Data Mining
Assignment 3

October 9, 2025

Viet Huy Duong
811433146

---

1. Table 1 shows a transaction database. Please answer the following questions. (10 pts)

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

Table 1. A Transaction Database

   (a) Please enumerate all association rules involving 3 items $\{Bread, Milk, Coke\}$.

   (b) Please compute the support and confidence of each association rule you listed in a.

(a) Please enumerate all association rules involving 3 items Bread, Milk, Coke. The association rules involve exactly 3 items Bread (i will call: B), Milk (i will call: M), Coke (i will call: C),with the antecedent and consequent parts combined being this non-empty set:

- $B \rightarrow M, C$
- $M \rightarrow B, C$
- $C \rightarrow B, M$
- $B, M \rightarrow C$
- $B, C \rightarrow M$
- $M, C \rightarrow B$

(b) Please compute the support and confidence of each association rule you listed in a.

The total number of transaction is 5. So:

- Support of rule $X \rightarrow Y$ = freq $(X \cup Y)$ / total_transaction.
- Confidence = support $(X \cup Y)$ / support X.

According to the data:

- $\{B, M, C\}$ is appeared at TID 1 and TID 5$\rightarrow$ freq = 2.
- Partitional Freq:
  - $\{B\} = 4$ (TID 1, 2, 4, 5)
  - $\{M\} = 4$ (TID 1, 3, 4, 5)
  - $\{C\} = 3$ (TID 1, 3, 5)
  - $\{B, M\} = 3$ (TID 1, 4, 5)
  - $\{B, C\} = 2$ (TID 1, 5)
  - $\{M, C\} = 3$ (TID 1, 3, 5)

- $B \rightarrow M, C$:
  - Support $= \frac{2}{5} = 0.4$
  - Confidence $= \frac{\frac{2}{5}}{\frac{4}{5}} = 0.5$
- $M \rightarrow B, C$:
  - Support $= 0.4$
  - Confidence $= \frac{\frac{2}{5}}{\frac{4}{5}} = 0.5$
- $C \rightarrow B, M$:
  - Support $= 0.4$
  - Confidence $= \frac{\frac{2}{5}}{\frac{3}{5}} = \frac{2}{3} \approx 0.667$
- $B, M \rightarrow C$:
  - Support $= 0.4$
  - Confidence $= \frac{\frac{2}{5}}{\frac{3}{5}} = \frac{2}{3} \approx 0.667$
- $B, C \rightarrow M$:
  - Support $= 0.4$
  - Confidence $= \frac{\frac{2}{5}}{\frac{2}{5}} = 1$
- $M, C \rightarrow B$:
  - Support $= 0.4$
  - Confidence $= \frac{\frac{2}{5}}{\frac{3}{5}} = \frac{2}{3} \approx 0.667$

2. 2. Please describe the anti-monotone property of the confidence for association rules, and discuss how to use this property to enable the pruning for the apriori algorithm. (15 pts)

The anti-monotone property of confidence for association rules applies specifically to rules generated from the same frequent itemset L (e.g., $L = A, B, C$). In this case, confidence is anti-monotone with respect to the number of items on the righthand side of the rule: as the size of the consequent (Y) increases (and the antecedent X decreases, since $X \cup Y = L$), confidence decreases or stays the same. This is because confidence = support(L) / support(X), where support(L) is fixed, and smaller X leads to higher support(X) due to the anti-monotone property of support.

So how to use that principle to pruning in **Apriori**:

Apriori first finds frequent itemsets based on the anti-monotone of support.

In the rule generation phase from the frequent itemset, use the rule generation tree. Start from the rule with small consequent (1 item), calculate the confidence. If the confidence is low, prune the branch that extends the larger consequent, because the confidence will decrease. This reduces the number of rules to check, optimizes.

3. Please choose one reference paper about association rule mining from the lecture slides, and present the motivation, problem definition, and their solutions in this paper. (40 pts)

I chose the paper: A Statistical Theory for Quantitative Association Rules by Yonatan Aumann and Yehuda Lindell, presented at the KDD-99 conference.

Motivation

Problem with Traditional Association Rules: Association rules, a popular data mining tool, were originally designed for categorical data, e.g., "X $\rightarrow$ Y" with high confidence and support. Algorithms like Apriori work well for this use case. However, most real-world databases contain quantitative attributes (e.g., age, salary, height), and extending the definition of categorical to quantitative is inefficient.

Limitations of previous approaches: The method of Srikant et al. 1996 [1] uses intervals to discretize quantitative data, leading to:

- Inaccurate or inaccurate results (e.g., the interval [100cm, 150cm] for

child height may include infants, but does not reflect the actual distribution).

- Exponential blowup, requiring a maxsup limit for filtering.
- Loss of information due to discretization, only finding an approximation of the best rule.
- Not capturing "interesting behavior" statistically, leading to redundant or misleading rules.

The authors argue that the goal of association rules is to detect extraordinary phenomena. They are inspired by categorical rules such as statistical correlations, and extend them quantitatively by using distributions instead of simple intervals. This makes it more efficient to mine real-world data, for example in the analysis of salaries, longevity, or writing habits.

Problem Definition

General structure of the rule: The rule has the form "population-subset $\rightarrow$ interesting-behavior", where:

- Left-hand side: Describes the population subset (profile), which can be a categorical attribute or a quantitative interval.
- Right-hand side: Describes the unusual behavior of that subset, using statistical measures of the distribution (such as mean, variance) of the quantitative attribute.

Interestingness: Behavior is interesting only if the distribution of the subset is significantly different from the rest of the population (complement). Use statistical tests for confirmation (e.g., Z-test for mean, F-test for variance), with confidence levels (usually 95%) and minimum differences (mindiff) to avoid the rule of triviality.

Specific types of rules:

- Categorical $\rightarrow$ Quantitative: Left is the categorical profile (e.g., sex = female), right is the mean/variance of one or more quantitative attributes (e.g., wage mean = \$7.90, overall = \$9.02).

- Quantitative → Quantitative: Left is the range on a quantitative attribute (e.g., age $\in [60, 80]$), right is the mean of another quantitative attribute. Rules must be "maximal" and "irreducible" to avoid redundant rules.

Sub-rules: Sub-rules of a basic rule, where the subset is smaller but the distribution is significantly different. This creates a hierarchy to provide comprehensive and accurate information (e.g. smoker → life expectancy = 60; smoker & wine-drinker → life expectancy = 70). Goal: Find all desired rules, including the basic rules and their sub-rules, without discretizing the data.

solution:

- **Algorithm for Quantitative → Quantitative (one attribute to one attribute):**
  - Sort the database by attribute on the left.
  - Use **Window** procedure: Maintain two windows (regions) A (irreducible, above/below average) and B (next to expand). Traverse linearly to find maximal and irreducible regions, check Z-test to confirm.
  - Call recursively to find sub-rules.
  - Complexity: $O(kn \log n + k^2 n)$ with k quantitative attributes, n transactions. Does not depend on minimum support, runs fast even with low support.
- **Algorithm for Categorical → Quantitative (many categoricals to many quantitative):**
  - Step 1: Find all frequent categorical sets using Apriori.
  - Step 2: Calculate mean/variance for each frequent set (use hash-tree, traverse only once).
  - Step 3: Browse the lattice of frequent sets to identify basic rules and sub-rules.
- Can be combined with Window for mixed profiles (categorical + a quantitative).

- **Statistical processing:** Use tests to filter out meaningless rules, avoid data explosion (e.g., remove 29,959 potential rules, keep only 354).
- **Experimental evaluation:**
  - On real data: Wages (1985 CPS, 534 transactions) and Linguistics (643 transactions, on non-native English writing habits).
  - Results: Find interesting rules (36% according to experts), many new rules are not found by traditional statistical tools like SPSS (e.g., Russians do not use "the" much without source text).
  - Compare to [7]: Their method generates more rules (about 6,000 vs 354), but only 1.2% interesting and often misleading due to the use of intervals.
  - Scalability: Runs fast (10–126 seconds for 10k–50k transactions), handles large data well.

4. ()

TID	Transactions
100	{A, B, E}
200	{B, D}
300	{A, B, E}
400	{A, C}
500	{B, C}
600	{A, C}
700	{A, B}
800	{A, B, C, E}
900	{A, B, C}
1000	{A, C, E}

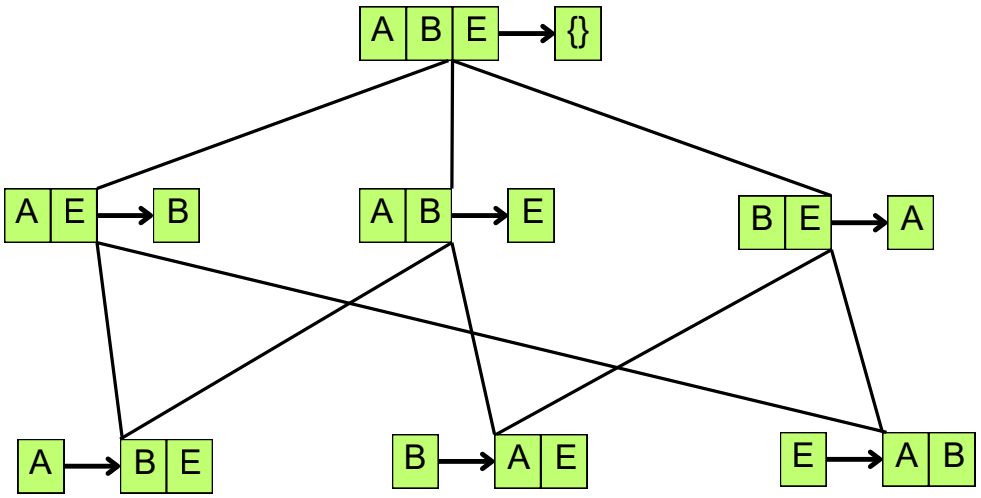
Table 2. Transaction Database for Bonus Question

(a) Please draw a lattice for association rules related to items A, B, E. [10 points]

(b) Compute the support and confidence of association rule AB → E in Table 2. [10 points]

a) Please draw a lattice for association rules related to items $\{A, B, E\}$

The lattice tree:



References

[1] R. Srikant, R. Agrawal, Mining quantitative association rules in large relational tables, ACM SIGMOD Record 25 (1) (1996) 1–10. doi:10.1145/235968.233311.