Association, Correlation and Causation

Review of Papers
"A Simulation Study of Three Related Causal Data mining Algorithms",
"Beyond Market Baskets: Generalizing Association Rules to Correlations"
"Scalable Techniques for Mining Causal Structures"

Meng Ding

Oct 24, 2002

Contents

Introduction – Association Rule

- **Based on support-confidence framework.**
  i_1 => i_2 if
  1. i_1 and i_2 occur together in at least s% of n transactions (support)
  2. and, of all the transactions containing i_1, at lease c% also contain i_2 (confidence)
- **Strong rules are NOT necessarily INTERESTING.**

Introduction – Association Rule

- **Example 1:**

Purchasing of Tea (t) and Coffee (c)

	c	$\bar{c}$	$\sum row$
t	20	5	25
$\bar{t}$	70	5	75
$\sum col$	90	10	100

Let's calculate the POTENTIAL association rule
t* => *c

Introduction – Correlation Rule

- **Strong association rules that are really INTERESTING.**
- **Definition:**

$$Corr_{A,B} = \frac{P(A \cup B)}{P(A)P(B)} = \begin{cases} < 1 & \text{Negatively Correlated} \\ = 1 & \textbf{Independent} \\ > 1 & \text{Positively Correlated} \end{cases}$$

Introduction – Correlation Rule

- **Correlation for Example 1:**

	c	$\bar{c}$	$\sum row$
t	20	5	25
$\bar{t}$	70	5	75
$\sum col$	90	10	100

	c	$\bar{c}$
t	.89	2
$\bar{t}$	1.04	.66

***t* => *c* is negatively correlated (0.89<1)**

Introduction – Correlation Rule

- **Correlation value alone can not tell if it is statistically significant.**

  χ^2 **(chi-square) statistics can be used to determine statistical significance**

- **Correlation is *upward closed.***

  Looking for Minimum Correlated Item-sets: item-sets with i variable are correlated but item-sets with i-1 variable are un-correlated.

Introduction – Causation

- **Causal Relationship between Item Sets.**

  Causal knowledge aids planning and decision making in almost all fields

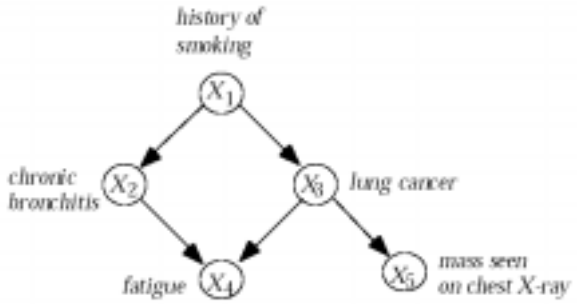
- **Correlation does not imply Causation, but may be helpful to constrain the possible Causal Relationships.**

  Simple Example: if item sets A and B are uncorrelated(independent), then there is no causal relationship between A and B.

Introduction – Causation

- **Causal Bayesian network**

  Acyclic, directed graph with each arc interpreting as direct causal influence between a parent node (variables) and a child node.



Contents

LCD Algorithm – Local Causality Discovery

- **Definition 1 (Markov Condition)**

  A variable is INDEPENDENT of its non-descendants, given just its parents.

  (Variables are independent ONLY IIF their independence is implied by the causal Markov condition.)

LCD Algorithm – Local Causality Discovery

- **CCC Causality**

  Suppose A, B, and C are three variables that are pair-wise dependent, and that A and C becomes independent when conditioned on B. Then we may infer that one of the following causal relations exists between A, B, and C:

  A ← B → C A → B → C A ← B ← C

  If through priori knowledge that A has no causes, then we will have the only possible relation:

  A → B → C

  That's to say: B causes C

LCD Algorithm – Local Causality Discovery

- **CCU Causality**

Suppose A, B, and C are three variables such that A and B are correlated, A and C are correlated, and B and C are uncorrelated, and that B and C become correlated when conditioned on A. Then we may infer that B and C cause A.

$$B \rightarrow A \leftarrow C$$

LCD Algorithm – Local Causality Discovery

- **Advantages**

CCC and CCU rules are local. They work only on three variables at the same time and do not explore the relationships between all variables of the causal Bayesian network. This makes the algorithm simple and fast.

LCD Algorithm – Local Causality Discovery

- **Limitations**

1. May produce disjoint picture of the causal relationship:
given W → X → Y → Z → A, the algorithm may produce X → A, X → Z, Y → Z, Y → A, Z → A.

2. Accumulated statistical error.

Contents

Determining Dependence & Independence

- **The χ^2 Test for Independence**

$$\chi^2 = \sum_{r \in R} \frac{(O(r) - E[r])^2}{E[r]}$$

$R = \{i_1, \bar{i}_1\} \times \cdots \times \{i_k, \bar{i}_k\}, r = r_1 \cdots r_k \in R.$ $E = Expectation$

E(r)= 9X(6/9)X(4/9)

	c	$\bar{c}$	$\sum row$
t	(1)	2	3
$\bar{t}$	4	(2)	**6**
$\sum col$	5	**4**	**9**

O(r) =1

Determining Dependence & Independence

- **The χ^2 is a normalized deviation from expectation, which is under assumption of independence.**

	c	$\bar{c}$	$\sum row$
t	1	2	3
$\bar{t}$	4	2	6
$\sum col$	5	4	9

real case

	c	$\bar{c}$	$\sum row$
t	1.67	1.33	3
$\bar{t}$	3.33	2.66	6
$\sum col$	5	4	9

expected case

χ^2 = 0.900

Contents

Algorithm Evaluation

- **Data Generation**

  Data is generated using existing causal Bayesian Network, e.g. ALARM monitoring system.

  Randomly choose node pairs from the existing causal network to form certain data size.

- **Evaluation Metrics**

Contents

- Introduction
- Causal Mining Algorithm – LCD
- Dependence Determination
- Algorithm Evaluation
- **Conclusion**

Conclusion

- **Relationship between Association, Correlation and Causation**
- **Causal Mining is a constraint-based mining approach**
- **For large data sets, LCD proves feasible and returns a large number of interesting causal relationships**
- **Finding causal relationship is very useful, but it is NOT easy. A lot of future work needs to be explored**

Thank you!