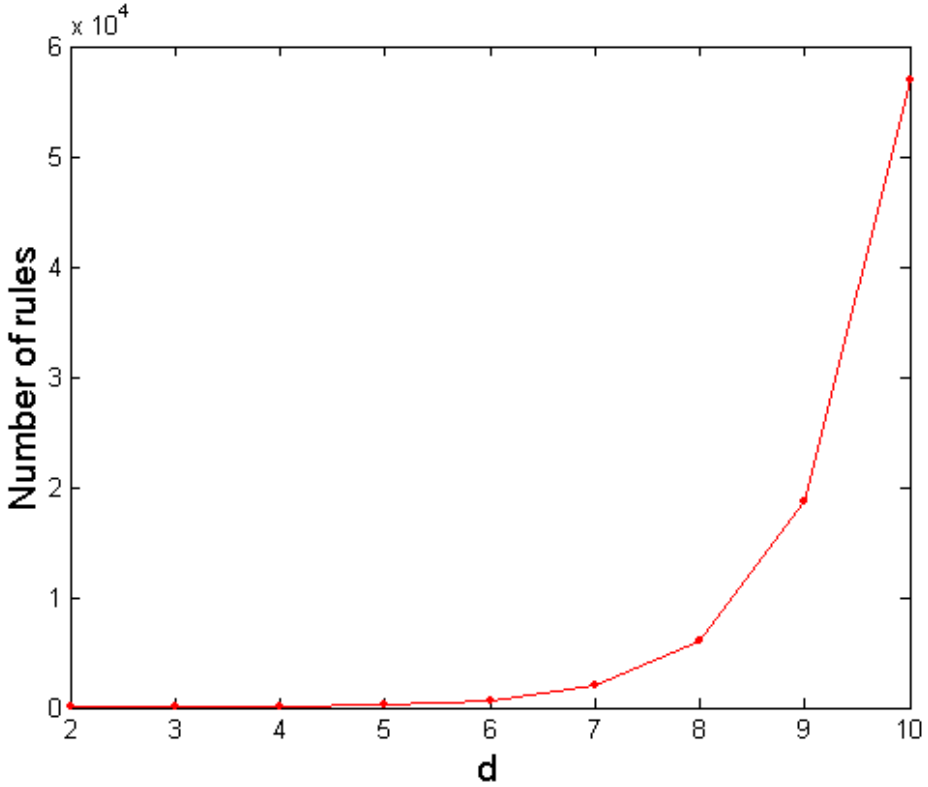
- Given a set of transactions *T* and 2 pre-set parameters *minsup*, *minconf*, the goal of association rule mining is to find all rules having
  - support ≥ *minsup* threshold
  - confidence ≥ *minconf* threshold

- Brute-force approach:
  - List all possible association rules
  - Compute the support and confidence for each rule
  - Prune rules that fail the *minsup* and *minconf* thresholds

  ⇒ Computationally prohibitive!

Computational Complexity

- Given *d* unique items:
  - Total number of itemsets = 2^d
  - Total number of possible association rules:



$$R = \sum_{k=1}^{d-1}\left[\binom{d}{k} \times \sum_{j=1}^{d-k}\binom{d-k}{j}\right]$$

$$= 3^d - 2^{d+1} + 1$$

If d=6, R = 602 rules

Outline

- Association Rule Overview
- Association Rule Algorithm
  - Frequent Itemset Generation
  - Rule Generation
- Association Rule Evaluation
- Summary

Mining Association Rules

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example of Rules:

{Milk,Diaper} → {Beer} (s=0.4, c=0.67)
{Milk,Beer} → {Diaper} (s=0.4, c=1.0)
{Diaper,Beer} → {Milk} (s=0.4, c=0.67)
{Beer} → {Milk,Diaper} (s=0.4, c=0.67)
{Diaper} → {Milk,Beer} (s=0.4, c=0.5)
{Milk} → {Diaper,Beer} (s=0.4, c=0.5)

Observations:

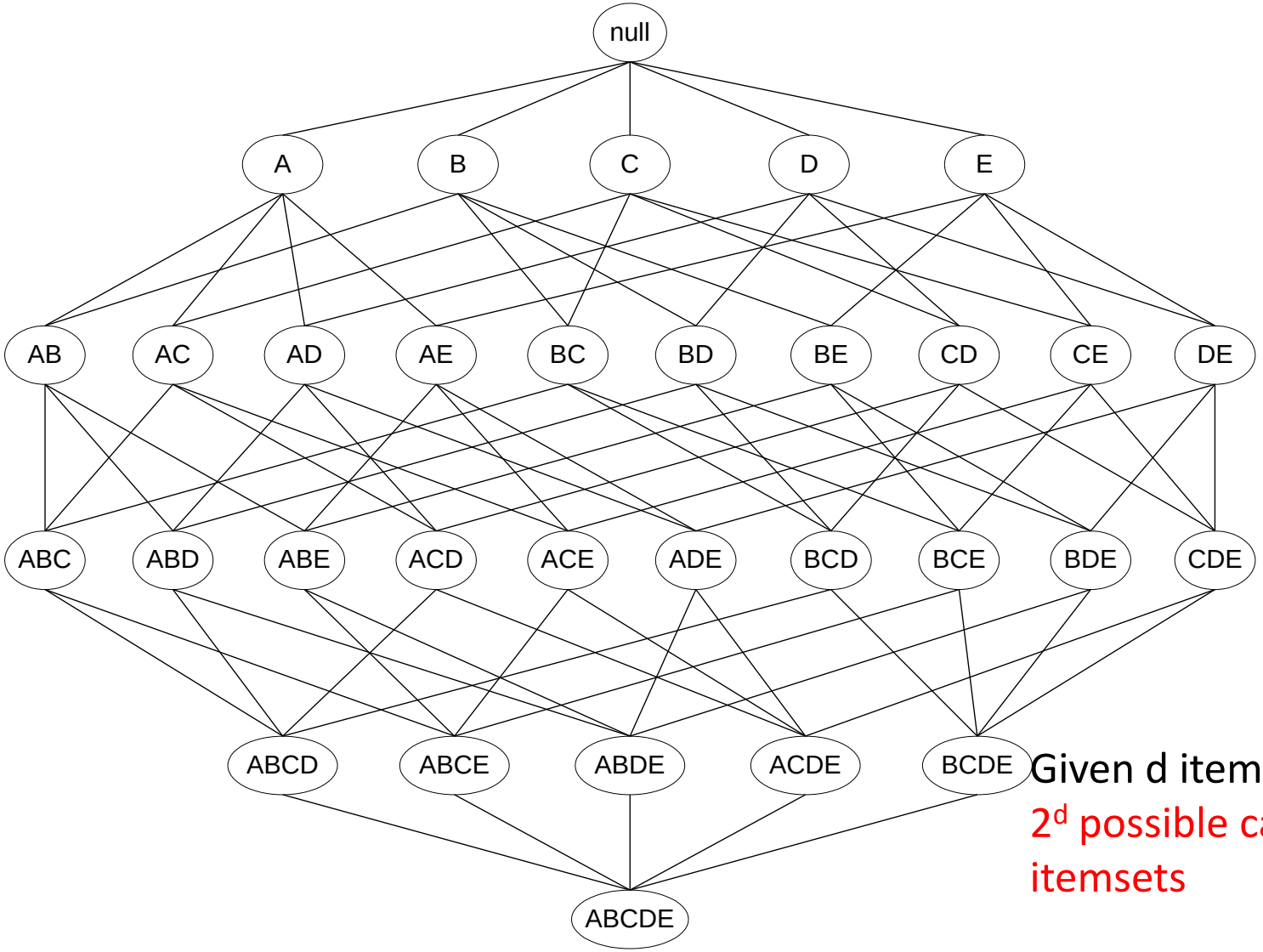
- All the above rules are binary partitions of the same itemset:
  {Milk, Diaper, Beer}
- Rules originating from the same itemset have identical support but can have different confidence
- Thus, we may decouple the support and confidence requirements

Mining Association Rules

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

- Frequent itemset generation is still computationally expensive

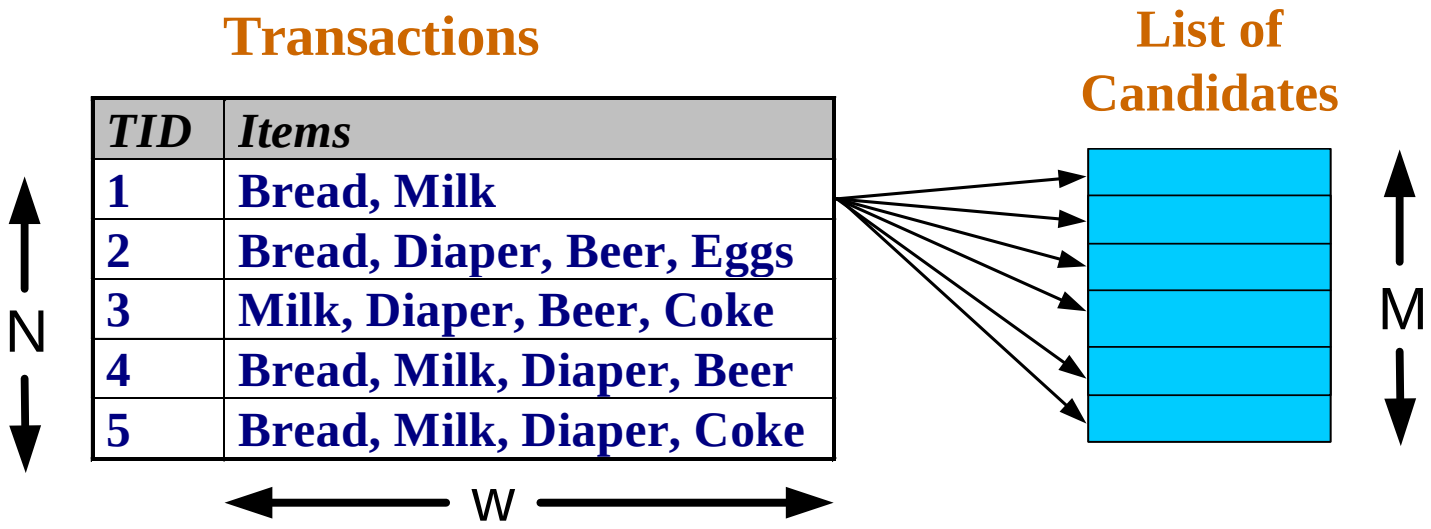
Frequent Itemset Generation



Given d items, there are 2^d possible candidate itemsets

Frequent Itemset Generation

- Brute-force approach:
  - Each itemset is a candidate frequent itemset
  - Count the support of each candidate by scanning the database



  - Match each transaction against every candidate
  - Complexity ~ O(NMw) => Expensive since $M = 2^d$!!!

Frequent Itemset Generation Strategies

- Reduce the number of candidates (*M*)
  - Complete search: $M=2^d$
  - Use pruning techniques to reduce M

- Reduce the number of transactions (*N*)
  - Reduce size of *N* as the size of itemset increases
  - No need to check every transaction for a given candidate itemset

- Reduce the number of comparisons (NM)
  - Use efficient data structures (hash tree, etc.) to store the candidates or transactions
  - No need to match every candidate against every transaction

Reducing Number of Candidates

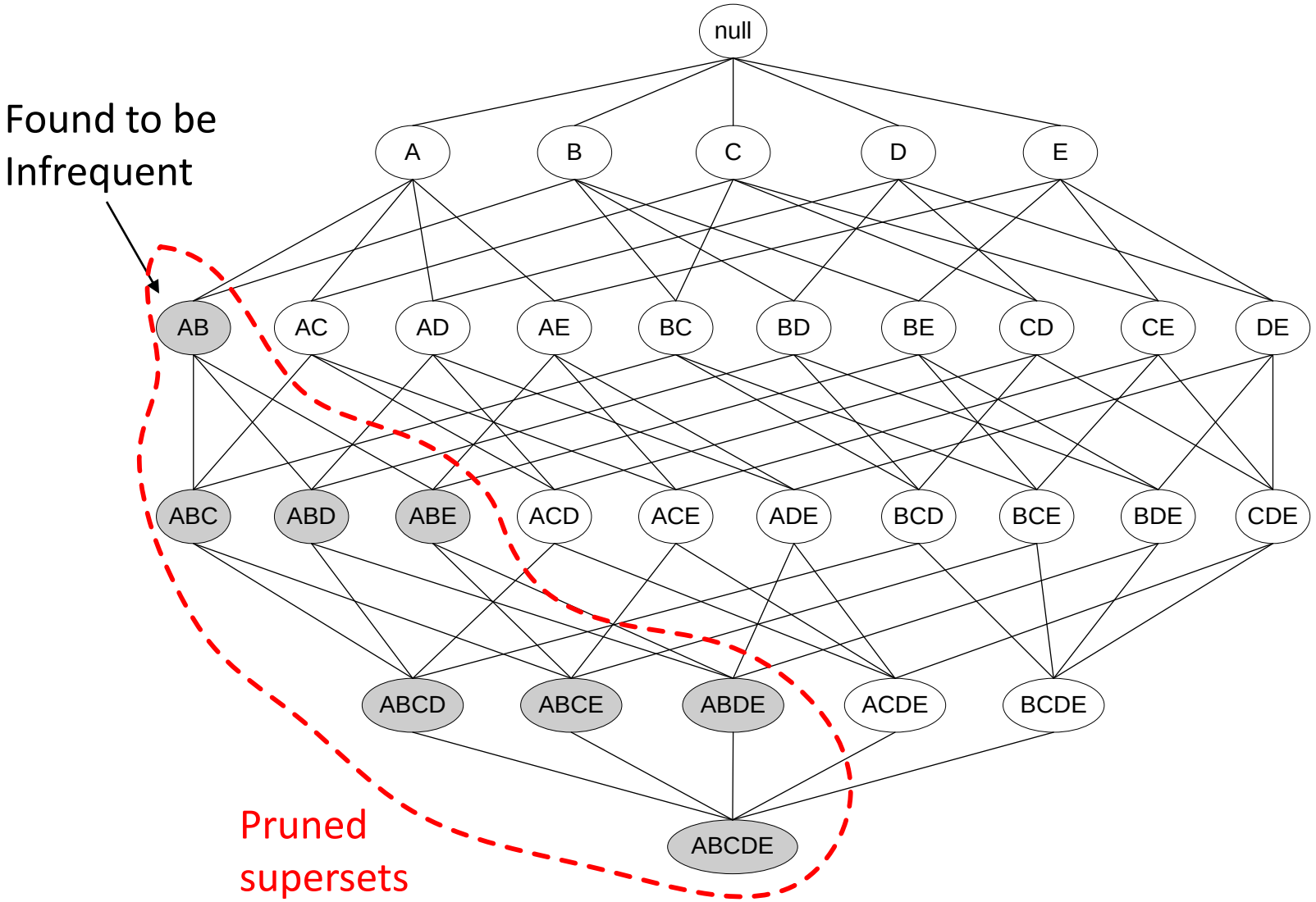
- Apriori principle:
  - If an itemset is frequent, then all of its subsets must also be frequent
  - Equivalently, If an itemset is infrequent, then all of its supersets must also be infrequent

- Apriori principle holds due to the following property of the support measure:

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

  - Called anti-monotone property of support
  - If *s(X) < minsup*, then *s(Y) <= s(X) < minsup*

Illustrating Apriori Principle



Illustrating Apriori Principle

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)

(No need to generate candidates involving Coke or Eggs)

Minimum Support = 3

Triplets (3-itemsets)

Itemset	Count
{Bread,Milk,Diaper}	3

If every subset is considered,

$$\binom{6}{1}+\binom{6}{2}+\binom{6}{3}=41$$

With support-based pruning,

$$\binom{6}{1}+\binom{4}{2}+1=6+6+1=13$$

Apriori Algorithm

- Let *k=1*
- Generate frequent itemsets of length *1*
- Repeat until no new frequent itemsets are identified
  - Generate length *(k+1)* candidate itemsets from length k frequent itemsets
  - Prune candidate itemsets containing subsets of length *k* that are infrequent
  - Count the support of each candidate by scanning the data
  - Eliminate candidates that are infrequent, leaving only those that are frequent

Factors Affecting Complexity

- Choice of minimum support threshold
  - lower support threshold results in more frequent itemsets, and consequently expensive computation
  - higher support threshold can miss itemsets involving interesting rare items (e.g., expensive products)
- Dimensionality (number of items) of the data set
  - more space is needed to store support count of each item
- Size of database
  - run time of algorithm may increase with number of transactions
- Average transaction width
  - may increase max length of frequent itemsets (number of subsets in a transaction increases with its width)

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:

ABC →D,	ABD →C,	ACD →B,	BCD →A,
A →BCD,	B →ACD,	C →ABD,	D →ABC
AB →CD,	AC → BD,	AD → BC,	BC →AD,
BD →AC,	CD →AB		

- For k-itemset, then there are $2^k - 2$ candidate association rules (ignoring $L \rightarrow \varnothing$ and $\varnothing \rightarrow L$)

- How to efficiently generate rules from frequent itemsets?

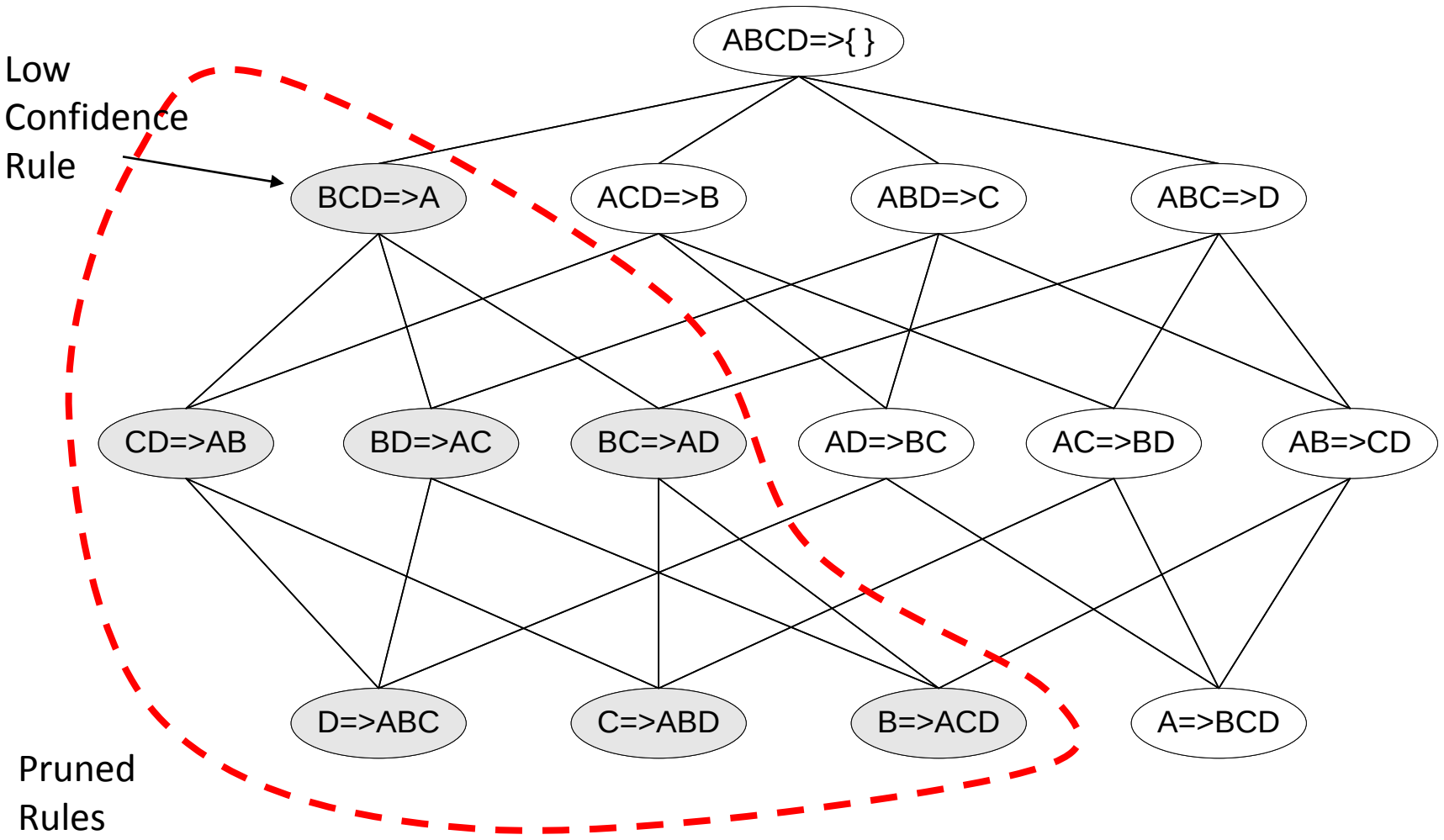
Rule Generation

- Recall: anti-monotone property of support
  - Given itemset *X, Y* $\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$
- Do we have similar property for confidence??
  - Can we guarantee: c(ABC $\rightarrow$D) $\geq$ c(AB $\rightarrow$D) or c(ABC $\rightarrow$D) $\leq$ c(AB $\rightarrow$D)?
  - NO!!
- However, confidence of rules generated from the same itemset has an anti-monotone property
  - E.g., L = {A,B,C,D} (σ(ABCD): # transactions containing A, B, C and D)

$$c(ABC \rightarrow D) = \frac{\sigma(ABCD)}{\sigma(ABC)} \quad c(AB \rightarrow CD) = \frac{\sigma(ABCD)}{\sigma(AB)} \quad c(A \rightarrow BCD) = \frac{\sigma(ABCD)}{\sigma(A)}$$

  - Guaranteed: c(ABC $\rightarrow$ D) $\geq$ c(AB $\rightarrow$ CD) $\geq$ c(A $\rightarrow$ BCD)
  - If c(ABC $\rightarrow$ D) < mincon, no need to check (AB $\rightarrow$ CD) and c(A $\rightarrow$ BCD)

Rule Generation for Apriori Algorithm



Outline

- Association Rule Overview
- Association Rule Algorithm
  - Frequent Itemset Generation
  - Rule Generation
- Association Rule Evaluation
- Summary

Evaluation of Association Rules

- Association rule algorithms tend to produce too many rules
  - many of them are "uninteresting"

- Measures of Interestingness
  - Subjective interestingness measure
  - Objective interestingness measure

Subjective Interestingness Measure

- A rule is considered subjectively uninteresting

  unless it

  - reveals unexpected information about the data, or
  - provides useful knowledge that can lead to profitable actions.

- E.g.,
  - {Butter} → {Bread}, not interesting
  - {Diaper} → {Beer}, interesting

- Defined based on domain information

Objective Interestingness Measure

- Without domain knowledge, how can we know some rules are not good (even with high confidence and support)?

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}	|T|

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

Used to define various measures

- Traditionally: support, confidence
- New: lift, etc.

Drawback of Confidence

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

Association Rule:
Tea $\rightarrow$ Coffee

Confidence = σ(Coffee, Tea) / σ(Tea) = P(Coffee|Tea)

= 15/20 = 0.75

but P(Coffee) = 90 / 100 = 0.9

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

A better rule: $\overline{\text{Tea}} \rightarrow$ Coffee

$\Rightarrow$ Confidence = σ(Coffee, $\overline{\text{Tea}}$) / σ($\overline{\text{Tea}}$) = P(Coffee|$\overline{\text{Tea}}$) = 75 / 80 = 0.9375

Correlation Measure

- Support and confidence measures are insufficient at filtering out uninteresting association rules
- Correlation measure is introduced to augment the framework
  - Given a rule A → B, the correlation between itemsets A and B is measured
- Many different correlation measures
  - Lift/Interest, χ^2, ϕ-coefficient, etc.
  - Most are based on statistical independence

Recall: Statistical Independence

- 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)

  - $P(S,B) = 420/1000 = 0.42$
  - $P(S) \times P(B) = 0.6 \times 0.7 = 0.42$

  - $P(S,B) = P(S) \times P(B)$ => Statistical independence
  - $P(S,B) > P(S) \times P(B)$ => Positively correlated
  - $P(S,B) < P(S) \times P(B)$ => Negatively correlated

Lift / Interest

- Given a rule A → B $$Lift = \frac{P(A,B)}{P(A)P(B)} = \frac{P(B|A)}{P(B)}$$

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

Association Rule:
Tea → Coffee

Confidence = P(Coffee|Tea) = 0.75, but P(Coffee) = 0.9

Lift = P(Coffee|Tea) / P(Coffee) = 0.75/0.9= 0.8333

→ < 1, therefore is negatively associated

→ thus should not be considered as interesting

Other Measures

symbol	measure	range	formula
ϕ	ϕ-coefficient	-1...1	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
Q	Yule's Q	-1 ...1	$\frac{P(A,B)P(\overline{A},\overline{B})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{A},\overline{B})+P(A,\overline{B})P(\overline{A},B)}$
Y	Yule's Y	-1 ...1	$\frac{\sqrt{P(A,B)P(\overline{A},\overline{B})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{A},\overline{B})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}}$
k	Cohen's	-1 ...1	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
PS	Piatetsky-Shapiro's	-0.25 ...0.25	$P(A,B)-P(A)P(B)$
F	Certainty factor	-1 ...1	$\max(\frac{P(B\|A)-P(B)}{1-P(B)}, \frac{P(A\|B)-P(A)}{1-P(A)})$
AV	added value	-0.5 ...1	$\max(P(B\|A)-P(B), P(A\|B)-P(A))$
K	Klosgen's Q	-0.33 ...0.38	$\sqrt{P(A,B)}\max(P(B\|A)-P(B), P(A\|B)-P(A))$
g	Goodman-kruskal's	0 ...1	$\frac{\Sigma_j \max_k P(A_j,B_k)+\Sigma_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-max_k P(B_k)}$
M	Mutual Information	0 ...1	$\frac{\Sigma_i\Sigma_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_J)}}{\min(-\Sigma_i P(A_i)\log P(A_i)\log P(A_i), -\Sigma_i P(B_i)\log P(B_i)\log P(B_i))}$
J	J-Measure	0 ...1	$\max(P(A,B)\log(\frac{P(B\|A)}{P(B)}) + P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}))$
			$P(A,B)\log(\frac{P(A\|B)}{P(A)}) + P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})$
G	Gini index	0 ...1	$\max(P(A)[P(B\|A)^2 + P(\overline{B}\|A)^2] + P(\overline{A}[P(B\|\overline{A})^2 + P(\overline{B}\|\overline{A})^2] - P(B)^2 - P(\overline{B})^2,$
			$P(B)[P(A\|B)^2 + P(\overline{A}\|B)^2] + P(\overline{B}[P(A\|\overline{B})^2 + P(\overline{A}\|\overline{B})^2] - P(A)^2 - P(\overline{A})^2)$
s	support	0 ...1	$P(A,B)$
c	confidence	0 ...1	$max(P(B\|A), P(A\|B))$
L	Laplace	0 ...1	$\max(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2})$
IS	Cosine	0 ...1	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
γ	coherence(Jaccard)	0 ...1	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
α	all_confidence	0 ...1	$\frac{P(A,B)}{\max(P(A),P(B))}$
o	odds ratio	0 ...∞	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(\overline{A},B)P(A,\overline{B})}$
V	Conviction	0.5 ...∞	$\max(\frac{P(A)P(\overline{B})}{P(A\overline{B})}, \frac{P(B)P(\overline{A})}{P(B\overline{A})})$
λ	lift	0 ...∞	$\frac{P(A,B)}{P(A)P(B)}$
S	Collective strength	0 ...∞	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})} \times \frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
χ^2	χ^2	0 ...∞	$\Sigma_i \frac{(P(A_i)-E_i)^2}{E_i}$

Outline

- Association Rule Overview
- Association Rule Algorithm
  - Frequent Itemset Generation
  - Rule Generation
- Association Rule Evaluation
- Summary

Summary

- Basic concepts:
  - Frequent item set
  - association rules
  - support-confident framework
- Algorithm
  - Apriori (Candidate generation & test)
- Evaluation
  - Weakness of support-confident framework
  - Correlation measure (Lift)

Demonstration

- Apriori
  - Weka\weather.nominal
    - Default setting
    - Set class = “play”
    - Increase “lowerBoundMinSupport”
    - Decrease “minMeric” (Confidence)
    - lowBoundMinSupport = 0.2; minMetric = 0.7
- Association rule in real-world application