Data Mining
Association Rules: Advanced Concepts and Algorithms

Lecture Notes for Chapter 7

Introduction to Data Mining

by

Tan, Steinbach, Kumar

Continuous and Categorical Attributes

How to apply association analysis formulation to variables other than asymmetric binary variables?

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:

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

Handling Categorical Attributes

- Transform categorical attribute into asymmetric binary variables
- Introduce a new "item" for each distinct attribute-value pair
  - Example: replace Browser Type attribute with
    - Browser Type = Internet Explorer
    - Browser Type = Mozilla
    - Browser Type = Mozilla

Handling Categorical Attributes

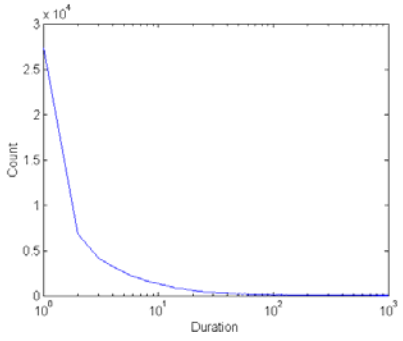
- Potential Issues
  - What if attribute has many possible values
    - Example: attribute country has more than 200 possible values
    - Many of the attribute values may have very low support
      - Potential solution: Aggregate the low-support attribute values
  - What if distribution of attribute values is highly skewed
    - Example: 95% of the visitors have Buy = No
    - Most of the items will be associated with (Buy=No) item
      - Potential solution: drop the highly frequent items

Handling Continuous Attributes

- Different kinds of rules:
  - Age$\in$[21,35) $\wedge$ Salary$\in$[70k,120k) $\rightarrow$ Buy
  - Salary$\in$[70k,120k) $\wedge$ Buy $\rightarrow$ Age: μ=28, σ=4
- Different methods:
  - Discretization-based
  - Statistics-based
  - Non-discretization based
    - minApriori

Handling Continuous Attributes

- Use discretization
- Unsupervised:
  - Equal-width binning
  - Equal-depth binning
  - Clustering
- Supervised:



Attribute values, v

Class	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9
Anomalous	0	0	20	10	20	0	0	0	0
Normal	150	100	0	0	0	100	100	150	100

bin_1 bin_2 bin_3

Discretization Issues

- Size of the discretized intervals affect support & confidence

  {Refund = No, (Income = $51,250)} → {Cheat = No}

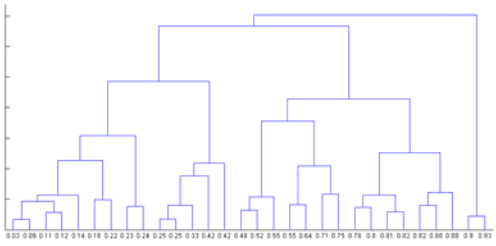
  {Refund = No, (60K ≤ Income ≤ 80K)} → {Cheat = No}

  {Refund = No, (0K ≤ Income ≤ 1B)} → {Cheat = No}

  - If intervals too small
    - may not have enough support
  - If intervals too large
    - may not have enough confidence
- Potential solution: use all possible intervals

Discretization Issues

- Execution time
  - If intervals contain n values, there are on average $O(n^2)$ possible ranges



- Too many rules

  {Refund = No, (Income = $51,250)} → {Cheat = No}

  {Refund = No, (51K ≤ Income ≤ 52K)} → {Cheat = No}

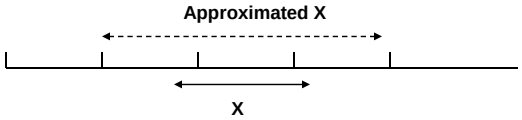
  {Refund = No, (50K ≤ Income ≤ 60K)} → {Cheat = No}

Approach by Srikant & Agrawal

- Preprocess the data
  - Discretize attribute using equi-depth partitioning
    - Use *partial completeness measure* to determine number of partitions
    - Merge adjacent intervals as long as support is less than max-support
- Apply existing association rule mining algorithms
- Determine interesting rules in the output

Approach by Srikant & Agrawal

- Discretization will lose information



  - Use *partial completeness measure* to determine how much information is lost

C: frequent itemsets obtained by considering all ranges of attribute values
P: frequent itemsets obtained by considering all ranges over the partitions

P is *K-complete* w.r.t C if $P \subseteq C$,and $\forall X \in C$, $\exists X' \in P$ such that:

1. X' is a generalization of X and support (X') $\leq K \times$ support(X) $(K \geq 1)$
2. $\forall Y \subseteq X$, $\exists Y' \subseteq X'$ such that support (Y') $\leq K \times$ support(Y)

Given *K (partial completeness level),* can determine number of intervals (N)

Interestingness Measure

{Refund = No, (Income = $51,250)} → {Cheat = No}

{Refund = No, (51K ≤ Income ≤ 52K)} → {Cheat = No}

{Refund = No, (50K ≤ Income ≤ 60K)} → {Cheat = No}

- Given an itemset: $Z = \{z_1, z_2, \ldots, z_k\}$ and its generalization $Z' = \{z_1', z_2', \ldots, z_k'\}$

  P(Z): support of Z

  $E_{Z'}(Z)$: expected support of Z based on Z'

$$E_{z'}(Z) = \frac{P(z_1)}{P(z_1')} \times \frac{P(z_2)}{P(z_2')} \times \cdots \times \frac{P(z_k)}{P(z_k')} \times P(Z')$$

  - Z is R-interesting w.r.t. Z' if $P(Z) \geq R \times E_{z'}(Z)$

Interestingness Measure

- For S: X → Y, and its generalization S': X' → Y'

  P(Y|X): confidence of X → Y

  P(Y'|X'): confidence of X' → Y'

  $E_{S'}(Y|X)$: expected support of Z based on Z'

$$E(Y \mid X) = \frac{P(y_1)}{P(y_1')} \times \frac{P(y_2)}{P(y_2')} \times \cdots \times \frac{P(y_k)}{P(y_k')} \times P(Y' \mid X')$$

- Rule S is R-interesting w.r.t its ancestor rule S' if
  - Support, $P(S) \geq R \times E_{s'}(S)$ or
  - Confidence, $P(Y|X) \geq R \times E_{s'}(Y|X)$

Statistics-based Methods

- Example:

  Browser=Mozilla $\wedge$ Buy=Yes $\rightarrow$ Age: μ=23

- Rule consequent consists of a continuous variable, characterized by their statistics
  - mean, median, standard deviation, etc.
- Approach:
  - Withhold the target variable from the rest of the data
  - Apply existing frequent itemset generation on the rest of the data
  - For each frequent itemset, compute the descriptive statistics for the corresponding target variable
    - Frequent itemset becomes a rule by introducing the target variable as rule consequent
  - Apply statistical test to determine interestingness of the rule

Statistics-based Methods

- How to determine whether an association rule interesting?
  - Compare the statistics for segment of population covered by the rule vs segment of population not covered by the rule:

    $A \Rightarrow B: \mu$ versus $\bar{A} \Rightarrow B: \mu'$

  - Statistical hypothesis testing:
    - Null hypothesis: H0: $\mu' = \mu + \Delta$
    - Alternative hypothesis: H1: $\mu' > \mu + \Delta$
    - Z has zero mean and variance 1 under null hypothesis

$$Z = \frac{\mu' - \mu - \Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Statistics-based Methods

- Example:

  r: Browser=Mozilla $\wedge$ Buy=Yes $\rightarrow$ Age: μ=23

  - Rule is interesting if difference between μ and μ' is greater than 5 years (i.e., Δ = 5)
  - For r, suppose n1 = 50, s1 = 3.5
  - For r' (complement): n2 = 250, s2 = 6.5

$$Z = \frac{\mu' - \mu - \Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{30 - 23 - 5}{\sqrt{\frac{3.5^2}{50} + \frac{6.5^2}{250}}} = 3.11$$

  - For 1-sided test at 95% confidence level, critical Z-value for rejecting null hypothesis is 1.64.
  - Since Z is greater than 1.64, r is an interesting rule

Min-Apriori (Han et al)

Document-term matrix:

TID	W1	W2	W3	W4	W5
D1	2	2	0	0	1
D2	0	0	1	2	2
D3	2	3	0	0	0
D4	0	0	1	0	1
D5	1	1	1	0	2

Example:

W1 and W2 tends to appear together in the same document

Min-Apriori

- Data contains only continuous attributes of the same "type"
  - e.g., frequency of words in a document

TID	W1	W2	W3	W4	W5
D1	2	2	0	0	1
D2	0	0	1	2	2
D3	2	3	0	0	0
D4	0	0	1	0	1
D5	1	1	1	0	2

- Potential solution:
  - Convert into 0/1 matrix and then apply existing algorithms
    - lose word frequency information
  - Discretization does not apply as users want association among words not ranges of words

Min-Apriori

- How to determine the support of a word?
  - If we simply sum up its frequency, support count will be greater than total number of documents!
    - Normalize the word vectors – e.g., using L_1 norm
    - Each word has a support equals to 1.0

TID	W1	W2	W3	W4	W5
D1	2	2	0	0	1
D2	0	0	1	2	2
D3	2	3	0	0	0
D4	0	0	1	0	1
D5	1	1	1	0	2

Normalize →

TID	W1	W2	W3	W4	W5
D1	0.40	0.33	0.00	0.00	0.17
D2	0.00	0.00	0.33	1.00	0.33
D3	0.40	0.50	0.00	0.00	0.00
D4	0.00	0.00	0.33	0.00	0.17
D5	0.20	0.17	0.33	0.00	0.33

Min-Apriori

- New definition of support:

$$\sup(C) = \sum_{i \in T} \min_{j \in C} D(i, j)$$

TID	W1	W2	W3	W4	W5
D1	0.40	0.33	0.00	0.00	0.17
D2	0.00	0.00	0.33	1.00	0.33
D3	0.40	0.50	0.00	0.00	0.00
D4	0.00	0.00	0.33	0.00	0.17
D5	0.20	0.17	0.33	0.00	0.33

Example:

Sup(W1,W2,W3)

= 0 + 0 + 0 + 0 + 0.17

= 0.17

Anti-monotone property of Support

TID	W1	W2	W3	W4	W5
D1	0.40	0.33	0.00	0.00	0.17
D2	0.00	0.00	0.33	1.00	0.33
D3	0.40	0.50	0.00	0.00	0.00
D4	0.00	0.00	0.33	0.00	0.17
D5	0.20	0.17	0.33	0.00	0.33

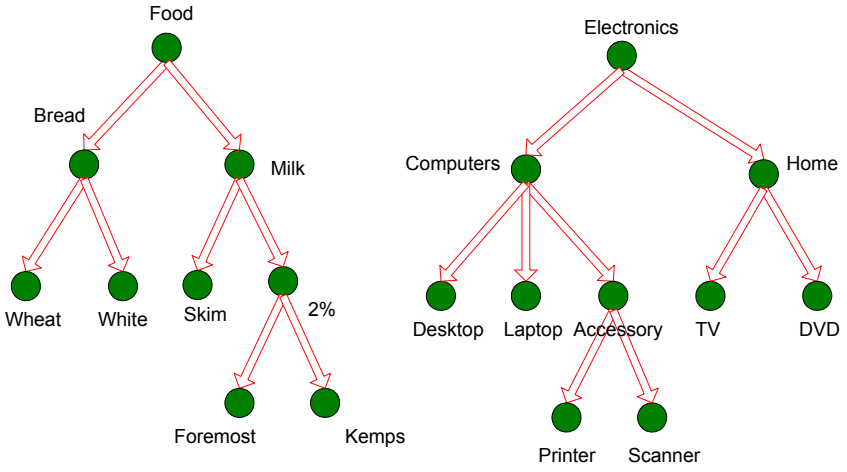
Example:

Sup(W1) = 0.4 + 0 + 0.4 + 0 + 0.2 = 1

Sup(W1, W2) = 0.33 + 0 + 0.4 + 0 + 0.17 = 0.9

Sup(W1, W2, W3) = 0 + 0 + 0 + 0 + 0.17 = 0.17

Multi-level Association Rules



Multi-level Association Rules

- Why should we incorporate concept hierarchy?
  - Rules at lower levels may not have enough support to appear in any frequent itemsets
  - Rules at lower levels of the hierarchy are overly specific
    - e.g., skim milk → white bread, 2% milk → wheat bread, skim milk → wheat bread, etc.

    are indicative of association between milk and bread

Multi-level Association Rules

- How do support and confidence vary as we traverse the concept hierarchy?
  - If X is the parent item for both X1 and X2, then $\sigma(X) \leq \sigma(X1) + \sigma(X2)$
  - If $\sigma(X1 \cup Y1) \geq$ minsup,
    and X is parent of X1, Y is parent of Y1
    then $\sigma(X \cup Y1) \geq$ minsup, $\sigma(X1 \cup Y) \geq$ minsup
    $\sigma(X \cup Y) \geq$ minsup
  - If $\text{conf}(X1 \Rightarrow Y1) \geq$ minconf,
    then $\text{conf}(X1 \Rightarrow Y) \geq$ minconf

Multi-level Association Rules

- Approach 1:
  - Extend current association rule formulation by augmenting each transaction with higher level items

  Original Transaction: {skim milk, wheat bread}

  Augmented Transaction:
  {skim milk, wheat bread, milk, bread, food}

- Issues:
  - Items that reside at higher levels have much higher support counts
    - if support threshold is low, too many frequent patterns involving items from the higher levels
  - Increased dimensionality of the data

Multi-level Association Rules

- Approach 2:
  - Generate frequent patterns at highest level first
  - Then, generate frequent patterns at the next highest level, and so on
- Issues:
  - I/O requirements will increase dramatically because we need to perform more passes over the data
  - May miss some potentially interesting cross-level association patterns