6

Association Analysis: Basic Concepts and Algorithms

1. For each of the following questions, provide an example of an association rule from the market basket domain that satisfies the following conditions. Also, describe whether such rules are subjectively interesting.

   (a) A rule that has high support and high confidence.
   Answer: Milk $\longrightarrow$ Bread. Such obvious rule tends to be uninteresting.

   (b) A rule that has reasonably high support but low confidence.
   Answer: Milk $\longrightarrow$ Tuna. While the sale of tuna and milk may be higher than the support threshold, not all transactions that contain milk also contain tuna. Such low-confidence rule tends to be uninteresting.

   (c) A rule that has low support and low confidence.
   Answer: Cooking oil $\longrightarrow$ Laundry detergent. Such low confidence rule tends to be uninteresting.

   (d) A rule that has low support and high confidence.
   Answer: Vodka $\longrightarrow$ Caviar. Such rule tends to be interesting.

2. Consider the data set shown in Table 6.1.

   (a) Compute the support for itemsets $\{e\}$, $\{b, d\}$, and $\{b, d, e\}$ by treating each transaction ID as a market basket.
   Answer:

Table 6.1. Example of market basket transactions.

Customer ID	Transaction ID	Items Bought
1	0001	$\{a, d, e\}$
1	0024	$\{a, b, c, e\}$
2	0012	$\{a, b, d, e\}$
2	0031	$\{a, c, d, e\}$
3	0015	$\{b, c, e\}$
3	0022	$\{b, d, e\}$
4	0029	$\{c, d\}$
4	0040	$\{a, b, c\}$
5	0033	$\{a, d, e\}$
5	0038	$\{a, b, e\}$

$$\begin{aligned} s(\{e\}) &= \frac{8}{10} = 0.8 \\ s(\{b,d\}) &= \frac{2}{10} = 0.2 \\ s(\{b,d,e\}) &= \frac{2}{10} = 0.2 \end{aligned} \tag{6.1}$$

(b) Use the results in part (a) to compute the confidence for the association rules $\{b, d\} \longrightarrow \{e\}$ and $\{e\} \longrightarrow \{b, d\}$. Is confidence a symmetric measure?

Answer:

$$\begin{aligned} c(bd \longrightarrow e) &= \frac{0.2}{0.2} = 100\% \\ c(e \longrightarrow bd) &= \frac{0.2}{0.8} = 25\% \end{aligned}$$

No, confidence is not a symmetric measure.

(c) Repeat part (a) by treating each customer ID as a market basket. Each item should be treated as a binary variable (1 if an item appears in at least one transaction bought by the customer, and 0 otherwise.)

Answer:

$$\begin{aligned} s(\{e\}) &= \frac{4}{5} = 0.8 \\ s(\{b,d\}) &= \frac{5}{5} = 1 \\ s(\{b,d,e\}) &= \frac{4}{5} = 0.8 \end{aligned}$$

(d) Use the results in part (c) to compute the confidence for the association rules $\{b, d\} \longrightarrow \{e\}$ and $\{e\} \longrightarrow \{b, d\}$.

Answer:

$$c(bd \longrightarrow e) = \frac{0.8}{1} = 80\%$$
$$c(e \longrightarrow bd) = \frac{0.8}{0.8} = 100\%$$

(e) Suppose s_1 and c_1 are the support and confidence values of an association rule r when treating each transaction ID as a market basket. Also, let s_2 and c_2 be the support and confidence values of r when treating each customer ID as a market basket. Discuss whether there are any relationships between s_1 and s_2 or c_1 and c_2.

Answer:

There are no apparent relationships between s_1, s_2, c_1, and c_2.

3. (a) What is the confidence for the rules $\emptyset \longrightarrow A$ and $A \longrightarrow \emptyset$?

Answer:

$c(\emptyset \longrightarrow A) = s(\emptyset \longrightarrow A)$.

$c(A \longrightarrow \emptyset) = 100\%$.

(b) Let c_1, c_2, and c_3 be the confidence values of the rules $\{p\} \longrightarrow \{q\}$, $\{p\} \longrightarrow \{q, r\}$, and $\{p, r\} \longrightarrow \{q\}$, respectively. If we assume that c_1, c_2, and c_3 have different values, what are the possible relationships that may exist among c_1, c_2, and c_3? Which rule has the lowest confidence?

Answer:

$c_1 = \frac{s(p \cup q)}{s(p)}$

$c_2 = \frac{s(p \cup q \cup r)}{s(p)}$

$c_3 = \frac{s(p \cup q \cup r)}{s(p \cup r)}$

Considering $s(p) \geq s(p \cup q) \geq s(p \cup q \cup r)$

Thus: $c1 \geq c2$ & $c3 \geq c2$.

Therefore $c2$ has the lowest confidence.

(c) Repeat the analysis in part (b) assuming that the rules have identical support. Which rule has the highest confidence?

Answer:

Considering $s(p \cup q) = s(p \cup q \cup r)$

but $s(p) \geq s(p \cup r)$

Thus: $c3 \geq (c1 = c2)$

Either all rules have the same confidence or c3 has the highest confidence.

(d) Transitivity: Suppose the confidence of the rules $A \longrightarrow B$ and $B \longrightarrow C$ are larger than some threshold, $minconf$. Is it possible that $A \longrightarrow C$ has a confidence less than $minconf$?

Answer:

Yes, It depends on the support of items A, B, and C.

For example:

s(A,B) = 60% s(A) = 90%
s(A,C) = 20% s(B) = 70%
s(B,C) = 50% s(C) = 60%

Let $minconf = 50\%$ Therefore:

$c(A \rightarrow B) = 66\% > minconf$
$c(B \rightarrow C) = 71\% > minconf$
But $c(A \rightarrow C) = 22\% < minconf$

4. For each of the following measures, determine whether it is monotone, anti-monotone, or non-monotone (i.e., neither monotone nor anti-monotone).

Example: Support, $s = \frac{\sigma(X)}{|T|}$ is anti-monotone because $s(X) \geq s(Y)$ whenever $X \subset Y$.

(a) A characteristic rule is a rule of the form $\{p\} \longrightarrow \{q_1, q_2, \ldots, q_n\}$, where the rule antecedent contains only a single item. An itemset of size k can produce up to k characteristic rules. Let ζ be the minimum confidence of all characteristic rules generated from a given itemset:

$$\begin{aligned} \zeta(\{p_1, p_2, \ldots, p_k\}) &= \min\big[\, c\big(\{p_1\} \longrightarrow \{p_2, p_3, \ldots, p_k\}\big), \ldots \\ &\qquad c\big(\{p_k\} \longrightarrow \{p_1, p_3 \ldots, p_{k-1}\}\big)\,\big] \end{aligned}$$

Is ζ monotone, anti-monotone, or non-monotone?

Answer:

ζ is an anti-monotone measure because

$$\zeta(\{A_1, A_2, \cdots, A_k\}) \geq \zeta(\{A_1, A_2, \cdots, A_k, A_{k+1}\}) \tag{6.2}$$

For example, we can compare the values of ζ for $\{A, B\}$ and $\{A, B, C\}$.

$$\begin{aligned} \zeta(\{A, B\}) &= \min\big(c(A \longrightarrow B), c(B \longrightarrow A)\big) \\ &= \min\big(\frac{s(A,B)}{s(A)}, \frac{s(A,B)}{s(B)}\big) \\ &= \frac{s(A,B)}{\max(s(A), s(B))} \end{aligned} \tag{6.3}$$

$$\begin{aligned}\zeta(\{A,B,C\}) &= \min\big(c(A \longrightarrow BC), c(B \longrightarrow AC), c(C \longrightarrow AB)\big) \\ &= \min\big(\frac{s(A,B,C)}{s(A)}, \frac{s(A,B,C)}{s(B)}, \frac{s(A,B,C)}{s(C)}\big) \\ &= \frac{s(A,B,C)}{\max(s(A), s(B), s(C))} \qquad (6.4)\end{aligned}$$

Since $s(A,B,C) \leq s(A,B)$ and $\max(s(A), s(B), s(C)) \geq \max(s(A), s(B))$, therefore $\zeta(\{A,B\}) \geq \zeta(\{A,B,C\})$.

(b) A discriminant rule is a rule of the form $\{p_1, p_2, \ldots, p_n\} \longrightarrow \{q\}$, where the rule consequent contains only a single item. An itemset of size k can produce up to k discriminant rules. Let η be the minimum confidence of all discriminant rules generated from a given itemset:

$$\begin{aligned}\eta(\{p_1, p_2, \ldots, p_k\}) &= \min\big[\, c\big(\{p_2, p_3, \ldots, p_k\} \longrightarrow \{p_1\}\big), \ldots \\ &\qquad c\big(\{p_1, p_2, \ldots p_{k-1}\} \longrightarrow \{p_k\}\big)\,\big]\end{aligned}$$

Is η monotone, anti-monotone, or non-monotone?

Answer:

η is non-monotone. We can show this by comparing $\eta(\{A,B\})$ against $\eta(\{A,B,C\})$.

$$\begin{aligned}\eta(\{A,B\}) &= \min\big(c(A \longrightarrow B), c(B \longrightarrow A)\big) \\ &= \min\big(\frac{s(A,B)}{s(A)}, \frac{s(A,B)}{s(B)}\big) \\ &= \frac{s(A,B)}{\max(s(A), s(B))} \qquad (6.5)\end{aligned}$$

$$\begin{aligned}\eta(\{A,B,C\}) &= \min\big(c(AB \longrightarrow C), c(AC \longrightarrow B), c(BC \longrightarrow A)\big) \\ &= \min\big(\frac{s(A,B,C)}{s(A,B)}, \frac{s(A,B,C)}{s(A,C)}, \frac{s(A,B,C)}{s(B,C)}\big) \\ &= \frac{s(A,B,C)}{\max(s(A,B), s(A,C), s(B,C))} \qquad (6.6)\end{aligned}$$

Since $s(A,B,C) \leq s(A,B)$ and $\max(s(A,B), s(A,C), s(B,C)) \leq \max(s(A), s(B))$, therefore $\eta(\{A,B,C\})$ can be greater than or less than $\eta(\{A,B\})$. Hence, the measure is non-monotone.

(c) Repeat the analysis in parts (a) and (b) by replacing the min function with a max function.

Answer:

Let

$$\zeta'(\{A_1, A_2, \cdots, A_k\}) = \max(\quad c(A_1 \longrightarrow A_2, A_3, \cdots, A_k), \cdots \\ c(A_k \longrightarrow A_1, A_3 \cdots, A_{k-1}))$$

$$\begin{aligned} \zeta'(\{A, B\}) &= \max\big(c(A \longrightarrow B), c(B \longrightarrow A)\big) \\ &= \max\big(\frac{s(A,B)}{s(A)}, \frac{s(A,B)}{s(B)}\big) \\ &= \frac{s(A,B)}{\min(s(A), s(B))} \end{aligned} \tag{6.7}$$

$$\begin{aligned} \zeta'(\{A, B, C\}) &= \max\big(c(A \longrightarrow BC), c(B \longrightarrow AC), c(C \longrightarrow AB)\big) \\ &= \max\big(\frac{s(A,B,C)}{s(A)}, \frac{s(A,B,C)}{s(B)}, \frac{s(A,B,C)}{s(C)}\big) \\ &= \frac{s(A,B,C)}{\min(s(A), s(B), s(C))} \end{aligned} \tag{6.8}$$

Since $s(A, B, C) \leq s(A, B)$ and $\min(s(A), s(B), s(C)) \leq \min(s(A), s(B))$, $\zeta'(\{A, B, C\})$ can be greater than or less than $\zeta'(\{A, B\})$. Therefore, the measure is non-monotone.

Let

$$\eta'(\{A_1, A_2, \cdots, A_k\}) = \max(\quad c(A_2, A_3, \cdots, A_k \longrightarrow A_1), \cdots \\ c(A_1, A_2, \cdots A_{k-1} \longrightarrow A_k))$$

$$\begin{aligned} \eta'(\{A, B\}) &= \max\big(c(A \longrightarrow B), c(B \longrightarrow A)\big) \\ &= \max\big(\frac{s(A,B)}{s(A)}, \frac{s(A,B)}{s(B)}\big) \\ &= \frac{s(A,B)}{\min(s(A), s(B))} \end{aligned} \tag{6.9}$$

$$\begin{aligned} \eta(\{A, B, C\}) &= \max\big(c(AB \longrightarrow C), c(AC \longrightarrow B), c(BC \longrightarrow A)\big) \\ &= \max\big(\frac{s(A,B,C)}{s(A,B)}, \frac{s(A,B,C)}{s(A,C)}, \frac{s(A,B,C)}{s(B,C)}\big) \\ &= \frac{s(A,B,C)}{\min(s(A,B), s(A,C), s(B,C))} \end{aligned} \tag{6.10}$$

Since $s(A,B,C) \leq s(A,B)$ and $\min(s(A,B), s(A,C), s(B,C)) \leq \min(s(A), s(B), s(C))$ $\leq \min(s(A), s(B))$, $\eta'(\{A,B,C\})$ can be greater than or less than $\eta'(\{A,B\})$. Hence, the measure is non-monotone.

5. Prove Equation 6.3. (Hint: First, count the number of ways to create an itemset that forms the left hand side of the rule. Next, for each size k itemset selected for the left-hand side, count the number of ways to choose the remaining $d-k$ items to form the right-hand side of the rule.)

Answer:

Suppose there are d items. We first choose k of the items to form the left-hand side of the rule. There are $\binom{d}{k}$ ways for doing this. After selecting the items for the left-hand side, there are $\binom{d-k}{i}$ ways to choose the remaining items to form the right hand side of the rule, where $1 \leq i \leq d-k$. Therefore the total number of rules (R) is:

$$\begin{aligned} R &= \sum_{k=1}^{d} \binom{d}{k} \sum_{i=1}^{d-k} \binom{d-k}{i} \\ &= \sum_{k=1}^{d} \binom{d}{k} \left(2^{d-k} - 1\right) \\ &= \sum_{k=1}^{d} \binom{d}{k} 2^{d-k} - \sum_{k=1}^{d} \binom{d}{k} \\ &= \sum_{k=1}^{d} \binom{d}{k} 2^{d-k} - \left[2^d + 1\right], \end{aligned}$$

where

$$\sum_{i=1}^{n} \binom{n}{i} = 2^n - 1.$$

Since

$$(1+x)^d = \sum_{i=1}^{d} \binom{d}{i} x^{d-i} + x^d,$$

substituting $x = 2$ leads to:

$$3^d = \sum_{i=1}^{d} \binom{d}{i} 2^{d-i} + 2^d.$$

Therefore, the total number of rules is:

$$R = 3^d - 2^d - \left[2^d + 1\right] = 3^d - 2^{d+1} + 1.$$

Table 6.2. Market basket transactions.

Transaction ID	Items Bought
1	{Milk, Beer, Diapers}
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{Milk, Diapers, Bread, Butter}
7	{Bread, Butter, Diapers}
8	{Beer, Diapers}
9	{Milk, Diapers, Bread, Butter}
10	{Beer, Cookies}

6. Consider the market basket transactions shown in Table 6.2.

   (a) What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)?

   Answer: There are six items in the data set. Therefore the total number of rules is 602.

   (b) What is the maximum size of frequent itemsets that can be extracted (assuming $minsup > 0$)?

   Answer: Because the longest transaction contains 4 items, the maximum size of frequent itemset is 4.

   (c) Write an expression for the maximum number of size-3 itemsets that can be derived from this data set.

   Answer: $\binom{6}{3} = 20$.

   (d) Find an itemset (of size 2 or larger) that has the largest support.

   Answer: {Bread, Butter}.

   (e) Find a pair of items, a and b, such that the rules $\{a\} \longrightarrow \{b\}$ and $\{b\} \longrightarrow \{a\}$ have the same confidence.

   Answer: (Beer, Cookies) or (Bread, Butter).

7. Consider the following set of frequent 3-itemsets:

$$\{1,2,3\}, \{1,2,4\}, \{1,2,5\}, \{1,3,4\}, \{1,3,5\}, \{2,3,4\}, \{2,3,5\}, \{3,4,5\}.$$

   Assume that there are only five items in the data set.

   (a) List all candidate 4-itemsets obtained by a candidate generation procedure using the $F_{k-1} \times F_1$ merging strategy.

   Answer:

   $\{1,2,3,4\}$,$\{1,2,3,5\}$,$\{1,2,3,6\}$.
   $\{1,2,4,5\}$,$\{1,2,4,6\}$,$\{1,2,5,6\}$.

$\{1,3,4,5\}$,$\{1,3,4,6\}$,$\{2,3,4,5\}$.
$\{2,3,4,6\}$,$\{2,3,5,6\}$.

(b) List all candidate 4-itemsets obtained by the candidate generation procedure in *Apriori*.

Answer:

$\{1,2,3,4\}$, $\{1,2,3,5\}$, $\{1,2,4,5\}$, $\{2,3,4,5\}$, $\{2,3,4,6\}$.

(c) List all candidate 4-itemsets that survive the candidate pruning step of the *Apriori* algorithm.

Answer:

$\{1,2,3,4\}$

8. The *Apriori* algorithm uses a generate-and-count strategy for deriving frequent itemsets. Candidate itemsets of size $k+1$ are created by joining a pair of frequent itemsets of size k (this is known as the candidate generation step). A candidate is discarded if any one of its subsets is found to be infrequent during the candidate pruning step. Suppose the *Apriori* algorithm is applied to the data set shown in Table 6.3 with $minsup = 30\%$, i.e., any itemset occurring in less than 3 transactions is considered to be infrequent.

Table 6.3. Example of market basket transactions.

Transaction ID	Items Bought
1	$\{a,b,d,e\}$
2	$\{b,c,d\}$
3	$\{a,b,d,e\}$
4	$\{a,c,d,e\}$
5	$\{b,c,d,e\}$
6	$\{b,d,e\}$
7	$\{c,d\}$
8	$\{a,b,c\}$
9	$\{a,d,e\}$
10	$\{b,d\}$

(a) Draw an itemset lattice representing the data set given in Table 6.3. Label each node in the lattice with the following letter(s):

- **N**: If the itemset is not considered to be a candidate itemset by the *Apriori* algorithm. There are two reasons for an itemset not to be considered as a candidate itemset: (1) it is not generated at all during the candidate generation step, or (2) it is generated during

the candidate generation step but is subsequently removed during the candidate pruning step because one of its subsets is found to be infrequent.

- **F**: If the candidate itemset is found to be frequent by the *Apriori* algorithm.
- **I**: If the candidate itemset is found to be infrequent after support counting.

Answer:

The lattice structure is shown below.

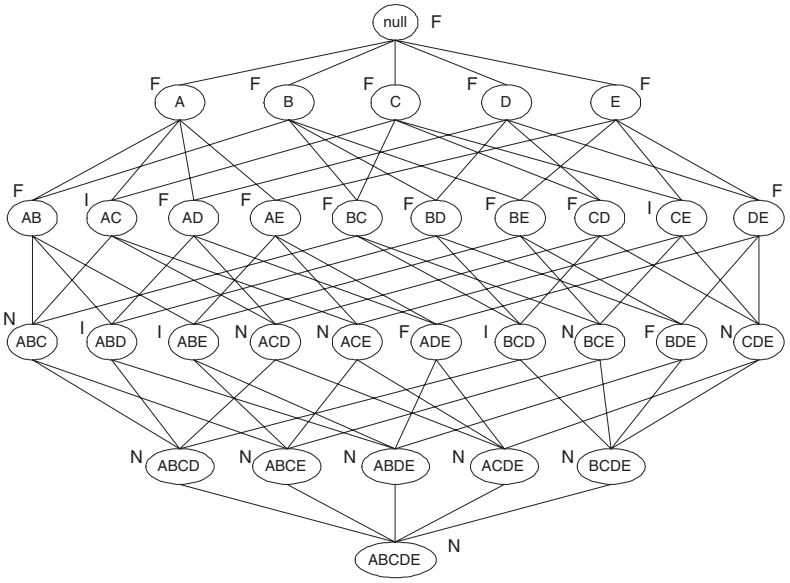

Figure 6.1. Solution.

(b) What is the percentage of frequent itemsets (with respect to all itemsets in the lattice)?

Answer:

Percentage of frequent itemsets = 16/32 = 50.0% (including the null set).

(c) What is the pruning ratio of the *Apriori* algorithm on this data set? (Pruning ratio is defined as the percentage of itemsets not considered to be a candidate because (1) they are not generated during candidate generation or (2) they are pruned during the candidate pruning step.)

Answer:

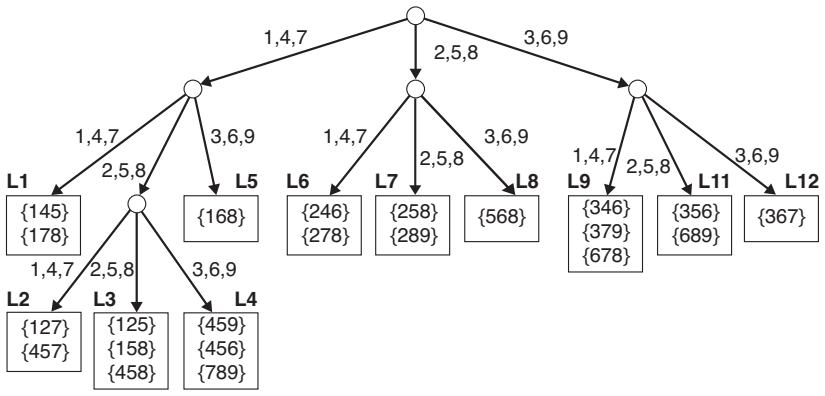

Figure 6.2. An example of a hash tree structure.

Pruning ratio is the ratio of N to the total number of itemsets. Since the count of $N = 11$, therefore pruning ratio is $11/32 = 34.4\%$.

(d) What is the false alarm rate (i.e, percentage of candidate itemsets that are found to be infrequent after performing support counting)?

Answer:

False alarm rate is the ratio of I to the total number of itemsets. Since the count of $I = 5$, therefore the false alarm rate is $5/32 = 15.6\%$.

9. The *Apriori* algorithm uses a hash tree data structure to efficiently count the support of candidate itemsets. Consider the hash tree for candidate 3-itemsets shown in Figure 6.2.

(a) Given a transaction that contains items $\{1, 3, 4, 5, 8\}$, which of the hash tree leaf nodes will be visited when finding the candidates of the transaction?

Answer:

The leaf nodes visited are L1, L3, L5, L9, and L11.

(b) Use the visited leaf nodes in part (b) to determine the candidate itemsets that are contained in the transaction $\{1, 3, 4, 5, 8\}$.

Answer:

The candidates contained in the transaction are $\{1, 4, 5\}$, $\{1, 5, 8\}$, and $\{4, 5, 8\}$.

10. Consider the following set of candidate 3-itemsets:

$$\{1, 2, 3\}, \{1, 2, 6\}, \{1, 3, 4\}, \{2, 3, 4\}, \{2, 4, 5\}, \{3, 4, 6\}, \{4, 5, 6\}$$

(a) Construct a hash tree for the above candidate 3-itemsets. Assume the tree uses a hash function where all odd-numbered items are hashed to the left child of a node, while the even-numbered items are hashed to the right child. A candidate k-itemset is inserted into the tree by hashing on each successive item in the candidate and then following the appropriate branch of the tree according to the hash value. Once a leaf node is reached, the candidate is inserted based on one of the following conditions:

Condition 1: If the depth of the leaf node is equal to k (the root is assumed to be at depth 0), then the candidate is inserted regardless of the number of itemsets already stored at the node.

Condition 2: If the depth of the leaf node is less than k, then the candidate can be inserted as long as the number of itemsets stored at the node is less than $maxsize$. Assume $maxsize = 2$ for this question.

Condition 3: If the depth of the leaf node is less than k and the number of itemsets stored at the node is equal to $maxsize$, then the leaf node is converted into an internal node. New leaf nodes are created as children of the old leaf node. Candidate itemsets previously stored in the old leaf node are distributed to the children based on their hash values. The new candidate is also hashed to its appropriate leaf node.

Answer:

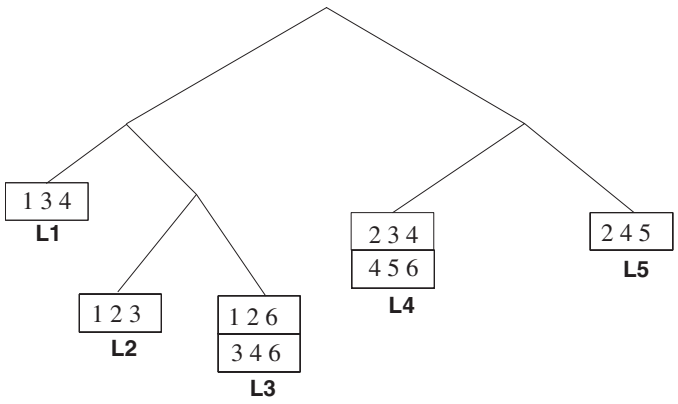

Figure 6.3. Hash tree for Exercise 10.

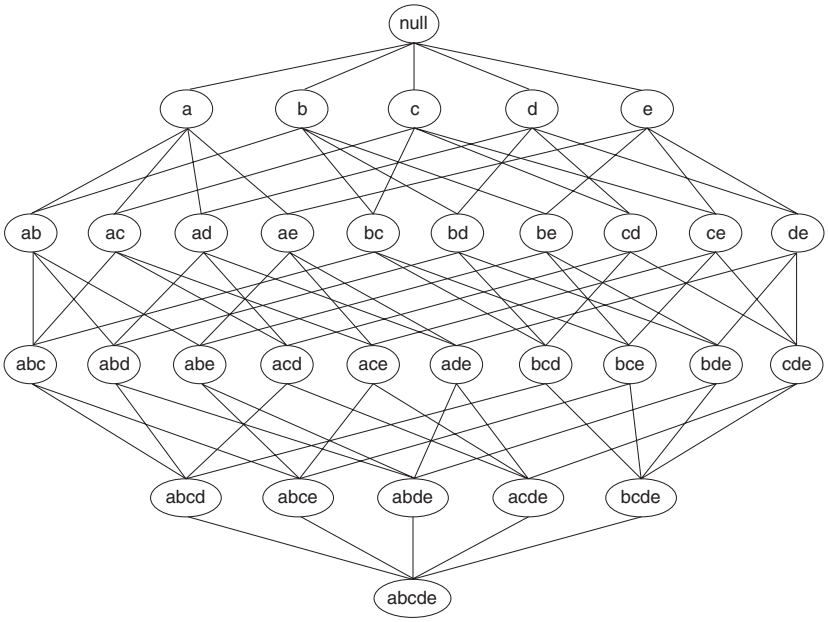

Figure 6.4. An itemset lattice

(b) How many leaf nodes are there in the candidate hash tree? How many internal nodes are there?

Answer: There are 5 leaf nodes and 4 internal nodes.

(c) Consider a transaction that contains the following items: $\{1, 2, 3, 5, 6\}$. Using the hash tree constructed in part (a), which leaf nodes will be checked against the transaction? What are the candidate 3-itemsets contained in the transaction?

Answer: The leaf nodes L1, L2, L3, and L4 will be checked against the transaction. The candidate itemsets contained in the transaction include {1,2,3} and {1,2,6}.

11. Given the lattice structure shown in Figure 6.4 and the transactions given in Table 6.3, label each node with the following letter(s):

- M if the node is a maximal frequent itemset,
- C if it is a closed frequent itemset,
- N if it is frequent but neither maximal nor closed, and
- I if it is infrequent.

Assume that the support threshold is equal to 30%.

Answer:

The lattice structure is shown below.

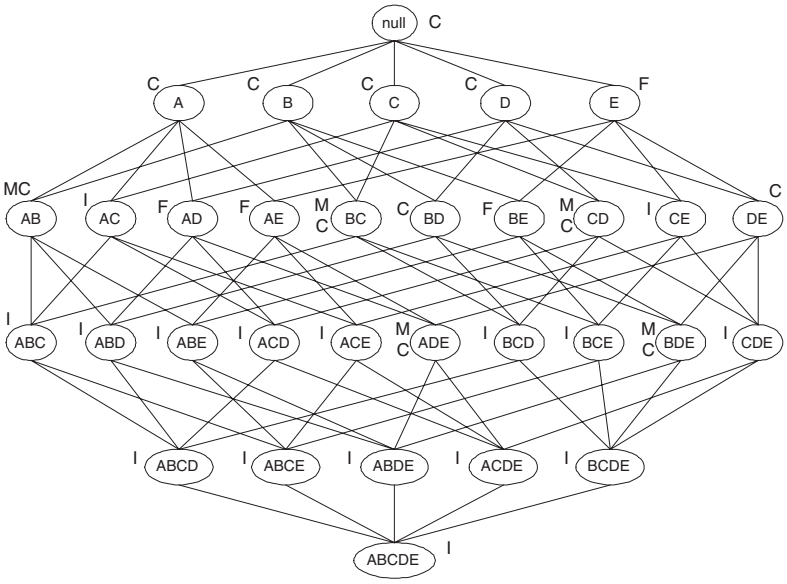

Figure 6.5. Solution for Exercise 11.

12. The original association rule mining formulation uses the support and confidence measures to prune uninteresting rules.

    (a) Draw a contingency table for each of the following rules using the transactions shown in Table 6.4.

    Rules: $\{b\} \longrightarrow \{c\}$, $\{a\} \longrightarrow \{d\}$, $\{b\} \longrightarrow \{d\}$, $\{e\} \longrightarrow \{c\}$, $\{c\} \longrightarrow \{a\}$.

    Answer:

	c	$\overline{c}$
b	3	4
$\overline{b}$	2	1

	d	$\overline{d}$
a	4	1
$\overline{a}$	5	0

	d	$\overline{d}$
b	6	1
$\overline{b}$	3	0

	c	$\overline{c}$
e	2	4
$\overline{e}$	3	1

	a	$\overline{a}$
c	2	3
$\overline{c}$	3	2

    (b) Use the contingency tables in part (a) to compute and rank the rules in decreasing order according to the following measures.

Table 6.4. Example of market basket transactions.

Transaction ID	Items Bought
1	$\{a, b, d, e\}$
2	$\{b, c, d\}$
3	$\{a, b, d, e\}$
4	$\{a, c, d, e\}$
5	$\{b, c, d, e\}$
6	$\{b, d, e\}$
7	$\{c, d\}$
8	$\{a, b, c\}$
9	$\{a, d, e\}$
10	$\{b, d\}$

i. Support.
Answer:

Rules	Support	Rank
$b \longrightarrow c$	0.3	3
$a \longrightarrow d$	0.4	2
$b \longrightarrow d$	0.6	1
$e \longrightarrow c$	0.2	4
$c \longrightarrow a$	0.2	4

ii. Confidence.
Answer:

Rules	Confidence	Rank
$b \longrightarrow c$	3/7	3
$a \longrightarrow d$	4/5	2
$b \longrightarrow d$	6/7	1
$e \longrightarrow c$	2/6	5
$c \longrightarrow a$	2/5	4

iii. Interest$(X \longrightarrow Y) = \frac{P(X,Y)}{P(X)}P(Y)$.
Answer:

Rules	Interest	Rank
$b \longrightarrow c$	0.214	3
$a \longrightarrow d$	0.72	2
$b \longrightarrow d$	0.771	1
$e \longrightarrow c$	0.167	5
$c \longrightarrow a$	0.2	4

iv. IS$(X \longrightarrow Y) = \frac{P(X,Y)}{\sqrt{P(X)P(Y)}}$.
Answer:

Rules	IS	Rank
$b \longrightarrow c$	0.507	3
$a \longrightarrow d$	0.596	2
$b \longrightarrow d$	0.756	1
$e \longrightarrow c$	0.365	5
$c \longrightarrow a$	0.4	4

v. Klosgen$(X \longrightarrow Y) = \sqrt{P(X,Y)} \times (P(Y|X) - P(Y))$, where $P(Y|X) = \frac{P(X,Y)}{P(X)}$.

Answer:

Rules	Klosgen	Rank
$b \longrightarrow c$	-0.039	2
$a \longrightarrow d$	-0.063	4
$b \longrightarrow d$	-0.033	1
$e \longrightarrow c$	-0.075	5
$c \longrightarrow a$	-0.045	3

vi. Odds ratio$(X \longrightarrow Y) = \frac{P(X,Y)P(\overline{X},\overline{Y})}{P(X,\overline{Y})P(\overline{X},Y)}$.

Answer:

Rules	Odds Ratio	Rank
$b \longrightarrow c$	0.375	2
$a \longrightarrow d$	0	4
$b \longrightarrow d$	0	4
$e \longrightarrow c$	0.167	3
$c \longrightarrow a$	0.444	1

13. Given the rankings you had obtained in Exercise 12, compute the correlation between the rankings of confidence and the other five measures. Which measure is most highly correlated with confidence? Which measure is least correlated with confidence?

    Answer:

    Correlation(Confidence, Support) = 0.97.

    Correlation(Confidence, Interest) = 1.

    Correlation(Confidence, IS) = 1.

    Correlation(Confidence, Klosgen) = 0.7.

    Correlation(Confidence, Odds Ratio) = -0.606.

    Interest and IS are the most highly correlated with confidence, while odds ratio is the least correlated.

14. Answer the following questions using the data sets shown in Figure 6.6. Note that each data set contains 1000 items and 10,000 transactions. Dark cells indicate the presence of items and white cells indicate the absence of

items. We will apply the *Apriori* algorithm to extract frequent itemsets with $minsup = 10\%$ (i.e., itemsets must be contained in at least 1000 transactions)?

(a) Which data set(s) will produce the most number of frequent itemsets?
Answer: Data set (e) because it has to generate the longest frequent itemset along with its subsets.

(b) Which data set(s) will produce the fewest number of frequent itemsets?
Answer: Data set (d) which does not produce any frequent itemsets at 10% support threshold.

(c) Which data set(s) will produce the longest frequent itemset?
Answer: Data set (e).

(d) Which data set(s) will produce frequent itemsets with highest maximum support?
Answer: Data set (b).

(e) Which data set(s) will produce frequent itemsets containing items with wide-varying support levels (i.e., items with mixed support, ranging from less than 20% to more than 70%).
Answer: Data set (e).

15. (a) Prove that the ϕ coefficient is equal to 1 if and only if $f_{11} = f_{1+} = f_{+1}$.
Answer:
Instead of proving $f_{11} = f_{1+} = f_{+1}$, we will show that $P(A, B) = P(A) = P(B)$, where $P(A, B) = f_{11}/N$, $P(A) = f_{1+}/N$, and $P(B) = f_{+1}/N$. When the ϕ-coefficient equals to 1:

$$\phi = \frac{P(A,B) - P(A)P(B)}{\sqrt{P(A)P(B)\big[1 - P(A)\big]\big[1 - P(B)\big]}} = 1$$

The preceding equation can be simplified as follows:

$$\begin{aligned}
\Big[P(A,B) - P(A)P(B)\Big]^2 &= P(A)P(B)\big[1 - P(A)\big]\big[1 - P(B)\big] \\
P(A,B)^2 - 2P(A,B)P(A)P(B) &= P(A)P(B)\big[1 - P(A) - P(B)\big] \\
P(A,B)^2 &= P(A)P(B)\big[1 - P(A) - P(B) + 2P(A,B)\big]
\end{aligned}$$

We may rewrite the equation in terms of $P(B)$ as follows:

$$P(A)P(B)^2 - P(A)\big[1 - P(A) + 2P(A,B)\big]P(B) + P(A,B)^2 = 0$$

The solution to the quadratic equation in $P(B)$ is:

$$P(B) = \frac{P(A)\beta - \sqrt{P(A)^2\beta^2 - 4P(A)P(A,B)^2}}{2P(A)},$$

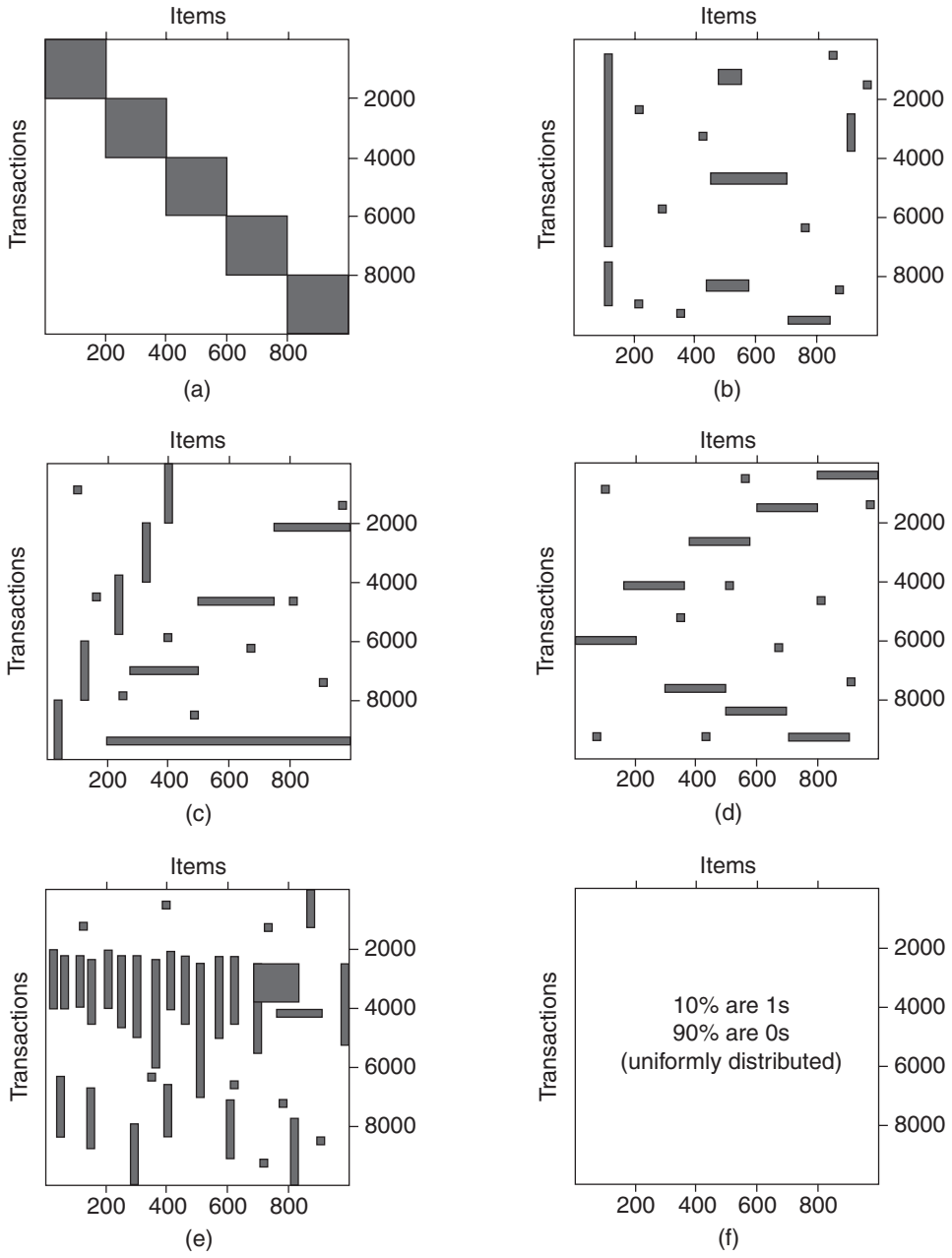

Figure 6.6. Figures for Exercise 14.

where $\beta = 1 - P(A) + 2P(A,B)$. Note that the second solution, in which the second term on the left hand side is positive, is not a feasible solution because it corresponds to $\phi = -1$. Furthermore, the solution for $P(B)$ must satisfy the following constraint: $P(B) \geq P(A,B)$. It can be shown that:

$$\begin{aligned} & P(B) - P(A,B) \\ = & \frac{1-P(A)}{2} - \frac{\sqrt{(1-P(A))^2 + 4P(A,B)(1-P(A))(1-P(A,B)/P(A))}}{2} \\ \leq & 0 \end{aligned}$$

Because of the constraint, $P(B) = P(A,B)$, which can be achieved by setting $P(A,B) = P(A)$.

(b) Show that if A and B are independent, then $P(A,B) \times P(A,\overline{B}) = P(A,\overline{B}) \times P(\overline{A},B)$.

Answer:

When A and B are independent, $P(A,B) = P(A) \times P(B)$ or equivalently:

$$\begin{aligned} & P(A,B) - P(A)P(B) = 0 \\ & P(A,B) - [P(A,B) + P(A,\overline{B})][P(A,B) + P(\overline{A},B)] = 0 \\ & P(A,B)[1 - P(A,B) - P(A,\overline{B}) - P(\overline{A},B)] - P(\overline{A},B)P(A,\overline{B}) = 0 \\ & P(A,B)P(\overline{A},\overline{B}) - P(\overline{A},B)P(A,\overline{B}) = 0. \end{aligned}$$

(c) Show that Yule's Q and Y coefficients

$$\begin{aligned} Q &= \left[\frac{f_{11}f_{00} - f_{10}f_{01}}{f_{11}f_{00} + f_{10}f_{01}}\right] \\ Y &= \left[\frac{\sqrt{f_{11}f_{00}} - \sqrt{f_{10}f_{01}}}{\sqrt{f_{11}f_{00}} + \sqrt{f_{10}f_{01}}}\right] \end{aligned}$$

are normalized versions of the odds ratio.

Answer:

Odds ratio can be written as:

$$\alpha = \frac{f_{11}f_{00}}{f_{10}f_{01}}.$$

We can express Q and Y in terms of α as follows:

$$\begin{aligned} Q &= \frac{\alpha - 1}{\alpha + 1} \\ Y &= \frac{\sqrt{\alpha} - 1}{\sqrt{\alpha} + 1} \end{aligned}$$

In both cases, Q and Y increase monotonically with α. Furthermore, when $\alpha = 0$, $Q = Y = -1$ to represent perfect negative correlation. When $\alpha = 1$, which is the condition for attribute independence, $Q = Y = 1$. Finally, when $\alpha = \infty$, $Q = Y = +1$. This suggests that Q and Y are normalized versions of α.

(d) Write a simplified expression for the value of each measure shown in Tables 6.11 and 6.12 when the variables are statistically independent.

Answer:

Measure	Value under independence
ϕ-coefficient	0
Odds ratio	1
Kappa κ	0
Interest	1
Cosine, IS	$\sqrt{P(A,B)}$
Piatetsky-Shapiro's	0
Collective strength	1
Jaccard	$0 \cdots 1$
Conviction	1
Certainty factor	0
Added value	0

16. Consider the interestingness measure, $M = \frac{P(B|A)-P(B)}{1-P(B)}$, for an association rule $A \longrightarrow B$.

(a) What is the range of this measure? When does the measure attain its maximum and minimum values?

Answer:

The range of the measure is from 0 to 1. The measure attains its maximum value when $P(B|A) = 1$ and its minimum value when $P(B|A) = P(B)$.

(b) How does M behave when $P(A, B)$ is increased while $P(A)$ and $P(B)$ remain unchanged?

Answer:

The measure can be rewritten as follows:

$$\frac{P(A,B) - P(A)P(B)}{P(A)(1 - P(B))}.$$

It increases when $P(A, B)$ is increased.

(c) How does M behave when $P(A)$ is increased while $P(A, B)$ and $P(B)$ remain unchanged?

Answer:

The measure decreases with increasing $P(A)$.

(d) How does M behave when $P(B)$ is increased while $P(A, B)$ and $P(A)$ remain unchanged?

Answer:

The measure decreases with increasing $P(B)$.

(e) Is the measure symmetric under variable permutation?

Answer: No.

(f) What is the value of the measure when A and B are statistically independent?

Answer: 0.

(g) Is the measure null-invariant?

Answer: No.

(h) Does the measure remain invariant under row or column scaling operations?

Answer: No.

(i) How does the measure behave under the inversion operation?

Answer: Asymmetric.

17. Suppose we have market basket data consisting of 100 transactions and 20 items. If the support for item a is 25%, the support for item b is 90% and the support for itemset $\{a, b\}$ is 20%. Let the support and confidence thresholds be 10% and 60%, respectively.

(a) Compute the confidence of the association rule $\{a\} \rightarrow \{b\}$. Is the rule interesting according to the confidence measure?

Answer:

Confidence is $0.2/0.25 = 80\%$. The rule is interesting because it exceeds the confidence threshold.

(b) Compute the interest measure for the association pattern $\{a, b\}$. Describe the nature of the relationship between item a and item b in terms of the interest measure.

Answer:

The interest measure is $0.2/(0.25 \times 0.9) = 0.889$. The items are negatively correlated according to interest measure.

(c) What conclusions can you draw from the results of parts (a) and (b)?

Answer:

High confidence rules may not be interesting.

(d) Prove that if the confidence of the rule $\{a\} \longrightarrow \{b\}$ is less than the support of $\{b\}$, then:

i. $c(\{\overline{a}\} \longrightarrow \{b\}) > c(\{a\} \longrightarrow \{b\})$,

ii. $c(\{\overline{a}\} \longrightarrow \{b\}) > s(\{b\})$,

where $c(\cdot)$ denote the rule confidence and $s(\cdot)$ denote the support of an itemset.

Answer:

Let

$$c(\{a\} \longrightarrow \{b\}) = \frac{P(\{a,b\})}{P(\{a\})} < P(\{b\}),$$

which implies that

$$P(\{a\})P(\{b\}) > P(\{a,b\}).$$

Furthermore,

$$c(\{\overline{a}\} \longrightarrow \{b\}) = \frac{P(\{\overline{a},b\})}{P(\{\overline{a}\})} = \frac{P(\{b\}) - P(\{a,b\})}{1 - P(\{a\})}$$

i. Therefore, we may write

$$\begin{aligned} c(\{\overline{a}\} \longrightarrow \{b\}) - c(\{a\} \longrightarrow \{b\}) &= \frac{P(\{b\}) - P(\{a,b\})}{1 - P(\{a\})} - \frac{P(\{a,b\})}{P(\{a\})} \\ &= \frac{P(\{a\})P(\{b\}) - P(\{a,b\})}{P(\{a\})(1 - P(\{a\}))} \end{aligned}$$

which is positive because $P(\{a\})P(\{b\}) > P(\{a,b\})$.

ii. We can also show that

$$\begin{aligned} c(\{\overline{a}\} \longrightarrow \{b\}) - s(\{b\}) &= \frac{P(\{b\}) - P(\{a,b\})}{1 - P(\{a\})} - P(\{b\}) \\ &= \frac{P(\{a\})P(\{b\}) - P(\{a,b\})}{1 - P(\{a\})} \end{aligned}$$

is always positive because $P(\{a\})P(\{b\}) > P(\{a,b\})$.

18. Table 6.5 shows a $2 \times 2 \times 2$ contingency table for the binary variables A and B at different values of the control variable C.

(a) Compute the ϕ coefficient for A and B when $C = 0$, $C = 1$, and $C = 0$ or 1. Note that $\phi(\{A,B\}) = \frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$.

Answer:

i. When $C = 0$, $\phi(A,B) = -1/3$.

ii. When $C = 1$, $\phi(A,B) = 1$.

iii. When $C = 0$ or $C = 1$, $\phi = 0$.

(b) What conclusions can you draw from the above result?

Answer:

The result shows that some interesting relationships may disappear if the confounding factors are not taken into account.

Table 6.5. A Contingency Table.

			A	
			1	0
C = 0	B	1	0	15
		0	15	30
C = 1	B	1	5	0
		0	0	15

Table 6.6. Contingency tables for Exercise 19.

	B	$\overline{B}$
A	9	1
$\overline{A}$	1	89

(a) Table I.

	B	$\overline{B}$
A	89	1
$\overline{A}$	1	9

(b) Table II.

19. Consider the contingency tables shown in Table 6.6.

(a) For table I, compute support, the interest measure, and the ϕ correlation coefficient for the association pattern {A, B}. Also, compute the confidence of rules $A \rightarrow B$ and $B \rightarrow A$.

Answer:

$s(A) = 0.1$, $s(B) = 0.9$, $s(A, B) = 0.09$.

$I(A, B) = 9$, $\phi(A, B) = 0.89$.

$c(A \longrightarrow B) = 0.9$, $c(B \longrightarrow A) = 0.9$.

(b) For table II, compute support, the interest measure, and the ϕ correlation coefficient for the association pattern {A, B}. Also, compute the confidence of rules $A \rightarrow B$ and $B \rightarrow A$.

Answer:

$s(A) = 0.9$, $s(B) = 0.9$, $s(A, B) = 0.89$.

$I(A, B) = 1.09$, $\phi(A, B) = 0.89$.

$c(A \longrightarrow B) = 0.98$, $c(B \longrightarrow A) = 0.98$.

(c) What conclusions can you draw from the results of (a) and (b)?

Answer:

Interest, support, and confidence are non-invariant while the ϕ-coefficient is invariant under the inversion operation. This is because ϕ-coefficient

takes into account the absence as well as the presence of an item in a transaction.

20. Consider the relationship between customers who buy high-definition televisions and exercise machines as shown in Tables 6.19 and 6.20.

    (a) Compute the odds ratios for both tables.

    Answer:

    For Table 6.19, odds ratio = 1.4938.

    For Table 6.20, the odds ratios are 0.8333 and 0.98.

    (b) Compute the ϕ-coefficient for both tables.

    Answer:

    For table 6.19, $\phi = 0.098$.

    For Table 6.20, the ϕ-coefficients are -0.0233 and -0.0047.

    (c) Compute the interest factor for both tables.

    Answer:

    For Table 6.19, $I = 1.0784$.

    For Table 6.20, the interest factors are 0.88 and 0.9971.

    For each of the measures given above, describe how the direction of association changes when data is pooled together instead of being stratified.

    Answer:

    The direction of association changes sign (from negative to positive correlated) when the data is pooled together.