Exercises of Data Mining

Li Yihai[1], Zhang Chao[2], and Shang Chenyang[3]

[123] *Institute of Mathematical Sciences From ShanXi University*

November 9, 2018

Contents

1 Preliminary Work

1.1 Data Matrix

Assume we observing children who have an allergic reaction to,say,tomato,apple,orange,cheese or milk.These observations are presented in data matrix as following table 1.

Table 1 : Data Matrix

Child	Tomato	Apple	Orange	Cheese	Milk
Anna	1	1	0	1	1
Aina	1	1	1	0	0
Naima	1	1	1	1	1
Rauha	0	1	1	0	1
Kai	0	1	0	1	1
Kille	1	1	0	0	1
Lempi	0	1	1	1	1
Ville	1	0	0	0	0
Ulle	1	1	0	1	1
Dulle	1	0	1	0	0
Dof	1	0	1	0	1
Kinge	0	1	1	0	1
Laade	0	1	0	1	1
Koff	1	1	0	0	1
Olvi	0	1	1	1	1

1.2 Definition

Definition 1.1. Atomic (Open) Formulas Child x is allergic to milk and Child y is allergic to cheese, write shorter Milk(x) and Cheese(y).

Definition 1.2. Unary Predicates Mikk(-),Cheese(-),Tomato(-),Orange(-)and Apple(-) are unary predicates of our observational language and x,y,z,$\cdots$ are variables.

Definition 1.3. Boolean Attributes Given the $0/1-$ data matrix, each pair of formulas called (also Boolean attributes)ϕ,ψ determines a four-fold frequency table of the form:

Table 2 : Four-fold Frequency

	ψ	$\neg\psi$	
ϕ	a	b	$a+b=r$
$\neg\phi$	c	c	$c+d=s$
	$a+c=k$	$b+d=l$	m

where m is the amount of rows in the data matrix,and
- a is the number of objects satisfying both ϕ and ψ.
- b is the number of objects satisfying both ϕ but not ψ.

• c is the number of objects not satisfying ϕ but ψ.
• d is the number of objects not satisfying ϕ nor ψ.
The truth value $v(\phi \sim \psi) =$ TRUE,FALSE is based on this table.

Definition 1.4. Several Possibilities of $\sim$:
• $\Rightarrow_{p,\text{Base}}$,where Base $\in \mathbb{N}$,$0 < p < 1$, p rational: $\phi(x) \Rightarrow_{p,\text{Base}} \psi(x)$.Read:$\phi(x)$ implies $\psi(x)$ with confidence p and support Base.
• Given a data matrix M, $v(\phi(x) \Rightarrow_{p,\text{Base}} \psi(x)) =$ TURE, iff $\frac{a}{a+b} \geq p$ and $a \geq$ Base.
• $\equiv_p$, where $0 < p \leq 1$, In any Model M,$v((\phi(x) \equiv_p \psi(x))) =$ TRUE iff $(a + d)/p(a + b + c + d)$ except for a case $(a + d) = 0$,$b + c \neq 0$;then $v((\phi(x) \equiv_p \psi(x))) =$ FAUSE.
• The exact truth definition of these quantifiers is the following

$$v((\phi(x) \sim_p \psi(x))) = \text{FAUSE}, \text{iff} \frac{a}{a+b} \geq \frac{(1+p)(a+c)}{(a+b+c+d)}, a \geq \text{Base} \qquad (1)$$

2 Solution

2.1 Question 1

In first exercise, we are asked to construct the four-fold frequency table for ϕ=Milk$(x)\wedge\neg$ Cheese(x),ψ=Apple$(x)\vee$Orange(x).

2.1.1 Solution of Question 1

From the question, ϕ represents a student who is allergic to milk and is not allergic to cheese, and ψ represents an allergy to apple or an allergy to orange. We import the data in to LISp_Miner, set Founded Implication $p = 1.000$ and Base= 5,Antecedent=ϕ=Milk$(x)\wedge\neg$ Cheese(x).Succedentψ=Apple$(x)\vee$Orange(x), based on the above relationship analysis, the when we get result in Figure 1 follow.

Table 3 : Four-fold Frequency Table for Question 1

	ψ	$\neg\psi$	
ϕ	5	0	5
$\neg\phi$	9	1	10
	14	1	15

This figure has meanings:with the 100% Confidence, we have conclusions below:
• 5 is the number of children satisfying both ϕ and ψ: There are five students who are allergic to milk but not to cheese, while have an allergy to apples or oranges or both.
• 0 is the number of child satisfying ϕ but not ψ: There is nobody satisfying the conditions.

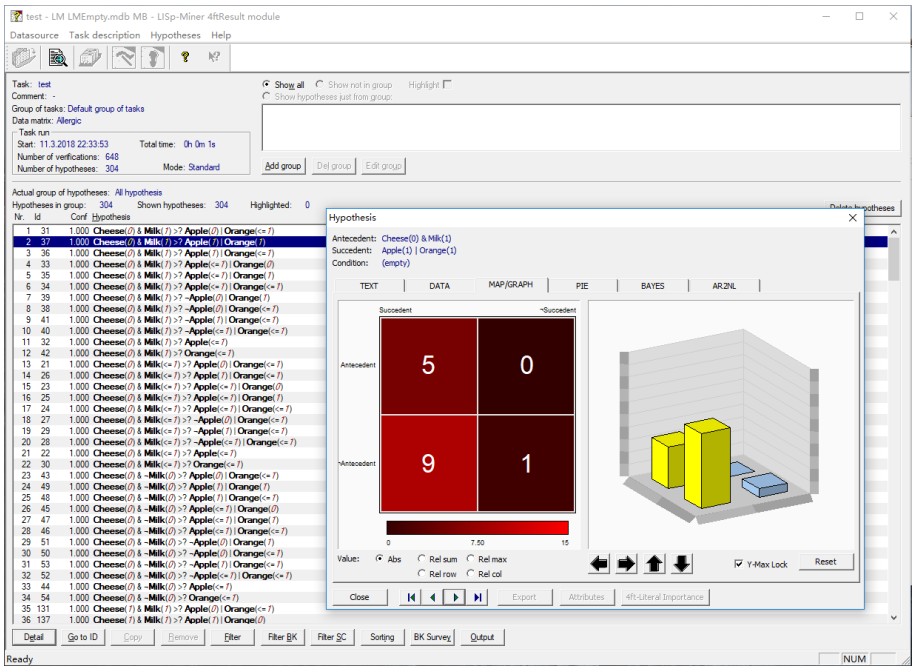

Figure 1: Hypotheses 1 From LISp_miner

• 9 is the number of children not satisfying ϕ but satisfying ψ: There are nine students who have allergic to apples or oranges or both, then when one of them is allergic to milk he is not allergic to cheese.
• 1 is the number of child not satisfying ϕ nor ψ: There is one students who has no allergy to apples or oranges, then when he is allergic to milk he is not allergic to cheese.

2.2 Question 2

In second exercise, we are asked to construct the four-fold frequency table for ϕ=Apple(x),ψ=Cheese(x)

2.2.1 Solution of Question 2

From the question, ϕ represents a student who is allergic to apple, and ψ represents an allergy to cheese. We import the data in to LISp_Miner, set Founded Implication $p = 1.000$ and Base= 5, based on the above relationship analysis, the when we get result in figure 2 follow.Antecedent=ϕ=Apple(x), Succedent = ψ=Cheese(x).
This figure has meanings:with the 100% Confidence, we have conclusions below:
• 7 is the number of children satisfying both ϕ and ψ: There are seven students who has no allergy to apples and cheese.

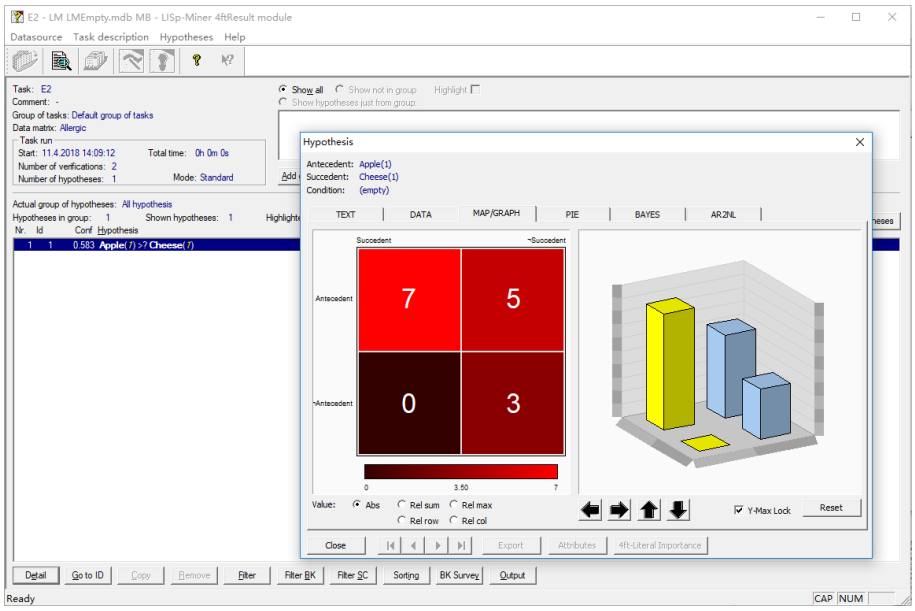

Figure 2: Conclusion From Lisp_miner

• 5 is the number of children satisfying ϕ but not ψ: There five students who has no allergy to cheese but is allergic to apples.
• 0 is the number of child not satisfying ϕ but satisfying ψ: There is nobody allergy satisfying the conditions.
• 3 is the number of children not satisfying ϕ nor ψ: There are three students not allergic to apples or cheese.

Table 4 : Four-fold Frequency for Question 2

	ψ	$\neg\psi$	
ϕ	7	5	12
$\neg\phi$	0	3	3
	7	8	15

2.3 Question 3

What is the truth value of $\text{Apple}(x) \Rightarrow_{0.7,4} \text{Cheese}(x)$?

2.3.1 Solution of Question 3

Assign Founded Implications to 0.70 and Base to 4.0 using LISp_Miner.
According to the hypotheses, we analyze the relationship between apple and cheese. We can get the result as the figure 3 showing, which mean there is no hypotheses is true.

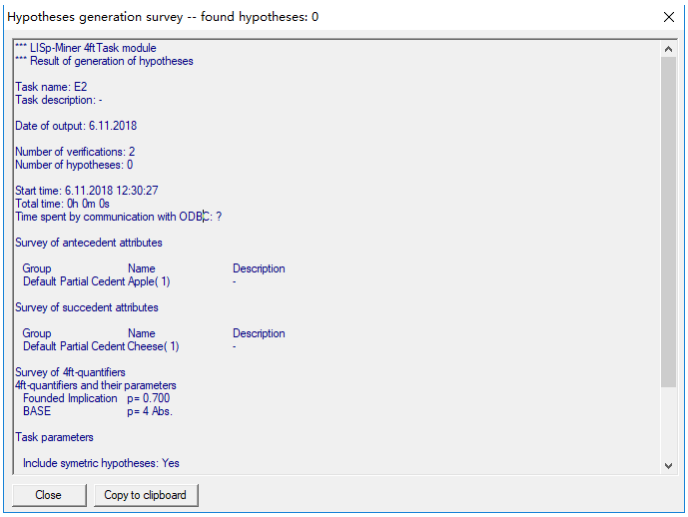

Figure 3: Hypotheses 3 From LISp_Miner

The analysis to results points that: assuming the people who are allergic to apples and cheese are a while people who are allergic to apples but not to cheese are b, we can get the conclusion that there are no hypotheses output from $\frac{a}{a+b} \geq 0.7$ and $a \geq 4$,also meaning

$$v(\text{Apple}(x) \Rightarrow_{0.7,4} \text{Cheese}(x)) = \texttt{FAUSE}. \tag{2}$$

Assign Founded Implications to minimum available value 0.01 and Base as same as above,using LISp_Miner.We can get the maximum value of $p_{max} = 0.583$ which make hypotheses true and the result as the figure 4 showing the range of p is $(0, 0.583]$.

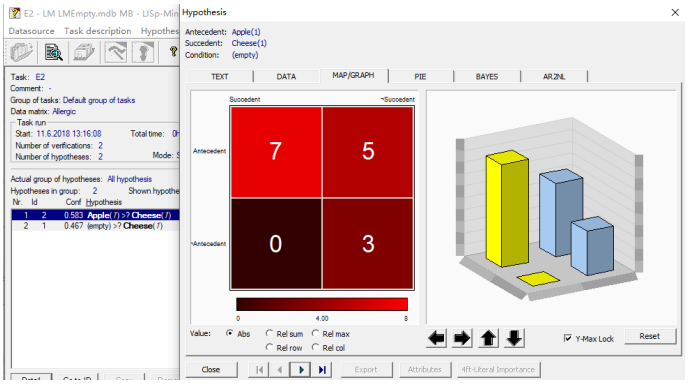

Figure 4: Hypotheses 3 $p = p_{max}$

Sign $f(x)_{0.7,4}$=$v(\text{Apple}(x) \Rightarrow_{0.7,4} \text{Cheese}(x)) = \texttt{FAUSE}$,then we have

$$f(x)_{0.7,4} = \begin{cases} \texttt{TRUE} & 0 < p \leq 0.583 \\ \texttt{FAUSE} & 0.583 < p \leq 1 \end{cases} \tag{3}$$

2.4 Question 4

What is the truth value of $\text{Apple}(x) \equiv_{0.6} \text{Cheese}(x)$?

2.4.1 Solution of Question 4

According to the question,we use LISp_Miner to do Basic Equivalence Quantifiers Analysis, and assign Founded Equivalence Quantifiers to 0.600. and according to the four-fold frequency table about $\text{Apple}(x)$ and $\text{Cheese}(x)$ built by third question which is table 4, in which we know $a = 7$,$b = 5$,$c = 0$,$d = 3$, we can calculate the p end up to the result by formula 4 below

$$\frac{a+d}{a+b+c+d} \geq p \tag{4}$$

Thus $p \in [0, \frac{2}{3}]$,also mean $v(\text{Apple}(x) \equiv_{0.6} \text{Cheese}(x)) = \texttt{TRUE}$.Alsp show in figure 5 below.

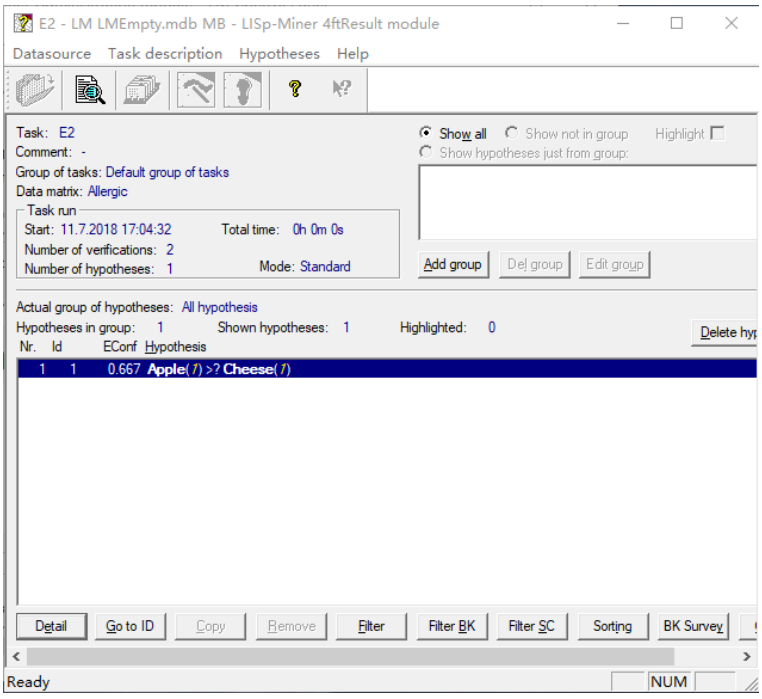

Figure 5: $v(\text{Apple}(x) \equiv_{0.6} \text{Cheese}(x)) = \texttt{TRUE}$

2.5 Question 5

Define p such that $v((\text{Apple}(x) \sim_p \text{Cheese}(x))=\texttt{TURE}$.

2.5.1 Solution of Question 5

According to the four-fold frequency table about Apple(x) and Cheese(x) built by third question, in which we know $a = 7$,$b = 5$,$c = 0$,$d = 3$, we can calculate the above average quantifiers end up to the result by formula

$$\frac{a}{a+b} \geq \frac{(1+p)(a+c)}{a+b+c+d}, a \geq \text{Base} \tag{5}$$

Thus $v(\text{Apple}(x) \equiv_p \text{Cheese}(x))=\texttt{TRUE}$ while $p \in (0, 0.250]$.

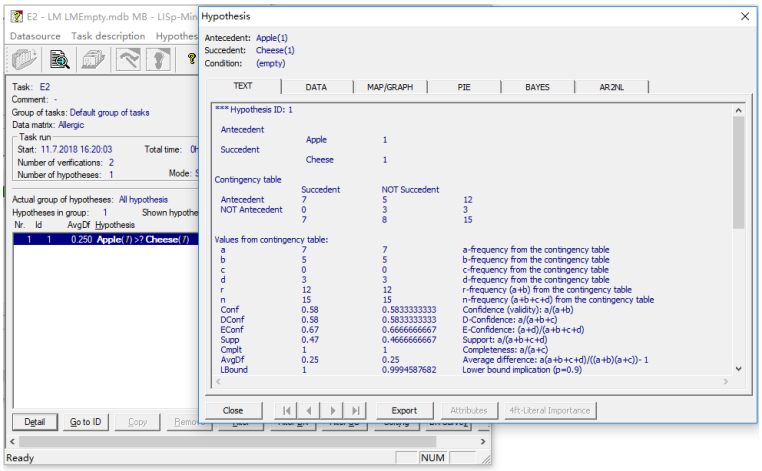

Figure 6: Hypotheses 5 $p = 0.200$ Confidence= 0.250