Föreläsning 10 - Sekvensanalys

Josef Wilzen

2022-09-27

Outline

Behandla kategoriska attribut

- Betrakta en mängd av flygbiljetter som sålts av ett flygbolag.
  - Transittyp (inrikes, utrikes, ej transit)
  - Avgångspunkt (Linköping, Sundsvall)

Behandla kategoriska attribut

- Transformera attribut till binär form.

Transaction ID	Transfer = No transfer	Transfer = Domestic	Transfer = International	Origin = Linköping	Origin = Sundsvall
1	0	1	0	1	0
2	1	0	0	0	0
3	0	0	1	0	1

- Alt: Transformera till transaktionsform om det behövs.

Problem med kategoriska data

- Om en attribut har många attributvärden leder detta till mindre frekventa binära attribut som inte uppfyller supporttröskeln.
  - Minska ej tröskeln för då ökar antalet frekventa regler väsentligt.
  - Lösning: Gruppera attributvärden på ett logiskt sätt, t.ex. städer i olika län eller regioner.
  - Skapa en "Övrigt" kategori
- Om det finns attributvärden med väldigt hög support, t.ex. pasta i Italienska stormarknader.
  - Ta bort dessa binära variabler.
- Komplexiteten ökar exponentiellt med antalet attributvärden
  - Genererar många fler kandidater till frekventa enhetsmängder.

Behandla kontinuerliga attribut

- Diskretisering (vanligaste metoden)
  - Dela upp attributsvärden i intervall genom olika metoder:
    - Lika bredd
    - Lika frekvens
    - Kluster
  - Skapa ett attribut för varje kategoriskt värde .
  - Vi måste bestämma antal intervall
- Ex: Ålder = [0,5), [5,12), [12,25), [25,40), [40,+Inf)

Problem med diskretisering

Age group	Chat = Yes	Chat = No	Support for Yes (%)	Confidence for Yes (%)
[12, 16)	12	13	0.048	0.480
[16, 20)	11	2	0.044	0.846
[20, 24)	11	3	0.044	0.786
[24, 28)	12	13	0.048	0.480
[28, 32)	14	12	0.056	0.538
[32, 36)	15	12	0.060	0.556
[36, 40)	16	14	0.064	0.533
[40, 44)	16	14	0.064	0.533
[44, 48)	4	10	0.016	0.286
[48, 52)	5	11	0.020	0.313
[52, 56)	5	10	0.020	0.333
[56, 60)	4	11	0.016	0.267
Sum	125	125		

- Om intervallen är för breda försvinner regler p.g.a. för låg konfidens.
- Om intervallen är för smala försvinner regler p.g.a. för låg support.

Sekvensanalys

- Associationsanalys + tidvariabel = Sekvensanalys
- Vi vill hitta **sekventiella mönster**
- Transaktionsdatabaser brukar innehålla en attribut som motsvarar tidpunkt, dvs. händelser är tidsmarkerade

Customer	Day	Purchased items
A	10	bread, diapers, beer
A	20	beer, milk
A	23	milk
B	11	diapers, beer
B	17	bread
B	21	diapers, milk, bread
B	28	milk, beer
C	14	milk, diapers, beer

Object	Time-stamp	Events
A	10	2, 3, 4
A	20	4, 1
A	23	1
B	11	3, 4
B	17	2
B	21	3, 1, 2
B	28	1, 4
C	14	1, 3, 4

Sekventiella mönster

En sekvens är en ordnad lista av element:

$$S = \{e_1, e_2, \ldots, e_n\}$$

där varje element

$$e_j = \{i_1, i_2, \ldots, i_k\}$$

är en händelsemängd som förknippas med ett givet objekt.

Sekventiella mönster

- Ex. 1. Köphistoria av en given kund där: element = transaktion = produkter köpta vid tidpunkt t
- Ex. 2. Webbaktivitet av en given användare där: element = sida som användaren besöker
- Ex. 3. Logg av händelser i en given kärnkraftsreaktor där: element = felmeddelande från sensorer i reaktorn

Delsekvens

- En delsekvens av s är en ordnad sekvens som består av element som ingår i s
- Ex:

Time-stamp	Element
10	{2, 3, 4}
20	{4, 1}
23	{1,3}

Exempel på delsekvenser:

<{2, 3}, {1}, {1}>

<{2}, {3}>

<{4}, {4}, {1}>

Utvinning av sekventiella mönster

- Objektsekvens är transaktionslistan som förknippas med ett objekt/individ
- Ex. Tabellen på nästa sida innehåller 5 objektsekvenser
- Support av S är andelen objektsekvenser som innehåller S.
- Hitta alla sekvenser som har $s(S) \geq minsup$

Utvinning av sekventiella mönster

Object	Time-stamp	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

Beräkna:

$\langle\{1, 2\}\rangle \quad s = 0.6$

$\langle\{1\}, \{2\}\rangle \qquad s = 0.8$

$\langle\{2\}, \{2\}\rangle \qquad s = 0.6$

Brute Force metodik

- Uppräkna alla möjliga sekvenser och beräkna supportnivån för varje blir dyrt!
  - 1-sekvenser
    $< i_1 >, < i_2 >, \ldots, < i_n >$
  - 2-sekvenser
    $< \{i_1, i_2\} >, < \{i_1, i_3\} >, \ldots, < \{i_{n-1}, i_n\} >$
  - $< \{i_1\}, \{i_1\} >, < \{i_1\}, \{i_3\} >, \ldots, < \{i_{n-1}\}, \{i_n\} >$
  - 3-sekvenser...

Brute Force problem

- Det finns betydligt fler kandidatsekvenser än kandidatenheter vid analys av frekventa enhetsmängder eftersom:
  - En sekvens kan innehålla ett element flera gånger
  - Elementföljden spelar roll, permutation istället för kombination.
- Detta innebär ännu större problem med dimensionalitet och transaktionsstorlek än icke-sekventiella metodiken.

Apriori-like algoritm

- Apriori-like algoritmen är utvecklad för sekvensanalys

Algorithm 7.1 *Apriori*-like algorithm for sequential pattern discovery.

1: $k = 1$.
2: $F_k = \{ i \mid i \in I \wedge \frac{\sigma(\{i\})}{N} \geq minsup \}$. {Find all frequent 1-subsequences.}
3: **repeat**
4: $k = k + 1$.
5: $C_k = \text{apriori-gen}(F_{k-1})$. {Generate candidate k-subsequences.}
6: **for** each data sequence $t \in T$ **do**
7: $C_t = \text{subsequence}(C_k, t)$. {Identify all candidates contained in t.}
8: **for** each candidate k-subsequence $c \in C_t$ **do**
9: $\sigma(c) = \sigma(c) + 1$. {Increment the support count.}
10: **end for**
11: **end for**
12: $F_k = \{ c \mid c \in C_k \wedge \frac{\sigma(c)}{N} \geq minsup \}$. {Extract the frequent k-subsequences.}
13: **until** $F_k = \emptyset$
14: Answer $= \bigcup F_k$.

- Candidate generation: liknar metoden som används för Apriori-algoritmen

Apriori-like algoritm

Exempel

Frekventa 3-sekvenser

<{1} {2} {3}>
<{1} {2, 5}>
<{1} {5} {3}>
<{2} {3} {4}>
<{2, 5} {3}>
<{3} {4} {5}>
<{5} {3, 4}>

Kandidat-framställning

<{1} {2} {3} {4}>
<{1} {2, 5} {3}>
<{1} {5} {3, 4}>
<{2} {3} {4} {5}>
<{2, 5} {3, 4}>

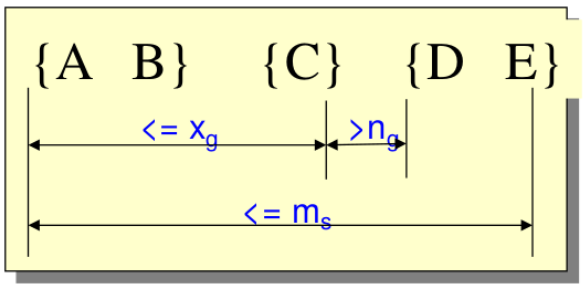
Kandidat-beskärning

<{1} {2, 5} {3}>

Tidsbegränsningar

- **Maxspan** – största tillåtna avstånd mellan första och sista händelsen i sekvensen
- **Mingap** – minsta tillåtna avståndet mellan intilliggande element
- **Maxgap** – största tillåtna avståndet mellan intilliggande element

Tidsbegränsningar



x_g: maxgap

n_g: mingap

m_s: maxspan

x_g = 2, n_g = 0, m_s= 4

Objektsekvens	Delsekvens	Stödjer?
< {2,4} {3,5,6} {4,7} {4,5} {8} >	< {6} {5} >	Yes
< {1} {2} {3} {4} {5}>	< {1} {4} >	No
< {1} {2,3} {3,4} {4,5}>	< {2} {3} {5} >	Yes
< {1,2} {3} {2,3} {3,4} {2,4} {4,5}>	< {1,2} {5} >	No

Projekt

- Se dokument för detaljer
- Ni ska analysera riktig data med metoder från kursen

Avslut

- Kurshemsidan
- Labben