Low Cost High Utility Pattern Mining

Jiaxuan Li[1]
Philippe Fournier-Viger[1]
Jerry Chun-Wei Lin[2]
Tin Truong Chi[3]

[1]Harbin Institute of Technology (Shenzhen), China
[2]University of Applied Sciences (HVL), Bergen, Norway
[3] University of Dalat, Vietnam

Motivation

- **High utility pattern mining**: to discover patterns that have a high utility.
- However, it ignores the cost or effort required to obtain these benefits.
- May find patterns that have:
  - **a high utility but a very high cost**
- **Cost of a pattern,** can be expressed in terms of aspects such as time, money, resources consumed and effort.

Sequential Activity Database (SADB)

A **sequence** is a series of activities, each having a cost

The **utility** of a sequence is a binary class or a positive number.

Sid	<Activity : cost>	Utility
S_1	<(a:4)(b:2)(e:4)(c:4)(d:5)>	Positive
S_2	<(b:3)(c:2)(f:1)(d:1)(e:2)>	Negative
S_3	<(a:2)(f:2)(e:1)(c:3)(d:5)>	Positive
S_4	<(a:2)(b:2)(c:1)(f:2)>	Negative

(e.g. cured or died after some medical treatments)

Sid	<Activity : cost>	Utility
S_1	<(a:4)(b:2)(e:4)(c:4)(d:5)>	40
S_2	<(b:3)(c:2)(f:1)(d:1)(e:2)>	50
S_3	<(a:2)(f:2)(e:1)(c:3)(d:5)>	60
S_4	<(a:2)(b:2)(c:1)(f:2)>	70

(e.g. score obtained at an exam)

Pattern Cost

Sid	<Activity : cost>	Utility
S_1	<(**a**:**4**)(**b**:**2**)(e:4)(c:4)(d:5)>	…
S_2	<(b:3)(c:2)(f:1)(d:1)(e:2)>	…
S_3	<(a:2)(f:2)(e:1)(c:3)(d:5)>	…
S_4	<(**a**:**2**)(**b**:**2**)(c:1)(f:2)>	…

Pattern Cost

$$c(p, S_s) = \sum_{v_i \in first(p, S_s)} c(v_i, S_s)$$

e.g. $c(ab)=c(4+2,S_1)+c(2+2,S_4)=10$

Average Cost

$$ac(p) = \frac{\sum_{p \subseteq S_s \wedge S_s \in SADB} c(p, S_s)}{|\sup(p)|}$$

e.g. $ac(ab)=c(6,S_1)+c(4,S_4) / 2=5$

Three problems

Discover **low-cost high utility patterns (LCHUPs)** when:

1. The **utility** is **binary classes**. Only records representing the **positive class** are used.
2. The **utility** is **binary classes.** All records are used.
3. The **utility** is **numeric values.**

(1) Positive Patterns in a Binary SADB

Find each LCHUP p such that:

$$\sup(p) \geq minsup$$

$$ac(p) \leq \text{maxcost}$$

Sid	<Activity : cost>	Utility
S_1	<(a:4)(b:2)(e:4)(c:4)(d:5)>	Positive
S_3	<(a:2)(f:2)(e:1)(c:3)(d:5)>	Positive

e.g. minsup=2 maxcost=7

Pattern	sup	ac	Pattern	sup	ac
a	2	3.0	e	2	2.5
c	2	3.5	ac	2	6.5
d	2	5.0	ae	2	5.5
ec	2	6.0			

(1) Limitations of positive patterns in a Binary SADB

1. Some positive patterns may be misleading to users as they may also appear in negative sequences.

2. The correlation between a pattern and utility is not measured.

(2) Correlated Patterns in a Binary SADB

Find each LCHUP p such that:

$$\sup(p) \geq minsup$$

$$ac(p) \leq \text{maxcost}$$

Sid	<Activity : cost>	Utility
S_1	<(**a:4**)(b:2)(e:4)(**c:4**)(d:5)>	Positive
S_2	<(b:3)(c:2)(f:1)(d:1)(e:2)>	Negative
S_3	<(**a:2**)(f:2)(e:1)(**c:3**)(d:5)>	Positive
S_4	<(**a:2**)(b:2)(**c:1**)(f:2)>	Negative

The **correlation** of a pattern p:

$$cor(p) = \frac{ac(D_p^+) - ac(D_p^-)}{Standard\ Deviationcost} \sqrt{\frac{\sup(D_p^+)}{|D_p|}\frac{\sup(D_p^-)}{|D_p|}}$$

$where, ac(D_p^+), ac(D_p^-)\ denotes\ pattern\ p's\ average\ cost\ in\ positive\ and\ negative\ sequences, respectively.$

(2) Correlated Patterns in a Binary SADB

Sid	<Activity : cost>	Utility
S_1	<(a:4)(b:2)(e:4)(c:4)(d:5)>	Positive
S_2	<(b:3)(c:2)(f:1)(d:1)(e:2)>	Negative
S_3	<(a:2)(f:2)(e:1)(c:3)(d:5)>	Positive
S_4	<(a:2)(b:2)(c:1)(f:2)>	Negative

e.g. *minsup = 3 maxcost = 10*

Pattern	sup	ac	cor	Pattern	sup	ac	cor
a	3	2.7	0.50	f	3	1.7	0.50
b	3	2.3	-0.50	ac	3	5.3	0.80
c	4	2.5	0.89	bc	3	4.7	0.76
d	3	3.7	1.00	cd	3	6.7	0.99
e	3	2.3	0.19				

(3)LCHU pattern mining in a Numeric SADB

Find each LCHUP p such that:

$$\sup(p) \geq minsup$$

$$ac(p) \leq \text{maxcost}$$

Sid	<Activity : cost>	Utility
S1	<(a:4)(b:2)(e:4)(c:4)(d:5)>	40
S2	<(b:3)(c:2)(f:1)(d:1)(e:2)>	50
S3	<(a:2)(f:2)(e:1)(c:3)(d:5)>	60
S4	<(a:2)(b:2)(c:1)(f:2)>	70

Utility of a pattern p:

$$u(p) = \frac{\sum_{p \subseteq S_s \wedge S_s \in SADB} su(S_s)}{|\sup(p)|}$$

where $su(S_s)$ *is the utility of the sequence,*
e.g. $su(S_1) = 40.$

Trade-off of a pattern p:

$$tf(p) = \frac{ac(p)}{u(p)}$$

(3)LCHU pattern mining in a Numeric SADB

Sid	<Activity : cost>	Utility
S1	<(a:4)(b:2)(e:4)(c:4)(d:5)>	40
S2	<(b:3)(c:2)(f:1)(d:1)(e:2)>	50
S3	<(a:2)(f:2)(e:1)(c:3)(d:5)>	60
S4	<(a:2)(b:2)(c:1)(f:2)>	70

minsup=3 maxcost=10

Utility:50		Utility:53		Utility:55		Utility:56		Utility:60	
pattern	**tf**	**pattern**	**tf**	**pattern**	**tf**	**pattern**	**tf**	**pattern**	**tf**
e	**0.05**	**b**	**0.04**	**c**	**0.05**	**a**	**0.05**	**f**	**0.03**
d	0.07	bc	0.09			ac	0.09		
cd	0.13								

How to find the patterns efficiently?

We introduce a lower-bound on the cost of a pattern p:

Average Supported Cost

$$ASC(p) = \frac{1}{minsup} \sum_{i=1,2,..,minsup} c(p, S_i)$$

where c(p,S_i) are sorted in ascending order.

e.g. *minsup = 2*

c(bc,S_1)=6, c(bc,S_2)=5, c(bc,S_4)=3

ASC(bc)=(3+5) / 2 = 4

Sid	<Activity : cost>	Utility
S_1	<(a:4)(**b:2**)(e:4)(**c:4**)(d:5)>	…
S_2	<(**b:3**)(**c:2**)(f:1)(d:1)(e:2)>	…
S_3	<(a:2)(f:2)(e:1)(c:3)(d:5)>	…
S_4	<(a:2)(**b:2**)(**c:1**)(f:2)>	…

How to find the patterns efficiently?

$$ASC(p) = \frac{1}{minsup} \sum_{i=1,2,..,minsup} c(p, S_i)$$

Properties of the ASC:

I. **Underestimation**: The ASC of a pattern p is smaller than or equal to its cost, $ASC(p) \leq c(p)$

II. **Monotonicity**: Let p_x and p_y be two patterns,
$If \;\; p_x \subset p_y \;\; then \;\; ASC(p_x) \leq ASC(p_y)$

III. **Pruning**: For a pattern p, if $ASC(p)$ > maxcost, then pattern p can be eliminated as well as its supersets.

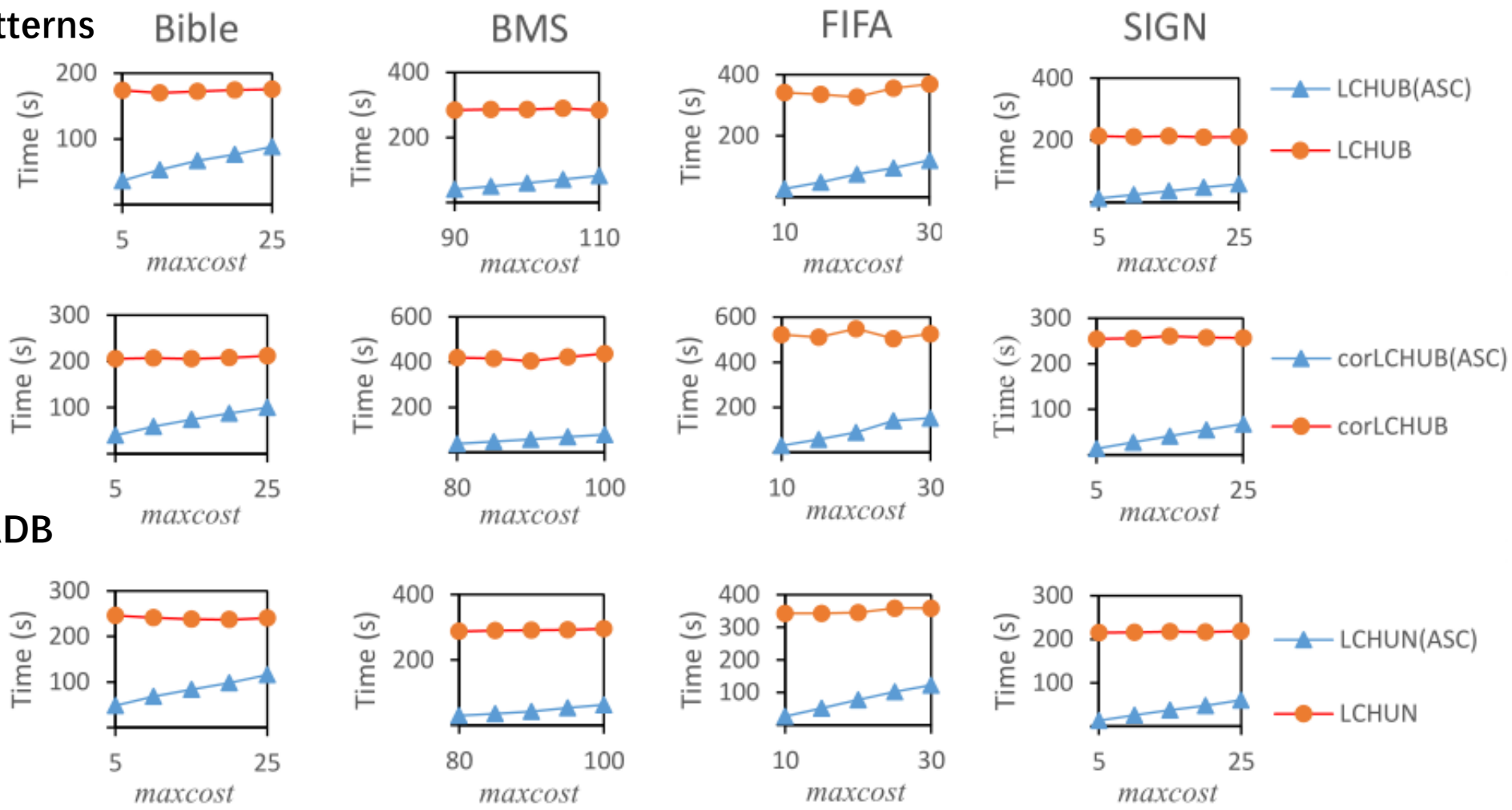
Performance Analysis

We varied the *maxcost* parameter on four benchmark databases (Bible, BMS, FIFA, SIGN)

Problem (1) positive patterns

Problem (2) positive or negative patterns

Problem (3) numeric SADB



Case study 1: binary e-learning SADB

Database

- **115 students**
- **A sequence is a series of learning sessions.**
- **Cost: time to complete a session.**
- **Utility: to *pass* or *fail* the final exam.**

Some patterns for *minsup = 0.5 maxcost = 600 s*

Pattern	Support	Avg. Cost	Correlation
(1)(6)	39	250	0.210
(1)(2)(5)(6)	34	485	0.209
(2)(6)	41	298	0.207
(1)(2)(6)	36	391	0.204
(2)(3)	40	284	0.001
(3)(4)(5)(6)	40	469	$5.32*10^{-4}$
(4)(5)	49	171	-0.109
(5)	53	96	-0.147

Case study 2: numeric e-learning SADB

- **A sequence is the activities of a learning session.**
- **The cost is the time to complete an activity.**
- **The utility is the score obtained at the final exam.**

Some patterns found for *minsup = 0.5 maxcost = 600 s*

```
Utility: 10, Trade-Off:2.01, (Study_Es_6_1 )(Deeds_Es_6_2 )
Utility: 11, Trade-Off:1.56, (Study_Es_6_2 )(Study_Es_6_3 )
Utility: 12, Trade-Off:0.69, (Study_Es_6_2 )
Utility: 13, Trade-Off:0.64, (Study_Es_6_3 )
Utility: 14, Trade-Off:0.62, (Deeds_Es_6_2 )
Utility: 15, Trade-Off:1.74, (Study_Es_6_2 )(Deeds_Es_6_2 )(Study_Es_6_3 )
Utility: 16, Trade-Off:3.94, (FSM_Es_6_1 )(FSM_Es_6_1 )(Deeds_Es_6_2 )(Study_Es_6_3 )
Utility: 17, Trade-Off:0.89, (Deeds_Es_6_2 )(Study_Es_6_3 )
Utility: 18, Trade-Off:1.04, (Study_Es_6_3 )(Study_Es_6_3 )
Utility: 20, Trade-Off:1.59, (Deeds_Es_6_1 )(Study_Es_6_3 )(Study_Es_6_3 )
Utility: 21, Trade-Off:4.48, (FSM_Es_6_3 )(Study_Es_6_3 )(Study_Es_6_3 )
Utility: 23, Trade-Off:1.15, (Deeds_Es_6_2 )(Study_Es_6_3 )(Study_Es_6_3 )
Utility: 24, Trade-Off:3.60, (FSM_Es_6_1 )(Deeds_Es_6_1 )(Study_Es_6_3 )(Study_Es_6_3 )
Utility: 28, Trade-Off:1.35, (Deeds_Es_6_1 )(Deeds_Es_6_2 )(Study_Es_6_3 )(Study_Es_6_3 )
```

Conclusion

- We proposed to mine **low cost-high utility patterns**.
- We defined three versions of this problem, which correspond to three different real-life scenarios.
- We defined a novel ASC lower-bound on the average cost of patterns to reduce the search space and discover patterns efficiently.
- A case study with e-learning data has shown that useful patterns can be found having a low cost and a high utility. Those patterns can provide insights to students and teachers about how to use learning material more efficiently.

Future Work

- Develop a tighter lower-bound to improve the performance
- Apply our approach to Activityset or sequential rule mining
- Add other constraints such as the length of patterns to the proposed model

Thank you for listening!