

Preview

Sequential Activity Database

- A **sequence** is a series of activities, each having a cost.
- The **utility** of a sequence is a **binary class**.
- Measure the **correlation** of a **pattern** p with the desirable outcome.

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)

What's new in this work?

- New problem was defined.
- A tighter bound to reduce the search space.
- A new measure to evaluate patterns.
- Two pruning strategies are designed.

Sequential Activity Database

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

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)

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. $\text{sup}(<\mathbf{ab}>) = |\{S_1, S_4\}| = 2$ 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 :

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

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

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

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

The *occupancy* 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 **occupancy** of a pattern p :

$$occup(p) = \frac{1}{sup(p)} \sum_{p \subseteq S_s \in SEL} \frac{|p|}{|S_s|}$$

$$occup(ab) = \frac{1}{2} \cdot \left(\underbrace{\frac{2}{5}}_{\text{Sequence 1}} + \underbrace{\frac{2}{4}}_{\text{Sequence 4}} \right) = \mathbf{0.45}$$

According to the previous experiments, this measure is used to remove patterns that are **short** and **non-representative** of the containing sequences.

Problem 2:

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

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

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

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

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

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

Problem 2:

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

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

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 **trade-off** between the **cost** and **utility** of a pattern p :

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

Average cost

Utility

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

Trade-off values are in the $(0, \infty]$ interval. Lower means more efficient.

More details...

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 :

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

$$u(\mathbf{ab}) = \underbrace{40}_{\text{Sequence 1}} + \underbrace{70}_{\text{Sequence 3}} / 2 = 55$$

Trade-off of a pattern p :

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

$$tf(\mathbf{ab}) = 5 / 55 = 0.09$$
$$tf(\mathbf{cd}) = 6.7 / 50 = 0.13$$

Thus, pattern (\mathbf{ab}) is more efficient than (\mathbf{cd}).

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 **tighter lower-bound** on the **average cost**:

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

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

e.g. For *minsup* = 2

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

$$AMSC(bc) = (3+5) / 2 = 4$$

How to reduce the search space? (2)

We use an upper bound on the occupancy of a pattern p :

$$uo(p) = \frac{1}{sup(p)} \cdot \max_{S_1, \dots, S_{sup(p)}} \sum_{i=1}^{sup(p)} \frac{psl[S_i] + ssl[S_i]}{sl[S_i]}$$

where $psl[S_i]$, $ssl[S_i]$ and $sl[S_i]$ is p 's length in S_i , the length of the subsequence after p in S_i , and S_i 's length, respectively.

e.g. $minsup = 2, p = \langle a, b, c \rangle$

$psl[S_1]=psl, [S_4]=3, ssl[S_1]=1, ssl[S_4]=1,$

$sl[S_1]=5, sl[S_4]=4,$

$$uo(p) = \frac{1}{2} \left(\frac{3+1}{5} + \frac{3+1}{4} \right) = 0.9$$

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 reduce the search space? (3)

We use an upper bound on the utility of a pattern p in a numeric DB:

$$upperu = \frac{1}{minsup} \sum_{i=1,2,\dots,n} u(p, S_i)$$

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. $minsup = 2$ $p = \langle a, b, c \rangle$

$u(p, S_1) = 40$ $u(p, S_4) = 70$

$upperu(p) = \frac{1}{2} (40 + 70) = 55$

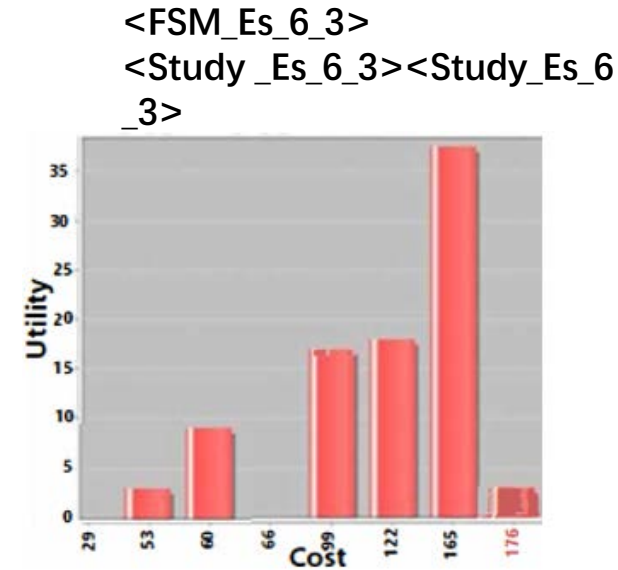
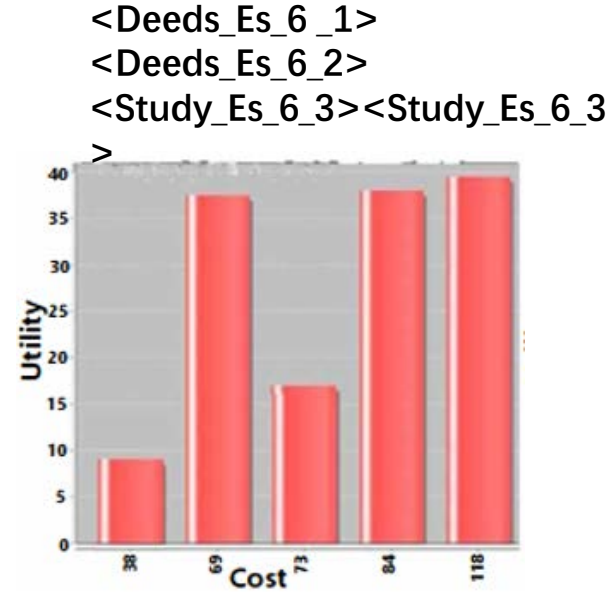
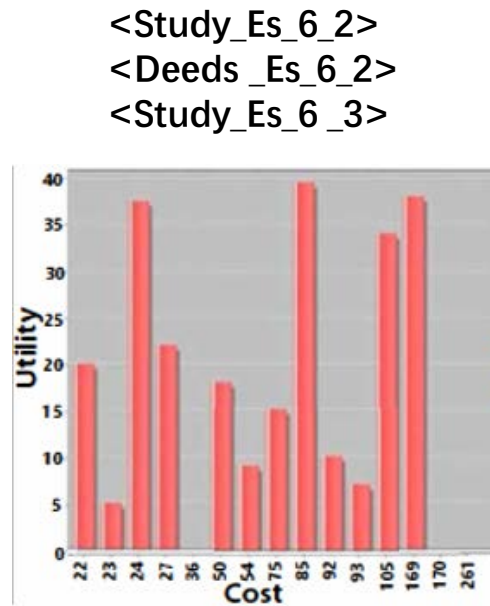
Case study 2: numeric e-learning DB

Cost-effective patterns found in learning session 6

Database

- A sequence is the learning activities of a session.
- **Cost**: time to complete an activity.
- **Utility**: the score at the final exam.

Utility	Pattern	trade-off	Average Cost	Support
1	$\langle Study_Es_6.1, Study_Es_6.1, Study_Es_6.1 \rangle$	48.0	57.6	5
2	$\langle Study_Es_6.1, Study_Es_6.1, Study_Es_6.3 \rangle$	15.0	33.0	5
4	$\langle Study_Es_6.1, Study_Es_6.2, Study_Es_6.2 \rangle$	7.0	32.8	6
5	$\langle Study_Es_6.1, Study_Es_6.1 \rangle$	5.1	27.6	9
6	$\langle Study_Es_6.1, Study_Es_6.1, Deeds_Es_6.1 \rangle$	6.0	40.5	6
7	$\langle Study_Es_6.2, Study_Es_6.2 \rangle$	2.9	20.7	11
8	$\langle Study_Es_6.2, Study_Es_6.2, Deeds_Es_6.2 \rangle$	3.6	31.3	6
9	$\langle Study_Es_6.1 \rangle$	1.2	11.0	20
10	$\langle Study_Es_6.1, Deeds_Es_6.2 \rangle$	2.1	21	13
11	$\langle Study_Es_6.2, Study_Es_6.3 \rangle$	1.56	18.2	16
12	$\langle Study_Es_6.2 \rangle$	0.69	8.9	25
13	$\langle Study_Es_6.3 \rangle$	0.64	8.52	25
14	$\langle Deeds_Es_6.2 \rangle$	0.62	9.1	28
15	$\langle Study_Es_6.2, Deeds_Es_6.2, Study_Es_6.3 \rangle$	1.7	27.0	10
16	$\langle FSM_Es_6.1, FSM_Es_6.1, Deeds_Es_6.2, Study_Es_6.3 \rangle$	3.9	64.2	5
17	$\langle Deeds_Es_6.2, Study_Es_6.3 \rangle$	0.89	15.6	16
18	$\langle Study_Es_6.3, Study_Es_6.3 \rangle$	1.0	18.8	9
20	$\langle Deeds_Es_6.1, Study_Es_6.3, Study_Es_6.3 \rangle$	1.6	32.7	7
21	$\langle FSM_Es_6.3, Study_Es_6.3, Study_Es_6.3 \rangle$	4.5	94.8	6
23	$\langle Deeds_Es_6.2, Study_Es_6.3, Study_Es_6.3 \rangle$	1.2	27.0	6
24	$\langle FSM_Es_6.1, Deeds_Es_6.1, Study_Es_6.3, Study_Es_6.3 \rangle$	3.6	86.3	6
28	$\langle Deeds_Es_6.1, Deeds_Es_6.2, Study_Es_6.3, Study_Es_6.3 \rangle$	1.35	38.0	5



$tr(<Study_Es_6_2> <Deeds_Es_6_2> <Study_Es_6_3>) = 1.74,$
cost / utility = 27 / 15

$tr(<Deeds_Es_6_1> <Deeds_Es_6_2> <Study_Es_6_3> <Study_Es_6_3>) = 1.35,$
cost / utility = 28 / 21

$tr(<FSM_Es_6_3> <Study_Es_6_3> <Study_Es_6_3>) = 4.5,$
cost / utility = 94.8 / 21

The smaller the trade-off, the more effective the pattern.