Characteristics of cosymmetric association rules

Michal Burda Marian Mindek Jana Šarmanová

VŠB – Technical University of Ostrava
Faculty of Electrical Engineering and Computer Science
Department of Computer Science

Dateso, 2005

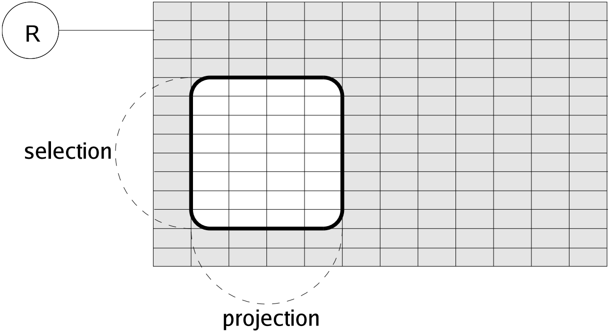
Outline

Outline

Probabilistic Logic of Typed Relations (PLTR)

- General language to express association rules of many types;
- Based on *Relational calculus*;
- Use of *probability* to express the intensity of rules;
- Formulae express rules found in data table as strong relationships between sub-tables.

Operations of Selection and Projection



Parts of typical PLTR Formula

$$R(\text{age} > 65)[\text{blood\_pressure}] >^{\star}_{mean} R(\text{age} < 21)[\text{blood\_pressure}]$$

typed relation – this notation expresses the source data the rules are mined from;

selection – pick up only the rows satisfying given condition;

projection – consider only the attributes listed in the brackets;

sub-relation – a part of typed relation described with relational operations;

relationship predicate – models the type of relationship between sub-tables.

Parts of typical PLTR Formula

$$R(\text{age} > 65)[\text{blood\_pressure}] >^{\star}_{mean} R(\text{age} < 21)[\text{blood\_pressure}]$$

typed relation – this notation expresses the source data the rules are mined from;

selection – pick up only the rows satisfying given condition;

projection – consider only the attributes listed in the brackets;

sub-relation – a part of typed relation described with relational operations;

relationship predicate – models the type of relationship between sub-tables.

Parts of typical PLTR Formula

$$R(\text{age} > 65)[\text{blood\_pressure}] >^{\star}_{mean} R(\text{age} < 21)[\text{blood\_pressure}]$$

typed relation – this notation expresses the source data the rules are mined from;

selection – pick up only the rows satisfying given condition;

projection – consider only the attributes listed in the brackets;

sub-relation – a part of typed relation described with relational operations;

relationship predicate – models the type of relationship between sub-tables.

Parts of typical PLTR Formula

$$R(\text{age} > 65)[\text{blood\_pressure}] >^{\star}_{mean} R(\text{age} < 21)[\text{blood\_pressure}]$$

typed relation – this notation expresses the source data the rules are mined from;

selection – pick up only the rows satisfying given condition;

projection – consider only the attributes listed in the brackets;

sub-relation – a part of typed relation described with relational operations;

relationship predicate – models the type of relationship between sub-tables.

Parts of typical PLTR Formula

$$R(\text{age} > 65)[\text{blood\_pressure}] >^{\star}_{mean} R(\text{age} < 21)[\text{blood\_pressure}]$$

typed relation – this notation expresses the source data the rules are mined from;

selection – pick up only the rows satisfying given condition;

projection – consider only the attributes listed in the brackets;

sub-relation – a part of typed relation described with relational operations;

relationship predicate – models the type of relationship between sub-tables.

Parts of typical PLTR Formula

$$R(\text{age} > 65)[\text{blood\_pressure}] >^{\star}_{mean} R(\text{age} < 21)[\text{blood\_pressure}]$$

typed relation – this notation expresses the source data the rules are mined from;

selection – pick up only the rows satisfying given condition;

projection – consider only the attributes listed in the brackets;

sub-relation – a part of typed relation described with relational operations;

relationship predicate – models the type of relationship between sub-tables.

Example of PLTR Formula

$$R(\text{age} > 65)[\text{blood\_pressure}] >^{\star}_{mean} R(\text{age} < 21)[\text{blood\_pressure}]$$

"Blood pressure of people older than 65 is in average significantly higher than blood pressure of people younger than 21."

Example of PLTR Formula

$$R(\text{age} > 65)[\text{blood\_pressure}] >^{\star}_{mean} R(\text{age} < 21)[\text{blood\_pressure}]$$

"Blood pressure of people older than 65 is in average significantly higher than blood pressure of people younger than 21."

Example of PLTR Formula

$$R(\text{age} > 65)[\text{blood\_pressure}] >^{\star}_{mean} R(\text{age} < 21)[\text{blood\_pressure}]$$

"Blood pressure of people older than 65 is in average significantly higher than blood pressure of people younger than 21."

Example of PLTR Formula

$$R(\text{age} > 65)[\text{blood\_pressure}] >^{\star}_{mean} R(\text{age} < 21)[\text{blood\_pressure}]$$

"Blood pressure of people older than 65 is in average significantly higher than blood pressure of people younger than 21."

Example of PLTR Formula

$$R(\text{age} > 65)[\text{blood\_pressure}] >^{\star}_{mean} R(\text{age} < 21)[\text{blood\_pressure}]$$

"Blood pressure of people older than 65 is in average significantly higher than blood pressure of people younger than 21."

Example of PLTR Formula

$$R(\text{age} > 65)[\text{blood\_pressure}] >^{\star}_{mean} R(\text{age} < 21)[\text{blood\_pressure}]$$

"Blood pressure of people older than 65 is in average significantly higher than blood pressure of people younger than 21."

Outline

1. Recall the logic of typed relations
   - Brief description of PLTR
2. The class of δ-cosymmetric rules
   - Motivation
   - Common properties
   - Examples

Motivation

- Many types of association rules in fact compare "something" against "something else".
- That is, two disjoint sets of objects are compared with respect to some attribute.
- What are their common properties?
- How to define the class of such association rules?

Example 1

"Non-smokers live in average longer."

In fact, the average life expectancy of smokers against the non-smokers is compared.

$$R(\text{smoker})[\text{life-expectancy}] <^{\star}_{mean} R(\neg\text{smoker})[\text{life-expectancy}]$$

Example 2

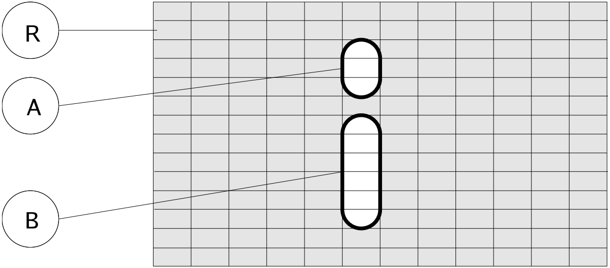
"The customer buying tequila often buys lemons, too."
(tequila $\Rightarrow$ lemon)

In fact, the probability of buying tequila and lemons is compared with the probability of buying tequila without lemons.

$R(\neg\text{lemon})[\text{tequila}] <^{\star}_{probability} R(\text{lemon})[\text{tequila}]$

General Schema of δ-Cosymmetric Rules

$$R(C_1)[X] <^{\star}_{some-characteristic} R(C_2)[X]$$



(here $A = R(C_1)[X]$ and $B = R(C_2)[X]$)

Outline

1. Recall the logic of typed relations
   - Brief description of PLTR

2. The class of δ-cosymmetric rules
   - Motivation
   - Common properties
   - Examples

Domain

Relationship predicate is a mapping that assigns truth value to several typed relations (data tables) given as arguments.

Domain of relationship predicate is a set of possible arguments.

Domain D of δ-cosymmetric rules should equal $D = K \times K$ for some $K \subseteq \mathcal{R}$, where $\mathcal{R}$ is a set of all typed relations.
That is, we can naturally ask for truth values of formulae

$$A <^\star B, \qquad B <^\star A, \qquad A <^\star A$$

if A, B are typed relations from K.

Minimum difference

Idea: Finding conditions for which some characteristic of some attribute is merely different does not always lead to interesting information.

Example: A group of people with life expectancy five days more than the rest population. It isn't interesting even if it passes a statistical test.

A δ-cosymmetric rule with minimum difference δ:

$$R(C_1)[X] <^{\star}_{\delta} R(C_2)[X]$$

Monotony

Idea: The increase of minimum difference δ leads to the reduction of the rule's probability.

Example: When it is very probable that Europeans are over *20 cm* taller than Asiatic, it is *even more* probable that Europeans are over *10 cm* taller than Asiatic.

Let F_1, F_2 be PLTR formulae. The fact that F_1 is at least as probable as F_2 is denoted with $F_1 \succeq F_2$.

$$\delta_1 < \delta_2 \ \Rightarrow\ \left(A <^{\star}_{\delta_1} B\right) \succeq \left(A <^{\star}_{\delta_2} B\right).$$

Non-symmetricity

Idea: Exchanging the direction of the relationship predicate negates the truth value.

Example: Let the following is very probable in data:
$R(\text{smoker})[\text{life-expect.}] <^{\star}_{mean} R(\neg\text{smoker})[\text{life-expect.}]$.
Then naturally, the probability of the rule
$R(\text{smoker})[\text{life-expect.}] >^{\star}_{mean} R(\neg\text{smoker})[\text{life-expect.}]$
should be very low.

$$B <^{\star} A \;\Leftrightarrow\; \neg\left(A <^{\star} B\right) \qquad \text{or} \qquad B <^{\star}_{\delta} A \;\Leftrightarrow\; \neg\left(A <^{\star}_{-\delta} B\right).$$

Quasi-transitivity

Idea: If $A <^{\star}_{\delta} B$ and $B <^{\star}_{\delta} C$ are rather probable then $A <^{\star}_{\delta} C$ isn't improbable.

Example: If the temperature in winter is very probably lower than in spring and if temperature in spring is very probably lower than in summer then also the winter's temperature is very probably lower than the summer's.

Problem: In special cases not satisfied. When using rank tests (e.g. Mann–Whitney's test), paradoxes may occur.

The Definition of δ-Cosymmetric Predicates
(The First Prototype)

Definition

A relationship predicate is called δ-cosymmetric if it has domain $D = K \times K$, where $K \subseteq \mathcal{R}$, and it satisfies conditions of monotony, non-symmetricity and quasi-transitivity.

Outline

1. Recall the logic of typed relations
   - Brief description of PLTR

2. The class of δ-cosymmetric rules
   - Motivation
   - Common properties
   - Examples

Aspin–Welch predicate I

- Aspin–Welch statistical test – two-sample test on means similar to Student's t test.
- Assumes the two random samples X and Y to be normally distributed (no need of equal variances).
- $H_0 : \mathrm{E}X - \mathrm{E}Y = \delta$ against $H_A : \mathrm{E}X - \mathrm{E}Y \neq \delta$

$$T = \frac{\bar{X} - \bar{Y} - \delta}{S}, \quad \text{where} \quad S = \sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}.$$

H_0 is rejected if $|T| \geq t_f(1 - \frac{\alpha}{2})$.

Aspin–Welch predicate II

Definition

Predicate $<^{\star}_{AW;\delta}$ is a function where a probability p is mapped the following way to each pair of typed relations $\langle X, Y \rangle$, which both are non-empty and both contain just one column.

$$<^{\star}_{AW;\delta} (X, Y) = p$$

for such p where $T = t_f(p)$ for T, f and t_f as above.

Aspin–Welch predicate III

Usage: Suppose we have a data table D about patients suffering certain disease. One may enquire the validity of the following rule:

$$D(\mathsf{sex} = \text{“male”})[\mathsf{pressure}] >_{AW;0} D(\mathsf{sex} = \text{“female”})[\mathsf{pressure}].$$

Theorem

Aspin–Welch relationship predicate $<^{\star}_{AW;\delta}$ *is* δ*-cosymmetric.*

Funded Implication I

- The rule $\varphi \Rightarrow_{p,base} \psi$ is true iff $\frac{a}{a+b} \geq p \wedge a \geq Base$.

Table: 4-field table of φ and ψ

	ψ	$\neg\psi$
φ	a	b
$\neg\varphi$	c	d

Funded Implication II

Definition

Let A and B be the typed relations, each containing exactly one column with values from the set $\{0, 1\}$ and let $\delta \in [-1, 1]$. Let $\mathrm{sum}(A)$ denotes the number of A's rows possessing "1". The *Funded predicate* $<^{\star}_{fnd;\delta}$ is defined:

$$
\begin{aligned}
>^{\star}_{fnd;\delta}(A, B) = 1 \quad & \text{iff} \quad \frac{\mathrm{sum}(A)}{\mathrm{sum}(A) + \mathrm{sum}(B)} > \frac{1+\delta}{2}, \\
>^{\star}_{fnd;\delta}(A, B) = \frac{1}{2} \quad & \text{iff} \quad \frac{\mathrm{sum}(A)}{\mathrm{sum}(A) + \mathrm{sum}(B)} = \frac{1+\delta}{2}, \\
>^{\star}_{fnd;\delta}(A, B) = 0 \quad & \text{iff} \quad \frac{\mathrm{sum}(A)}{\mathrm{sum}(A) + \mathrm{sum}(B)} < \frac{1+\delta}{2}.
\end{aligned}
$$

Funded Implication III

Theorem

The Funded predicate $<^{\star}_{fnd;\delta}$ is δ-cosymmetric.

That is, the rule

$$\varphi \Rightarrow_{p,0} \psi$$

equals to

$$R(\psi)[\varphi] >^{\star}_{fnd;(2p-1)} R(\neg\psi)[\varphi].$$

Summary

This paper has presented:

- Brief description of PLTR language for association rules expression;
- δ-cosymmetric rules as a general notion of many association rule types.

For Further Reading

- Michal Burda, Marian Mindek, Jana Šarmanová.
  Using relational operations to express association rules.
  To appear in the proceedings of SYRCoDIS, Russia, 2005.

- Michal Burda, Martin Hynar, Jana Šarmanová.
  Pravděpodobnostní logika typovaných relací.
  Znalosti poster proceedings, Slovakia, 2005.

- Yonatan Aumann and Yehuda Lindell.
  A Statistical Theory for Quantitative Association Rules.
  Knowledge Discovery and Data Mining, 1999.