Додатне технике правила придруживања

Ненад Митић

Математички факултет
nenad@matf.bg.ac.rs

Непрекидни и категорички атрибути

- Претходно приказани алгоритми су применљиви на податке представљене у облику асиметричних бинарних атрибута
- Шта ако су подаци представљени категоричким, непрекидним или симетричним бинарним атрибутима
- Потребне су модификације алгоритама

Индекс	Име	Презиме	Пол	Назив програма	Назив предмета	Статус	Датум полагања	Оцена
20150400	Ања	Јевтић	ж	Информатика	Алгебра 1	о	2016-02-19	10
20150220	Урош	Рачић	м	Математика	Геометрија 2	о	2016-02-03	10
20150220	Урош	Рачић	м	Математика	Програмирање 1	о	2016-02-02	10
20150220	Урош	Рачић	м	Математика	Страни језик	о	2016-01-30	10
...	...	...	.	...	...	.	...	...

Непрекидни и категорички атрибути

- Поступак са категоричким атрибутима
- Трансформисати категоричке атрибуте у асиметричне бинарне променљиве
- Могући проблеми
  - Шта ако атрибут има велики број могућих вредности (нпр. атрибут држава има више од 200 могућих вредности)
  - Велики број вредности атрибута може да има малу подршку
  - Могуће решење: агрегирати вредности таквих атрибута

Непрекидни и категорички атрибути

- Шта ако је расподела вредности атрибута јако одскаче на једну страну?
- Нпр. 95% студената има за оцену 10 вредност предмета "Страни језик"
- Могуће решење: избрисати ставке са високом учесталошћу
- Поступак са непрекидним атрибутима?
- Различите врсте правила

  [STATUS=o]+[OCENA=6]+[OCENA=7]+[DATPOLAGANJA >= 736950 AND < 737450]

  ⟶ [NAZIV_PREDMETA=Програмирање 2]

Поступак са непрекидним атрибутима

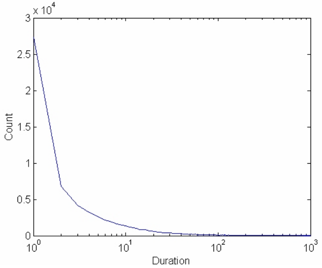
Различите методе

- Засноване на дискретизацији
- Статистички засноване методе
- Засноване на не-дискретизацији
  - *minApriori*

Методе засноване на дискретизацији

Дискретизација

- Без надзора
  - Делови исте ширине
  - Делови исте дубине
  - Кластеровање



- Под надзором

На основу вредности атрибута A

Класа	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9
Аномалија	0	0	20	10	20	0	0	0	0
Нормалних	150	100	0	0	0	100	100	150	100

deo1 deo2 deo3

Проблеми код дискретизације

- Величина дискретизационих интервала утице на подршку и поузданост
- Ако је интервал јако мали подршка може да буде недовољна
- Ако је интервал јако велики поузданост може да буде недовољна
- Решење: користити све могуће интервале (ефикасност?)

Проблеми код дискретизације

- Време извршавања
  - Ако интервал садржи n вредности, у просеку постоји $O(n^2)$ могућих уређења
- Велики број правила
- Илустрација проблема: правила придруживања над табелом *rezultati_ispita*

Статистички засноване методе

- Primer: [status=o]+[datpolaganja >= 736950 and < 737450] $\longrightarrow$ [ocena]:$\mu = 7.5$
- Последично правило - непрекидна променљиве карактерисана статистиком (средина, медијана, stddev, ...)
- Приступ
  - Издвајање циљног атрибута из остатка података
  - Применити постојеће генерисање честог скупа ставки на остатак података
  - За сваку честу ставку израчунати описну статистику за одговарајућу циљну променљиву
    - Чест скуп ставки постаје правило за укључивање циљног атрибута као последичног правила
  - Применити статистичке тестове ради одређивања интересантности правила

Статистички засноване методе

- Како одредити да ли је правило придруживања интересантно?
- Поредити статистику дела популације покривену правилом у односу на део популације који није покривен правилом: $A \longrightarrow B : \mu$ prema $\overline{A} \longrightarrow B : \mu'$
- Статистичко тестирање хипотеза:
    - Нулта хипотеза: $H0 : \mu' = \mu + \Delta$
    - Алтернативна хипотеза: $H1 : \mu' > \mu + \Delta$

Статистички засноване методе

Да би се одредило која је хипотеза важећа рачуна се Z статистика

$$Z = \frac{\mu' - \mu - \Delta}{\sqrt{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})}}$$

где је

- n_1 број трансакција које подржавају A, n_2 број трансакција које не подржавају A
- s_1 стандардна девијација циљног атрибута у трансакцијама које подржавају A
- s_2 стандардна девијација циљног атрибута у трансакцијама које не подржавају A

Према нултој хипотези Z има средину 0 и варијансу 1. Ако је $Z > Z_0$ где је Z_0 критична вредност за одређен ниво поузданости, тада се нулта хипотеза одбацује

Статистички засноване методе

Пример: нека је правило

[Назив_предмета='Анализа 1']+[ознакарока='Јун'] $\longrightarrow$ [поени]:$\mu = 48$

подржано од стране 50 студената, и нека је стандардна девијација броја њихових поена 3.5. Са друге стране, за комплементаран скуп од 200 студената који не подржавају ово правило средња вредност броја поена је 55, а стандардна девијација је 6.5. Нека је правило интересантно ако је разлика између μ' и μ већа од 5 поена ($\Delta = 5$). Вредност Z је

$$Z = \frac{55 - 48 - 5}{\sqrt{(\frac{3.5^2}{50} + \frac{6.5^2}{200})}} = \frac{2}{\sqrt{(0.245 + 0.21125}} = \frac{2}{\sqrt{(0.45625}} = \frac{2}{0,675462804} = 2,960932842$$

За 1-страни тест са поузданошћу од 95% критична вредност за одбацивање нулте хипотезе 1.64. Како је $Z > 1.64$ нулта хипотеза се одбацује ===> правило је интересантно

Методе засноване на не-дискретизацији

- Постоје случајеви када је интереснатније наћи везе између непрекидних атрибута него између њихових дискретних интервала
- Пример: на основу табеле појављивања речи у тексту може се закључити да W_1 и W_2 имају тенденцију да се појављују заједно у истом документу

	W_1	W_2	W_3	W_4	W_5
D_1	2	2	0	0	1
D_2	0	0	1	2	2
D_3	2	3	0	0	0
D_4	0	0	1	0	1
D_5	1	1	1	0	2

Методе засноване на не-дискретизацији

- Подаци садрже само непрекидне атрибуте истог "типа"
- Пример: фреквенција речи у неком документу

	W_1	W_2	W_3	W_4	W_5
D_1	2	2	0	0	1
D_2	0	0	1	2	2
D_3	2	3	0	0	0
D_4	0	0	1	0	1
D_5	1	1	1	0	2

- Могуће решење:
  - Конвертовати садржај у 0/1 матрицу где је 1 нормализована вредност која прелази одређен праг и применити постојеће алгоритме (губи се информација о фреквенцији речи)
  - Дискретизација је често непримењива пошто корисници желе везе између речи а не између броја појављивања речи

Min-Apriori

- Како одредити подршку за реч?
  - Ако се саберу фреквенције подршка ће бити већа од укупног броја докумената!
  - Нормализује се вектор речи, нпр. употребом L_1 норме
  - Свака реч има подршку једнаку 1.0

	W_1	W_2	W_3	W_4	W_5			W_1	W_2	W_3	W_4	W_5
D_1	2	2	0	0	1	$\longrightarrow$	D_1	0.40	0.33	0.00	0.00	0.17
D_2	0	0	1	2	2	$\longrightarrow$	D_2	0.00	0.00	0.33	1.00	0.33
D_3	2	3	0	0	0	$\longrightarrow$	D_3	0.40	0.50	0.00	0.00	0.00
D_4	0	0	1	0	1	$\longrightarrow$	D_4	0.00	0.00	0.33	0.00	0.17
D_5	1	1	1	0	2	$\longrightarrow$	D_5	0.20	0.17	0.33	0.00	0.33

Min-Apriori

- Ставка је скуп речи
- Подршка представља меру колико су речи придружене једна другој
- Подршка скупа речи C у скупу докумената T се рачуна као

$$sup(C) = \sum_{i \in T} \min_{j \in C} D(i,j)$$

Пример: $sup(W_1, W_2, W_3) = 0 + 0 + 0 + 0 + 0.17 = 0.17$

	W_1	W_2	W_3	W_4	W_5
D_1	0.40	0.33	0.00	0.00	0.17
D_2	0.00	0.00	0.33	1.00	0.33
D_3	0.40	0.50	0.00	0.00	0.00
D_4	0.00	0.00	0.33	0.00	0.17
D_5	0.20	0.17	0.33	0.00	0.33

Min-Apriori

Ова мера подршке се назива *Min-Apriori* и има особине да подршка

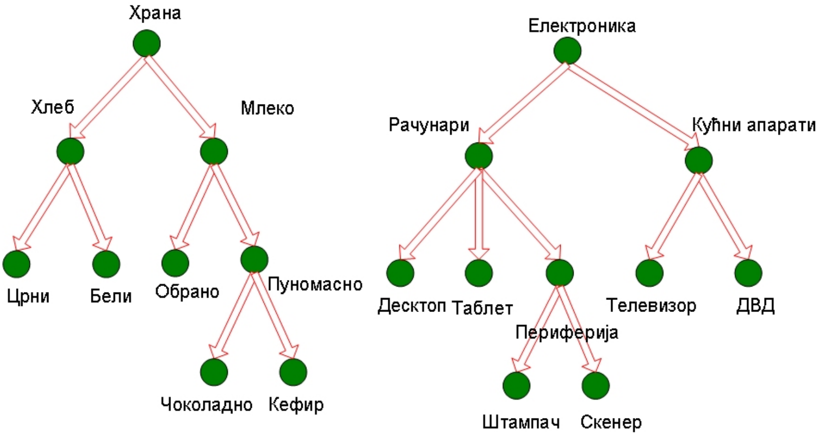
- монтоно расте како расте нормализована фреквенција речи
- монтоно расте како расте број документата који садрже реч
- монотоно опада како расте број речи у скупу ставки - анти-монотоност

Пример

- $sup(W_1) = 0.4 + 0 + 0.4 + 0 + 0.2 = 1.0$
- $sup(W_1, W_2) = 0.33 + 0 + 0.4 + 0 + 0.17 = 0.9$
- $sup(W_1, W_2, W_3) = 0 + 0 + 0 + 0 + 0.17 = 0.17$

	W_1	W_2	W_3	W_4	W_5
D_1	0.40	0.33	0.00	0.00	0.17
D_2	0.00	0.00	0.33	1.00	0.33
D_3	0.40	0.50	0.00	0.00	0.00
D_4	0.00	0.00	0.33	0.00	0.17
D_5	0.20	0.17	0.33	0.00	0.33

Правила придруживања између више нивоа



Правила придруживања између више нивоа

- Зашто се укључује хијерархија концепата?
- Правила на нижим нивоима можда немају довољну подршку да се јаве у честим скуповима података
- Правила на нижим нивоима су превише специфична. Нпр. обрано млеко $\longrightarrow$ бели хлеб, пуномасно млеко $\longrightarrow$ црни хлеб, кефир $\longrightarrow$ цри хлеб, ... су индикатори правила придруживања између млека и хлеба
- Ако се обилази дрво хијерархије концепата тада важи
  1. Ако је X родитељ ставка за X_1 и X_2 тада важи $\sigma(X) \leq \sigma(X_1) + \sigma(X_2)$
  2. Ако је $\sigma(X_1 \cup Y_1) \geq \text{minsup}$ и X је родитељ од X_1 и Y је родитељ од Y_1 тада $\sigma(X \cup Y_1) \geq \text{minsup}$, $\sigma(X_1 \cup Y) \geq \text{minsup}$, и $\sigma(X \cup Y) \geq \text{minsup}$
  3. Ако $\text{conf}(X_1 \longrightarrow Y_1) \geq \text{minconf}$ тада $\text{conf}(X_1 \longrightarrow Y) \geq \text{minconf}$

Правила придруживања између више нивоа

Приступ 1:

- Проширити текуће правило придруживања проширивањем сваке трансакције са ставком са вишег нивоа
  - Оригинална трансакције: {обрано млеко, бели хлеб}
  - Проширена трансакција: {обрано млеко, бели хлеб, млеко, хлеб, храна}
- Проблеми:
  - Ставке које се налазе на вишем нивоу имају много виши ниво подршке
    - ако је подршка јако мала велики број честих образаца укључује ставке са вишег нивоа
  - Повећава се димензионалност података

Правила придруживања између више нивоа

Приступ 2:

- Формирати честе обрасце прво на највишем нивоу
- Затим формирати цесте обрасце на наредним највишим нивоима, итд.

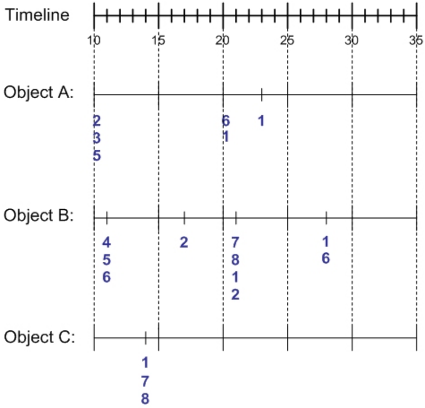
Проблеми:

- У/И захтеви се драматично повећавају због потребе вишеструког пролажења кроз податке
- Могу да се изгубе неке од потенцијално занимљивих веза између образаца на различитим нивоима

Низ података

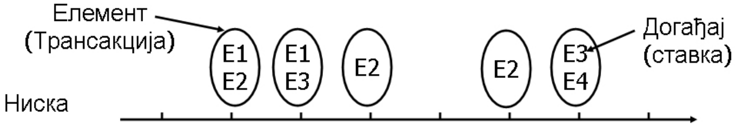
Низ у бази:

Object	Timestamp	Events
A	10	2, 3, 5
A	20	6, 1
A	23	1
B	11	4, 5, 6
B	17	2
B	21	7, 8, 1, 2
B	28	1, 6
C	14	1, 8, 7



Примери низа података

Niz u bazi	Niz podataka	Element (Transakcija)	Događaj (stavka)
Kupac	Istorija kupovanja datog kupca	Skup stavki kupljen od strane kupca u trenutku t	Knjige, tekući proizvodi, CDovi, itd.
Veb podaci	Pregledanje aktivnosti pojedinačnog posetioca Veba	Skup datoteka pregledan od strane posecioca Veba posle pritiska na tipku miša	Glavna strana, strana sa indeksima, kontakt informacije, itd.
Podaci o događajima	Istorija događaja formirana pomoću datog senzora	Događaju uočeni senzorom u trenutku t	Tipovi alarma koje je formirao senzor
Niske genoma	DNK niske pojedinačnih vrsta	Elementi DNK niski	Baze A,T,G,C



Формална дефиниција ниске

- Ниска је уређена листа елемената (трансакција) $S =< e_1 e_2 e_3 \ldots >$
- Сваки елемент садржи скуп догађаја (ставки) $e_i = \{i_1 i_2 \ldots i_k\}$
- Сваком елементу се додељује одређено време или место
- Број елемената у нисци s одређује дужину ниске $\mid s \mid$
- к-ниска је ниска која садржи к догађаја (ставки)
- Пример ниске: $<$ Анализа1, Анализа 2, Анализа 3$>$

Формална дефиниција подниске

Дефиниција: Ниска $< a_1 a_2 \ldots a_n >$ је садржана у нисци $< b_1 b_2 \ldots b_m >$ $(m \geq n)$ ако постоје цели бројеви $i_1 < i_2 < \ldots < i_n$ такви да важи $a_1 \subseteq b_{i_1}, a_2 \subseteq b_{i_2}, \ldots, a_n \subseteq b_{i_n}$

Ниска података	Подниска	Садржи
<{2,4} {3,5,6} {8} >	<{2} {3,5}>	Да
<{1,2} {3,4}>	<{1} {2}>	Не
<{2,4} {2,4} {2,5} >	<{2} {4}>	Да
<{2,4} {2,5} {4,5}>	<{2} {4} {5}>	Не
<{2,4} {2,5} {4,5}>	<{2} {5} {5}>	Да
<{2,4} {2,5} {4,5}>	<{2,4,5}>	Не

Дефиниција истраживања секвенцијалних образаца

- Задато
  - база са нискама
  - кориснички дефинисана најмања подршка minsup
- Подршка ниске w је количник броја ниски података које садрже w у односу на укупан број ниски
- Секвенцијални образац је честа подниска (т.ј. подниска чија је подршка $\geq$ minsup)
- Циљ: Наћи све подниске које имају подршку $\geq$ minsup
- Из дате ниске дужине n може да се изведе $\binom{n}{k}$ k-подниски

Истраживање секвенцијалних образаца - пример

Object	Timestamp	Events
A	1	1,2,4
A	2	2,3
A	3	5
B	1	1,2
B	2	2,3,4
C	1	1, 2
C	2	2,3,4
C	3	2,4,5
D	1	2
D	2	3, 4
D	3	4, 5
E	1	1, 3
E	2	2, 4, 5

Minsup = 50%

Primeri čestih podniski:

< {1,2} >	s=60%
< {2,3} >	s=60%
< {2,4}>	s=80%
< {3} {5}>	s=80%
< {1} {2} >	s=80%
< {2} {2} >	s=60%
< {1} {2,3} >	s=60%
< {2} {2,3} >	s=60%
< {1,2} {2,3} >	s=60%

Секвенцијални обрасци / потрошачка корпа

Секвенце

Купац	Датум	Ставке
A	10	2, 3, 5
A	20	1,6
A	23	1
B	11	4, 5, 6
B	17	2
B	21	1,2,7,8
B	28	1, 6
C	14	1,7,8

Подаци из потрошачке корпе

Догађај
2, 3, 5
1,6
1
4,5,6
2
1,2,7,8
1,6
1,7,8

{2} -> {1} $conf(\{2\} \rightarrow \{1\}) = \dfrac{\sigma(\{2\}\{1\})}{\sigma(\{2\})}$

(1,8) -> (7) $conf(1,8) \rightarrow (7)) = \dfrac{\sigma(1{,}7{,}8)}{\sigma(\{1{,}8\})}$

Издвајање секвенцијалних образаца

- Дато је н догађаја: $i_1, i_2, \dots i_n$
- Кандидатске 1-подниске:
  $< \{i_1\} >, < \{i_2\} >, < \{i_3\} >, \dots, < \{i_n\} >$
- Кандидатске 2-подниске:
  $< \{i_1, i_2\} >, < \{i_1, i_3\} >, ..., < \{i_1\}\{i_1\} >,$
  $< \{i_1\}\{i_2\} >, ..., < \{i_{n-1}\}\{i_n\} >$
- Кандидатске 3-подниске:
  $< \{i_1, i_2, i_3\} >, < \{i_1, i_2, i_4\} >, ..., < \{i_1, i_2\}\{i_1\} >, < \{i_1, i_2\}\{i_2\} >$
  $, ..., < \{i_1\}\{i_1, i_2\} >, < \{i_1\}\{i_1, i_3\} >,$
  $..., < \{i_1\}\{i_1\}\{i_1\} >, < \{i_1\}\{i_1\}\{i_2\} >, ...$

Издвајање секвенцијалних образаца

- Нека су дата два догађаја: a и b
- Кандидатске 1-подниске:
  $<\{a\}>, <\{b\}>$
- Кандидатске 2-подниске:
  $<\{a\}\{a\}>, <\{a\}\{b\}>, <\{b\}\{a\}>, <\{b\}\{b\}>, <\{a, b\}>$
- Кандидатске 3-подниске:
  $<\{a\}\{a\}\{a\}>, <\{a\}\{a\}\{b\}>, <\{a\}\{b\}\{a\}>, <\{a\}\{b\}\{b\}>,$
  $<\{b\}\{b\}\{b\}>, <\{b\}\{b\}\{a\}>, <\{b\}\{a\}\{b\}>, <\{b\}\{a\}\{a\}>$
  $<\{a, b\}\{a\}>, <\{a, b\}\{b\}>, <\{a\}\{a, b\}>, <\{b\}\{a, b\}>$

Формирање секвенцијалних образаца

Алгоритам

- Корак 1: Направити први пролаз кроз базу ниски D ради добијања свих 1-елемент честих подниски
- Корак 2: Понављати поступак све док има нових честих подниски
  - Формирање кандидата: спојити парове честих подниски нађених у (к-1)-овом пролазу ради формирања кандидатских ниски које садрже к догађаја
  - Поткресивање списка кандидата: поткресати скуп кандидатских к-ниски које садрже ретке (к-1) подниске
  - Израчунавање подршке: направити нови пролаз кроз базу ниски D ради налажења подршке за преостале кандидатске ниске
  - Уклањање кандидата: уклонити кандидатске к-ниске чија је подршка мања од minsup

Формирање кандидата

Основни случај (к=2)

- Спајањем две честе 1-ниске $<\{i_1\}>$ и $<\{i_2\}>$ се формирају две кандидатске 2-ниске: $<\{i_1\}\{i_2\}>$ и $<\{i_1 i_2\}>$

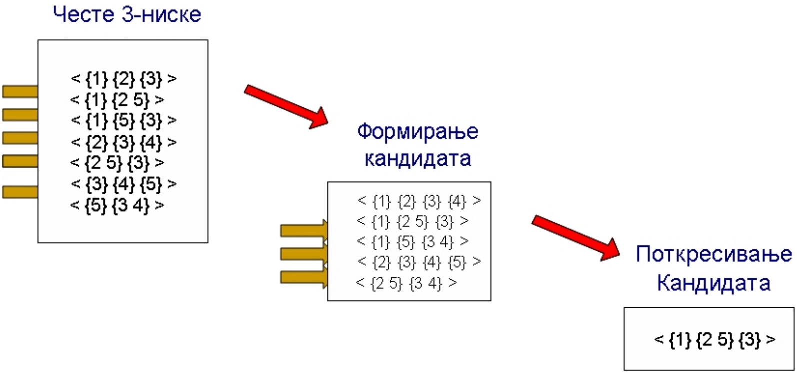
Општи случај (к>2)

- Честа (к-1)-ниска w_1 се спаја са другом честом (к-1)-ниском w_2 и формира кандидатска к-ниска ако је подниска добијена уклањањем првог догађаја из w_1 иста као и подниска добијена уклањањем последњег догађаја из w_2
  - Резултујућа кандидатска ниска је добијена проширењем ниске w_1 последњим догађајем из ниске w_2. Ако последња два догађаја из w_2 припадају истом елементу, тада последњи догађај из w_2 постаје део последњег елемента у w_1
  - У супротном, последњи догађај из w_2 постаје посебан елемент додат на крај w_1

Примери формирање кандидата

- Спајањем ниски $w_1 =< \{123\}\{46\} >$ и $w_2 =< \{23\}\{46\}\{5\} >$ се формира кандидатска ниска $< \{123\}\{46\}\{5\} >$ јер последњи елемент из w_2 (5) има само један догађај
- Спајањем ниски $w_1 =< \{1\}\{23\}\{4\} >$ и $w_2 =< \{23\}\{45\} >$ се формира кандидатска ниска $< \{1\}\{23\}\{45\} >$ јер последња два догађаја из w_2 (4 и 5) припадају истом елементу
- Спајањем ниски $w_1 =< \{1\}\{23\}\{4\} >$ и $w_2 =< \{23\}\{4\}\{5\} >$ се формира кандидатска ниска $< \{1\}\{23\}\{4\}\{5\} >$ јер последња два догађаја из w_2 (4 и 5) не припадају истом елементу
- Не могу да се споје ниске $w_1 =< \{1\}\{26\}\{4\} >$ и $w_2 =< \{1\}\{2\}\{45\} >$ да би се добила кандидатска ниска $< \{1\}\{26\}\{45\} >$ јер би, у случају да је цео поступак коректан, ниска добила спајањем ниске w_1 са ниском $< \{26\}\{45\} >$

Примери формирање кандидата



Налажење секвенцијалних образаца - алгоритам

Алгоритам за налажење секвенцијалних образаца - верзија слична Apriori

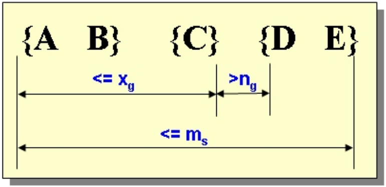
```
k=1
F_k = {i | i ∈ I ∧ σ({i})/N ≥ minsup}  {Naci sve ceste 1−podniske}
repeat
    k=k+1
    c_k = apriori_generisan(F_{k−1})  {Formirati kandidatske k−podniske}
    for svaka_sekvenca_podataka t ∈ T do
        C_t=podniska(C_k, t)  {Naci sve kandidate iz t}
        for svaka_kandidatska k−podniska c ∈ C_t do
            σ(c) = σ(c)+1  {Povecanje podrske}
        end for
    end for
    F_k = {c | c ∈ C_k ∧ σ({c})/N ≥ minsup}  {izdvajanje cestih k−podsniski}

until F_k = ∅
Rezultat= ⋃F_k
```

Временска ограничења

- Као један од услова да би ниска била честа може се поставити временско ограничење у коме се ниска појављује
- Ограничење може да укључи најмању и највећу вредност временског интервала између два појављивања ниске
- Интервал може да се односи на разлику између појављивања прве и последње ставке у комплетној секвенци или на најмању/највећу разлику између појављивања две ниске, или на временски прозор који представља разлику између појављивања прве/последње ставке у појединачној нисци

Временска ограничења



x_g: максимални јаз (max-gap)

n_g: минимални јаз (min-gap)

m_s: максимални размак

$x_g = 2$, $n_g = 0$, $m_s = 4$

Ниска података	Подниска	Садржи?
<{2,4} {3,5,6} {4,7} {4,5} {8} >	<{6} {5}>	Да
<{1} {2} {3} {4} {5} >	<{1} {4}>	Не
<{1} {2,3} {3,4} {4,5} >	<{2} {3} {5}>	Да
<{1,2} {3} {2,3} {3,4} {2,4} {4,5} >	<{1,2} {5}>	Не

Секвенцијални обрасци са временским ограничењима

- Приступ 1
  - Истраживати секвенцијалне обрасце без временских ограничења
  - Додатно обрадити откривене обрасце
- Приступ 2
  - Модификовати претходне алгоритме да директно поткресују кандидате који крше временска ограничења
  - Да ли још увек важи Априори принцип?

Априори принцип за низ података

Object	Timestamp	Events
A	1	1,2,4
A	2	2,3
A	3	5
B	1	1,2
B	2	2,3,4
C	1	1, 2
C	2	2,3,4
C	3	2,4,5
D	1	2
D	2	3, 4
D	3	4, 5
E	1	1, 3
E	2	2, 4, 5

Претпоставка:

x_g = 1 (max-gap)

n_g = 0 (min-gap)

m_s = 5 (maximum span)

minsup = 60%

<{2} {5}> подршка = 40%

или

<{2} {3} {5}> подршка = 60%

Проблем постоји због ограничења максималног јаза (*max-gap*)

Проблем се не јавља ако је максимални јаз бесконачан

Непрекидне поднискe

Важење Априори принципа се обезбеђује увођењем концепта *непрекидних подниски*

Ниска s је *непрекидна подниска* од $w = < e_1 \ e_2 \ \ldots \ e_k >$ ако важи

- s је добијено из w брисањем догађаја или из e_1 или из e_k, или
- s је добијено из w брисањем догађаја из неког елемента $e_i \in w$ који садржи најмање два догађаја, или
- s је непрекидна подниска од t и t је непрекидна подниска од w (рекурзивна дефиниција)

Ниска података s	Образац t	t непрекидна подниска s
<{1} {2,3} >	<{1} {2}>	Да
<{1,2} {2} {3} >	<{1} {2}>	Да
<{3,4} {1,2} {2,3} {4} >	<{1} {2}>	Да
<{1} {3} {2}>	<{1} {2}>	Не
<{1,2} {1} {3} {2} >	<{1} {2}>	Не

Модификовани Априори принцип

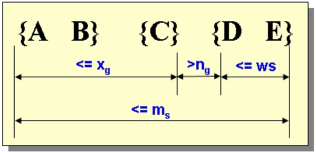
Дефиниција (модификовани Априори принцип): Ако је k-ниска честа тада су и све њене непрекидне $k-1$ ниске честе

Применом модификованог Априори принципа на истраживање секвенцијалних образаца разматрају се и поткресују кандидатске ниске

- Без ограничења на величину максималног јаза
  - Разматрају се све (к-1) подниске и кандидатска к-ниска се поткресује ако је најмање једна од њених (к-1)-подниски ретка
- Са ограничењем на величину максималног јаза
  - Разматрају се само непрекидне подниске и кандидатска к-ниска се поткресује ако је најмање једна непрекидна (к-1)-подниска ретка

Ограничења величине прозора

Додатно ограничење - величина прозора којим се дефинише највећи дозвољени временски размак између првог и последњег појављивања догађаја у елементима секвенцијалног обрасца. Прозор величине 0 означава да се сви догађаји у истом елементу дешавају истовремено



x_g: максимални јаз (max-gap)

n_g: минимални јаз (min-gap)

ws: величина прозора

m_s: максимални размак

$x_g = 2$, $n_g = 0$, ws = 1, $m_s = 5$

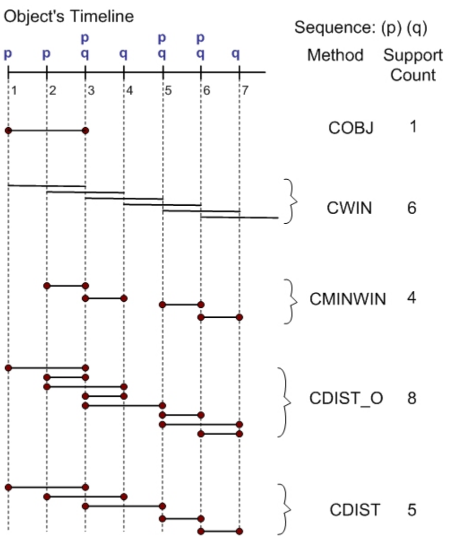
Ниска података	Образац	Садржи?
<{2,4} {3,5,6} {4,7} {4,6} {8} >	<{3,5}>	Не
<{2,4} {3,5,6} {4,7} {4,5} {8} >	<{3,5}>	Да
<{1} {2} {3} {4} {5}>	<{1,2} {3}>	Да
<{1} {2} {3} {4} {5}>	<{1,2} {3,4}>	Не
<{1,2} {2,3} {3,4} {4,5}>	<{1,2} {3,4}>	Да

Ограничења величине прозора

ws=2, mingap=0, maxgap=3, maxspan=∞

Ниска података	Образац	Садржи?
<{1,3} {3,4} {4} {5} {6,7} {8} >	<{3,4} {5}>	Да
<{1,3} {3,4} {4} {5} {6,7} {8} >	<{4,6} {8}>	Да
<{1,3} {3,4} {4} {5} {6,7} {8} >	<{3,4,6} {8}>	Не
<{1,3} {3,4} {4} {5} {6,7} {8} >	<{1,3,4} {6,7,8}>	Не

Алтернативне шеме пребројавања ставки



Assume:

x_g = 2 (max-gap)

n_g = 0 (min-gap)

ws = 0 (window size)

m_s = 2 (maximum span)

Ретки обрасци

Дефиниција (ретки обрасци): Редак образац је скуп ставки или правило чија је подршка мања од задатог прага *minsup*

Примери:

- Истовремена продаја DVD и VCR снимача је релативно ретка. Ове ставке су негативно корелисане и представљају конкурентске ставке
- Истовремено уписивање изборних предмета Анализа 4, Алгебра 2 и Дискретне структуре 3 на 4. години студија (И смера) је ретко. Ови предмети су међусобно конкурентни и њихова заједничка појава је негативно корелисана
- Ако је {Пожар = Да} честа ставка али {Пожар = Да, Аларм = Не} је ретка, ово правилио је важна ртка ставка јер означава грешку у алармном систему

Негативни обрасци

Нека је $I = \{i_1, i_2, \ldots, i_d\}$ скуп ставки. негативна ставка $\overline{i_k}$ означава одсуство ставке i_k из дате трансакције.

Дефиниција (негативан скуп ставки): Негативан скуп ставки X је скуп ставки који има следеће особине

- $X = A \cup \overline{B}$ где је A скуп позитивних ставки, $\overline{B}$ је скуп негативних ставки, $\mid \overline{B} \mid \geq 1$
- $sup(X) \geq minsup$

Дефиниција (негативно правило придруживања): Негативно правило придруживања је правило придруживања које има следеће особине

- правило је издвојено из негативног скупа ставки
- подршка правила је $\geq minsup$
- поузданост правила је $\geq minconf$

Негативно корелисани обрасци

Дефиниција (негативно корелисани обрасци): Скуп ставки $X = \{x_1, x_2, \ldots, x_k\}$ је негативно корелисан ако

$$sup(X) < \prod_{j=1}^{k} sup(x_j) = sup(x_1) \times sup(x_2) \times \ldots \times sup(x_k)$$

где $sup(x_j)$ означава подршку ставке x_j

Дефиниција (негативно корелисано правило придруживања, делимични услов): Правило придруживања $X \longrightarrow Y$ је негативно корелисано ако

$$sup(X \cup Y) < sup(X) \times sup(Y)$$

где су X и Y дисјунктни скупови ставки

Негативно корелисани обрасци

Дефиниција (негативно корелисано правило придруживања, пуни услов): Правило придруживања $X \longrightarrow Y$ је негативно корелисано ако

$$sup(X \cup Y) < \prod_j sup(x_j) \times \prod_j sup(y_j)$$

где су X и Y дисјунктни скупови ставки

Како су ставке у X и Y међусобно позитивно корелисане, практичније је користити делимичан уместо пуног услова негативне корелисаности правила придруживања. На пример, иако је правило

{наочаре, марамице за брисање стакала} $\longrightarrow$ {контактна сочива, течност за сочива}

негативно корелисано, ставке на левој и десној страни су међусобно корелисане и у случају примене пуног услова корелисаности ово правило се не би јавило

Негативно корелисани обрасци

Kako je

$$sup(X \cup Y) - sup(X) \times sup(Y)$$
$$= sup(X \cup Y) - \left[sup(X \cup Y) + sup(X \cup \overline{Y})\right] \times \left[sup(X \cup Y) + sup(\overline{X} \cup Y)\right]$$
$$= sup(X \cup Y) \times \left[1 - sup(X \cup Y) - sup(X \cup \overline{Y}) - sup(\overline{X} \cup Y)\right] - sup(X \cup \overline{Y}) \times sup(\overline{X} \cup Y)$$
$$= sup(X \cup Y) \times sup(\overline{X} \cup \overline{Y}) - sup(X \cup \overline{Y}) \times sup(\overline{X} \cup Y)$$

Услов негативне корелације може да се запише као

$$sup(X \cup Y) \times sup(\overline{X} \cup \overline{Y}) < sup(X \cup \overline{Y}) \times sup(\overline{X} \cup Y)$$

Поређење образаца

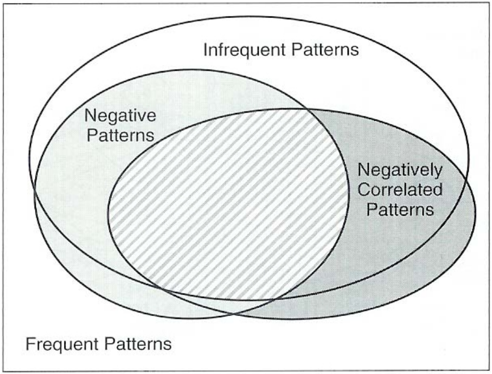
- Многи ретки обрасци имају одговарајуће негативне обрасце
- Многи негативно корелисани обрасци такође имају одговарајуће негативне обрасце
- Што ме мања подршка за $X \cup Y$ образац је више негативно корелисан
- Ретки негативно корелисани обрасци су интересантнији од честих негативно корелисаних образаца

Табела контингената за правило придруживања $X \longrightarrow Y$

	Y	$\overline{Y}$	
X	$sup(X \cup Y)$	$sup(X \cup \overline{Y})$	$sup(X)$
$\overline{X}$	$sup(\overline{X} \cup Y)$	$sup(\overline{X} \cup \overline{Y})$	$sup(\overline{X})$
	$sup(Y)$	$sup(\overline{Y})$	

Ако су X и Y негативно корелисани тада $X \cup \overline{Y}$, $\overline{X} \cup Y$ или оба морају да имају релативно високу подршку

Поређење


Infrequent Patterns
Negative Patterns
Negatively Correlated Patterns
Frequent Patterns

Технике за истраживање негативних образаца

Техника заснована на симетричним бинарним променљивим

TID	Items
1	{A,B}
2	{A,B,C}
3	{C}
4	{B,C}
5	{B,D}

Original Transactions

TID	A	$\overline{A}$	B	B	C	$\overline{C}$	D	D
1	1	0	1	0	0	1	0	1
2	1	0	1	0	1	0	0	1
3	0	1	0	1	1	0	0	1
4	0	1	1	0	1	0	0	1
5	0	1	1	0	0	1	1	0

Transactions with Negative Items

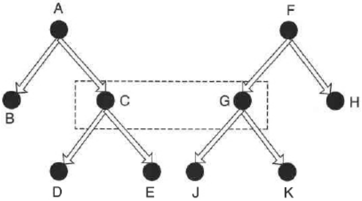
Проблеми

- Велики број ставки (уместо решетке са 2^d ставки добија се решетка са 2^{2d})
- Поткресивање на основу подршке не може да се примени (за сваку ставку или x или $\overline{x}$ имају подршку $\geq 50\%$
- Дужина трансакције се повећава када се укључе негативне ставке (на бар укупан броја ставки)

Технике засноване на очекиваној подршци

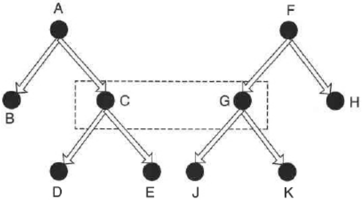
- Технике засноване на хијерархији концепата
- Технике засноване на индиректном придруживању

Технике засноване на хијерархији концепата



Ако је ставка {C,G} честа, а ставка {D,J} није, тада D и J формирају интересантан образац

Технике засноване на хијерархији концепата

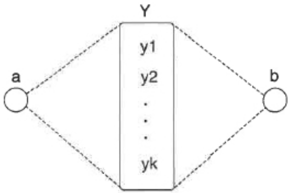


$$\varepsilon(s(E,J)) = s(C,G)x\frac{s(E)}{s(C)}x\frac{s(J)}{s(G)}$$

$$\varepsilon(s(C,J)) = s(C,G)x\frac{s(J)}{s(G)}$$

$$\varepsilon(s(C,H)) = s(C,G)x\frac{s(H)}{s(G)}$$

Технике засноване на индиректном придруживању



Пар ставки a и b је индиректно придружен преко медијатора ('комшија' који је довољно близу) Y ако важи

- подршка за пар $sup(\{a, b\}) < t_s$
- $\exists Y \neq \emptyset$ тако да
  - $s(\{a\} \cup Y) \geq t_f$ и $s(\{b\} \cup Y) \geq t_f$
  - $d(\{a\} \cup Y) \geq t_d$ $d(\{a\} \cup Y) \geq t_d$ где је $d(X, Z)$ мера придруживања X и Z

Класификација помоћу правила придруживања

Добијање правила за класификацију

- Скуп правила са великом поузданошћу
- На десном делу правила само једна ставка
- Десни део правила представља класу која се предвиђа
- Ставке на левом делу правила чине атрибуте на основу којих се врши предвиђање