

Overview

- Combine a utility and cost model to mine cost-effective guidance patterns in E-learning event logs.
- Design measures to evaluate the correlation between patterns and the binary utility.
- Design a pruning strategy to improve algorithm's performance.
- A case study in real life E-learning dataset shows some interesting pattern.
- Detailed experiments and the visualization patterns' properties demonstrate our models' accuracy and usefulness.

High-utility sequential pattern mining

Input

Quantitative sequences with purchase quantities (internal utility)
sequence 1: $\langle (a, 3), (b, 3), (c, 1), (b, 4) \rangle$
sequence 2: $\langle (a, 1), (e, 3) \rangle$
sequence 3: $\langle (a, 6), (c, 7), (b, 8), (d, 9) \rangle$
sequence 4: $\langle (b, 3), (c, 1) \rangle$
Unit profits (external utility)
$a = 5\$, b = 1\$, c = 2\$, d = 1\$$

a minimum utility threshold (e.g. $minutil = 30$)

Output

All **sequences** having a *utility* $\geq minutil$)

High-utility sequential pattern mining

Input

Quantitative sequences with purchase quantities (internal utility)

sequence 1: $\langle (a, 3), (b, 3), (c, 1), (b, 4) \rangle$

sequence 2: $\langle (a, 1), (e, 3) \rangle$

sequence 3: $\langle (a, 6), (c, 7), (b, 8), (d, 9) \rangle$

sequence 4: $\langle (b, 3), (c, 1) \rangle$

Unit profits (external utility)

$a = 5\$$ $b = 1\$$, $c = 2\$$, $d = 1\$$

a minimum utility threshold (e.g. $minutil = 30$)

Output

All **sequences** having a *utility* $\geq minutil$)

The **sequence** $\langle ab \rangle$ is a high utility pattern because:

$$u(\langle ab \rangle) = \underbrace{3 \times 5 + 3 \times 1}_{\text{Sequence 1}} + \underbrace{6 \times 5 + 8 \times 1}_{\text{Sequence 3}} = 56 > minutil$$

Limitations

- **High utility pattern mining** aims at discovering patterns that have a high utility.
- But 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:** *time, money, resources consumed or effort.*

Our proposal: Find Cost-effective Patterns →

Sequential Activity Database

- A **sequence** is a series of activities, each having a cost.
- The **utility** of a sequence is a **binary class**.

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)

Two problems

Discover **low-cost high utility patterns** 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.

The *support* measure

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)>	...

The **support** of a pattern p :

$$\text{sup}(p) = |\{S_s | p \subseteq S_s \in DB\}|$$

(number of sequences containing p)

e.g. $\text{sup}(\langle \mathbf{ab} \rangle) = |\{S_1, S_4\}| = 2$ sequences

This measure is used to remove noise.

The *cost* measure

Sid	<Activity : cost>	...
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)>	...

The **cost** of a pattern p :

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

$$c(\mathbf{ab}, S_1) = 4 + 2 = 6$$

This measure is used to assess the effort or resource spent.

The **average cost** of a pattern p :

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

$$ac(\mathbf{ab}) = \underbrace{6}_{\text{Sequence 1}} + \underbrace{4}_{\text{Sequence 4}} / 2 = 5$$

Sequence 1 Sequence 4

Problem1:

Positive Patterns in a binary DB

Find each p such that:

$$\text{sup}(p) \geq \text{minsup}$$

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

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			

Problem1:

Limitations of positive patterns in a binary DB

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.

Problem 2:

Finding all cost-effective patterns in a **binary DB**

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

A pattern p is **cost-effective** if:

$$\text{sup}(p) \geq \text{minsup}$$

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

$$\text{occup}(p) \geq \text{minoccup}$$

Furthermore, we measure the **correlation** of a pattern p with the desirable outcome:

$$\text{cor}(p) = \frac{\text{ac}(D_p^+) - \text{ac}(D_p^-)}{\text{Std}} \sqrt{\frac{|D_p^+| |D_p^-|}{|D_p^+ \cup D_p^-|}} \in [-1, 1]$$

a positive correlation is desirable

Pattern	support	average cost	correlation
<ac>	3	5.3	0.80

More details...

The **correlation** of a pattern p :

$$cor(p) = \frac{ac(D_p^+) - ac(D_p^-)}{Std} \sqrt{\frac{|D_p^+||D_p^-|}{|D_p^+ \cup D_p^-|}}$$

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

- $ac(D_p^+) - ac(D_p^-)$, indicates the difference in terms of average cost for positive and negative sequences.
- Std , standard deviation of the cost to avoid absolute values.
- $\sqrt{\frac{|D_p^+||D_p^-|}{|D_p^+ \cup D_p^-|}}$, measures distribution difference to indicate patterns' effect on the outcome.
- The cor measure values are in the $[-1,1]$ interval.
- The greater positive(negative) the cor measure is, the more a pattern is correlated with a positive (negative) utility.

How to reduce the search space? (1)

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)>	...

We propose a **lower-bound** on the **average cost**:

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

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

e.g. For *minsup* = 2

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

$$ASC(bc) = (3+5) / 2 = 2.67$$

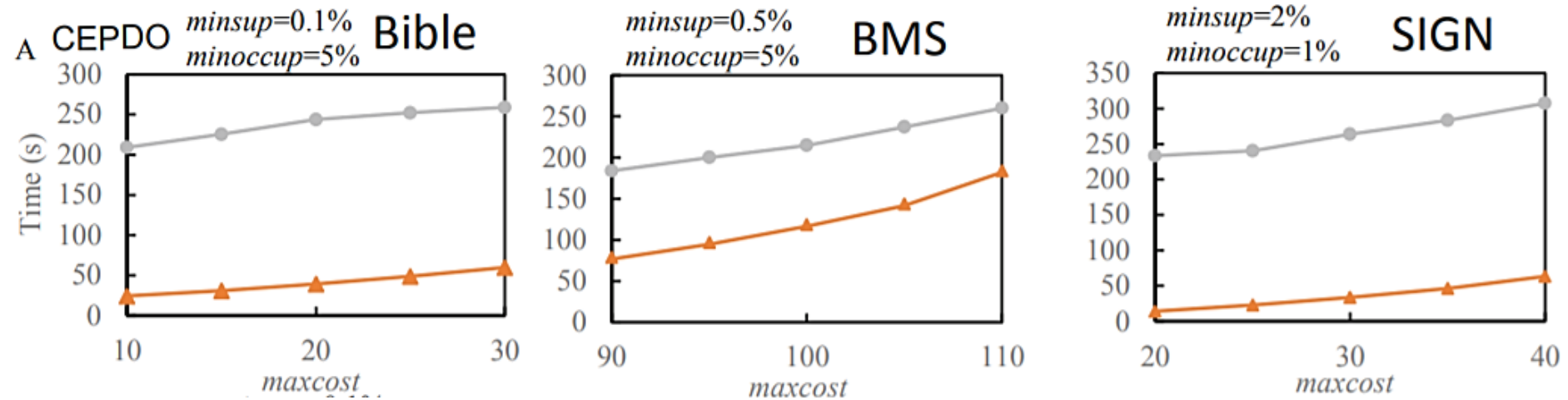
Properties of ASC

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

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 supersequences.

Execution times



gray line: no pruning strategy / orange line: using *AMSC*

results: up to 10 times faster

BMS, Bible and SIGN are benchmark datasets

Case Study: data review

- Six session, each session contains 115 students' study records.
- In each session, 15 activities and 13 features in system are documented.

e.g

1	1, 1, Es, Other, 2.10.2014 11:25:33, 2.10.2014 11:25:34, 0, 0, 0, 0, 0, 84, 0
2	1, 1, Es, Aulaweb, 2.10.2014 11:25:35, 2.10.2014 11:25:42, 218, 0, 0, 4, 0, 397, 0
3	1, 1, Es, Blank, 2.10.2014 11:25:43, 2.10.2014 11:25:43, 0, 0, 0, 0, 0, 59, 0
4	1, 1, Es, Deeds, 2.10.2014 11:25:44, 2.10.2014 11:26:17, 154117, 6, 0, 8, 0, 1581, 4
5	1, 1, Es, Other, 2.10.2014 11:26:18, 2.10.2014 11:26:18, 0, 0, 0, 2, 0, 103, 0
6	1, 1, Es, Other, 2.10.2014 11:26:19, 2.10.2014 11:26:27, 460, 0, 0, 4, 0, 424, 8
7	1, 1, Es, Blank, 2.10.2014 11:26:28, 2.10.2014 11:26:28, 0, 0, 0, 1, 0, 93, 0
8	1, 1, Es, Deeds, 2.10.2014 11:26:29, 2.10.2014 11:26:29, 0, 0, 0, 1, 0, 75, 0
9	1, 1, Es, Aulaweb, 2.10.2014 11:26:30, 2.10.2014 11:26:33, 0, 0, 0, 2, 0, 238, 0
10	1, 1, Es, Deeds, 2.10.2014 11:26:34, 2.10.2014 11:26:41, 4933, 0, 0, 2, 0, 268, 0
11	1, 1, Es, Other, 2.10.2014 11:26:42, 2.10.2014 11:26:47, 3212, 0, 0, 4, 0, 275, 2
12	1, 1, Es, Aulaweb, 2.10.2014 11:26:48, 2.10.2014 11:27:0, 1174, 3, 0, 2, 0, 596, 0
13	1, 1, Es, Aulaweb, 2.10.2014 11:27:1, 2.10.2014 11:27:4, 63, 0, 0, 2, 0, 297, 0
14	1, 1, Es, Other, 2.10.2014 11:27:5, 2.10.2014 11:27:10, 7834, 0, 0, 0, 0, 142, 0
15	1, 1, Es, Aulaweb, 2.10.2014 11:27:11, 2.10.2014 11:27:15, 0, 4, 0, 2, 0, 342, 0
16	1, 1, Es, Other, 2.10.2014 11:27:16, 2.10.2014 11:27:17, 140, 0, 0, 0, 0, 99, 0
17	1, 1, Es_1_1, Study_Es_1_1, 2.10.2014 11:27:18, 2.10.2014 11:27:45, 165188, 0, 0, 4, 0, 715, 0
18	1, 1, Es_1_1, Deeds_Es_1_1, 2.10.2014 11:27:46, 2.10.2014 11:27:49, 234, 0, 0, 2, 0, 214, 0
19	1, 1, Es_1_1, Study_Es_1_1, 2.10.2014 11:27:50, 2.10.2014 11:27:50, 0, 0, 0, 0, 0, 77, 0
20	1, 1, Es_1_1, Deeds_Es_1_1, 2.10.2014 11:27:51, 2.10.2014 11:33:57, 11510470, 0, 0, 230, 54, 16970, 7
21	1, 1, Es_1_1, Properties, 2.10.2014 11:33:58, 2.10.2014 11:33:59, 31, 0, 0, 0, 0, 189, 0
22	1, 1, Es_1_1, Deeds_Es_1_1, 2.10.2014 11:34:0, 2.10.2014 11:36:3, 2396565, 5, 0, 50, 0, 2811, 0
23	1, 1, Es_1_1, Other, 2.10.2014 11:36:4, 2.10.2014 11:36:12, 11543, 0, 0, 2, 0, 356, 0
24	1, 1, Es_1_1, Deeds_Es_1_1, 2.10.2014 11:36:13, 2.10.2014 11:36:55, 819129, 0, 0, 2, 0, 116, 0

Case study 1: binary e-learning DB

Database

- 115 students
- A **sequence** is a series of learning sessions, e_1 to e_6 .
- **Cost**: time to complete a session.
- **Utility**: to *pass* or *fail* the final exam.

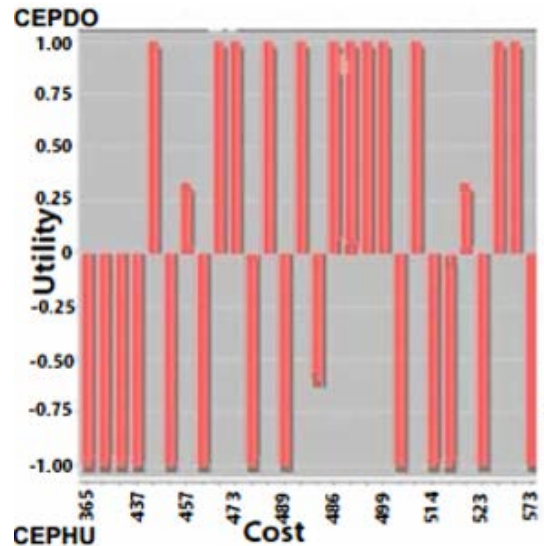
Observation: now the point threshold for passing the exam is set to be 60, when the threshold **decrease**, the **correlation** values **increase**.

Cost-efficient patterns

Pattern	Correlation	Average Cost	Support
$\langle e_1, e_6 \rangle$	0.210	250.2	39
$\langle e_1, e_2, e_5, e_6 \rangle$	0.209	485.7	34
$\langle e_2, e_6 \rangle$	0.208	298.4	41
$\langle e_1, e_2, e_6 \rangle$	0.204	391.9	36
$\langle e_1, e_5, e_6 \rangle$	0.194	344.3	37
$\langle e_6 \rangle$	0.193	157.2	50
$\langle e_1, e_4 \rangle$	-0.004	169.1	41
$\langle e_1, e_5 \rangle$	0.002	186.0	41
$\langle e_2, e_3 \rangle$	0.001	284.1	40
$\langle e_3, e_4, e_5, e_6 \rangle$	0.001	469.5	40
$\langle e_1, e_4, e_5 \rangle$	0.003	263.2	38
$\langle e_1, e_2, e_4 \rangle$	-0.003	311.5	36
$\langle e_2, e_3, e_4 \rangle$	-0.005	358.2	38
$\langle e_5 \rangle$	-0.147	96.3	53
$\langle e_4, e_5 \rangle$	-0.109	171.0	49
$\langle e_1, e_3 \rangle$	-0.099	234.6	37
$\langle e_1, e_3, e_4 \rangle$	-0.081	311.2	35

Visualization and Interpretability

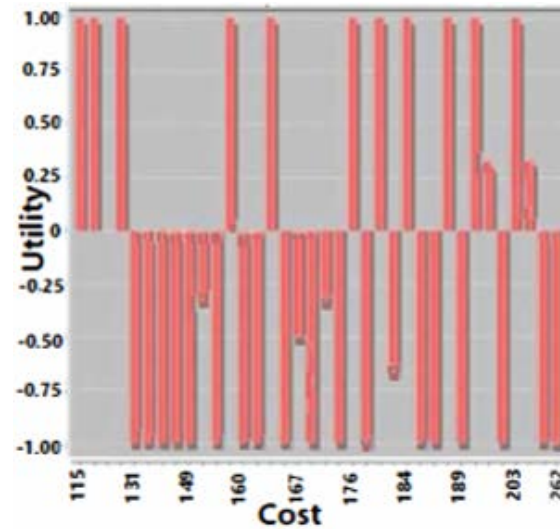
$\langle e_1, e_6 \rangle$



$$\text{cor}(\langle e_1, e_6 \rangle) = 0.210$$

positive
correlation

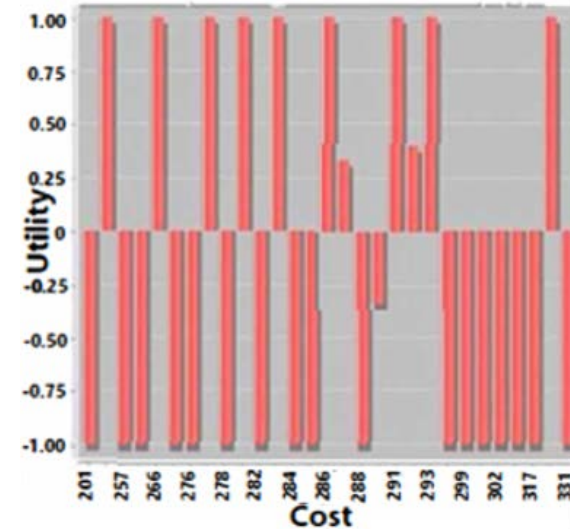
$\langle e_4, e_5 \rangle$



$$\text{cor}(\langle e_4, e_5 \rangle) = -0.109$$

negative correlation

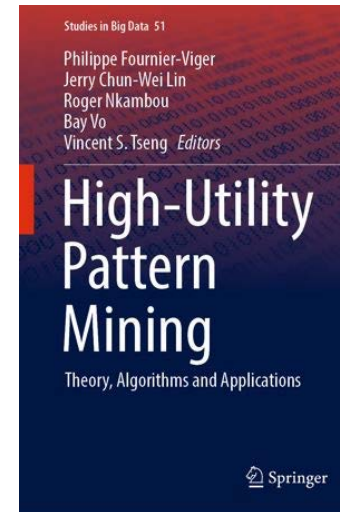
$\langle e_2, e_3 \rangle$



$$\text{cor}(\langle e_2, e_3 \rangle) = 0.001$$

no correlation

The difference distribution in terms of cost shows the measure's rationality.



Open source Java data mining software, 150 algorithms

<http://www.phillippe-fournier-viger.com/spmf/>