Metody reprezentacji zbiorów częstych

Marzena Kryszkiewicz
Politechnika Warszawska

Wprowadzenie do metod reprezentowania zbiorów częstych

Czy niezbędna jest znajomość wszystkich częstych zbiorow?

- Liczba zbiorów częstych może być olbrzymia.
- Czas ich wykrywania może być bardzo długi.
- Niekiedy wystarczająca jest znajomość małego podzbioru rodziny zbiorów częstych! W szczególności *reprezentatywne reguły asocjacyjne* i *minimalne nieredundantne* można wyprowadzić bezpośrednio ze zwięzłych reprezentacji zbiorów częstych opartych na tzw. *zbiorach zamkniętych* i *generatorach*.

3

Prosty przykład wnioskowania o wsparciach zbiorów

- Niech $sup(\{ac\}) = 3$ oraz $sup(\{abcde\}) = 3$.
- Informacja ta jest wystarczająca do określenia wsparcia zbioru $\{abce\}$ bez dostępu do zbioru danych, co wynika z poniższej obserwacji:

  $$3 = sup(\{ac\}) \geq sup(\{abce\}) \geq sup(\{abcde\}) = 3.$$

  Stąd,

  $$sup(\{ac\}) = sup(\{abce\}) = sup(\{abcde\}) = 3.$$
- W ogólności, jeśli wiadomo, że $X \subseteq Y$ oraz $sup(X) = sup(Y) = k$, to dla każdego zbioru Z spełniającego warunek: $X \subseteq Z \subseteq Y$, jego wsparcie także jest równe k.

4

Bezstratne reprezentacje zbiorów częstych

- *Reprezentacja zbiorów* częstych jest nazywana *bezstratną,* jeżeli umożliwia wyprowadzenie i wyznaczenie wsparcia wszystkich zbiorów częstych bez dostępu do zbioru danych.
- Do bezstratnego reprezentowania zbiorów częstych wykorzystuje się m.in. następujące specyficzne zbiory pozycji:
  - *zbiory zamknięte* (ang. *closed itemsets*),
  - *generatory* (ang. *generators*),
- Wnioskowanie o wsparciach zbiorów prowadzi się z wykorzystaniem zależności pomiędzy tymi zbiorami a ich nadzbiorami lub podzbiorami.

5

Wsparcia i tidlisty podzbiorów/nadzbiorów

- *Tidlista zbioru X* jest oznaczana jako $t(X)$ i definiowana jako lista identyfikatorów tych transakcji w zbiorze danych D, które zawierają X.
- **Lemat.** Niech $X \subseteq Y$. Wtedy:

  $$t(X) = t(Y) \Leftrightarrow sup(X) = sup(Y).$$

  Dowód:

  (⇒). Trywialny.

  (⇐). Niech $X \subseteq Y$ oraz $sup(X) = sup(Y)$.

  Wtedy, $t(X) \supseteq t(Y)$ oraz $|t(X)| = |t(Y)|$.

  Stąd, $t(X) = t(Y)$.

6

Reprezentacja oparta na częstych zbiorach zamkniętych (CR)

Domknięcie zbioru

- *Domknięcie zbioru X* jest oznaczane jako $\gamma(X)$ i definiowane następująco:

  $\gamma(X) = \bigcap\{T \in D \cup \{I\} \mid T \supseteq X\}$,

  gdzie *I* jest *uniwersum*, czyli zbiorem wszystkich pozycji.
- **Uwaga:** Dla każdego zbioru pozycji *X* istnieje dokładnie jeden zbiór pozycji, będący domknięciem zbioru *X*!
- **Własność.** *Y* jest domknięciem zbioru $X \Leftrightarrow Y$ jest największym nadzbiorem zbioru *X*, spełniającym warunek:

  $sup(Y) = sup(X)$.

8

Zbiory zamknięte i wnioskowanie o wsparciach zbiorów ich przy użyciu

- *Zbiór* jest definiowany jako *zamknięty* (ang. *closed itemset*), jeśli jest równy swojemu domknięciu.
- **Ważna własność zbiorów zamkniętych:** Informacja o wsparciach wszystkich zbiorów zamkniętych jest wystarczająca do wyznaczenia wsparcia każdego zbioru *X* w 2^I, a mianowicie:

  $sup(X) = \max\{sup(Y) \mid Y \text{ jest zamknięty} \wedge Y \supseteq X\}$.

9

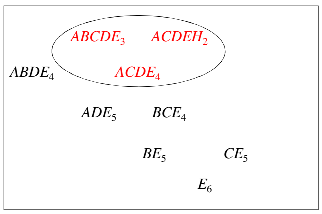
Reprezentacja oparta na częstych zbiorach zamkniętych (CR)

- *Reprezentacja oparta na częstych zbiorach zamkniętych* (CR) składa się ze wszystkich częstych zbiorów zamkniętych wraz z informacją o ich wsparciach.

10

Przykład: Wnioskowanie z użyciem CR

- CR for *minSup*=2.



- {*AF*} nie ma żadnego nadzbioru w CR. Stąd domknięcie zbioru {*AF*} nie jest zbiorem częstym i w konsekwencji: $sup(\{AF\}) \leq minSup$.
- $sup(\{ACD\}) = \max(sup(\{ACDE\}), sup(\{ABCDE\}), sup(\{ACDEH\})) = 4$.

11

Algorytm CHARM - wyznaczanie CR

Id	Transaction
T1	{abcde}
T2	{abcdef}
T3	{abcdehi}
T4	{abe}
T5	{bcdehi}

$h(X \cup Y) = \Sigma_{i \in t(X \cup Y)}\, i$

- $t(X) = t(Y)$:
  - usuń z drzewa węzeł z etykietą *Y*;
  - rozszerz etykietę węzła *X* i etykiety jego dzieci o *Y* (czyli zastąp każde wystąpienie *X* przez $X \cup Y$).
- $t(X) \subset t(Y)$:
  - rozszerz etykietę węzła *X* i etykiety jego dzieci o *Y* (czyli zastąp każde wystąpienie *X* przez $X \cup Y$).
- $t(X) \supset t(Y)$:
  - usuń z drzewa węzeł z etykietą *Y*;
  - dodaj węzeł z etykietą $X \cup Y$ jako dziecko węzła *X*.
- wpp.:
  - dodaj węzeł z etykietą $X \cup Y$ jako dziecko węzła *X*.

12

Algorytm dCHARM - wyznaczanie CR

Id	Transaction
T1	{abcde}
T2	{abcdef}
T3	{abcdehi}
T4	{abe}
T5	{bcdehi}

- $d(X \cup Y) = d(Y) \setminus d(X)$
- $sup(X \cup Y) = sup(X) - |d(X \cup Y)|$
- $h(X \cup Y) = h(X) - (\Sigma_{i \in d(X \cup Y)}\, i)$

- $d(X) = d(Y)$:
  - usuń z drzewa węzeł z etykietą Y;
  - rozszerz etykietę węzła X i etykiety jego dzieci o Y (czyli zastąp każde wystąpienie X przez $X \cup Y$).
- $d(X) \supset d(Y)$:
  - rozszerz etykietę węzła X i etykiety jego dzieci o Y (czyli zastąp każde wystąpienie X przez $X \cup Y$).
- $d(X) \subset d(Y)$:
  - usuń z drzewa węzeł z etykietą Y;
  - dodaj węzeł z etykietą $X \cup Y$ jako dziecko węzła X.
- wpp.:
  - dodaj węzeł z etykietą $X \cup Y$ jako dziecko węzła X.

13

Reprezentacja generatorowa (GR)

Generator zbioru

- Y jest definiowany jako *generator zbioru* X, jeżeli jest minimalnym podzbiorem zbioru X, spełniającym warunek:
  $$\gamma(Y) = \gamma(X).$$
- **Uwaga:** Dla każdego zbioru pozycji X istnieje co najmniej jeden zbiór pozycji, będący generatorem zbioru X!
- **Własność.** Y jest generatorem zbioru $X \Leftrightarrow Y$ jest minimalnym podzbiorem zbioru X, spełniającym warunek:
  $$sup(Y) = sup(X).$$

15

Generator (kluczowy)

- Zbiór X jest definiowany jako *generator* (*kluczowy*), jeżeli generatorem zbioru X jest X.
- **Twierdzenie.**
  - Wszystkie podzbiory generatora są generatorami.
  - Wszystkie nadzbiory zbioru nie będącego generatorem nie są generatorami.

16

Nadzbiory nie-generatorów

Twierdzenie.
Jeżeli X nie jest generatorem, to $\forall Y \supset X$, Y nie jest generatorem.

Dowód. Niech X nie będzie generatorem i $Y \supset X$. Wtedy:
$\exists X' \in G(X)$ taki, że $X' \subset X$ oraz
$\exists Z \neq \emptyset$ taki, że $Z = Y \setminus X$
$\Rightarrow t(X') = t(X)$ oraz
$t(Y) = t(X \cup Z) = t(X) \cap t(Z) = t(X') \cap t(Z) = t(X' \cup Z)$
$\Rightarrow sup(Y) = sup(X \cup Z) = sup(X' \cup Z)$ oraz
$Y = X \cup Z \supset X' \cup Z$
$\Rightarrow Y$ nie jest generatorem.

17

Podzbiory generatorów

Twierdzenie A. Jeżeli X nie jest generatorem, to $\forall Y \supset X$, Y nie jest generatorem.

Twierdzenie B.
Jeżeli X jest generatorem, to $\forall Y \subset X$, Y jest generatorem.

Dowód (przez zaprzeczenie).
Niech X będzie generatorem, $Y \subset X$ oraz Y nie będzie generatorem. Wtedy zgodnie z Twierdzeniem A, wszystkie właściwe nadzbiory zbioru Y (w tym zbiór X) nie są generatorami.
Stąd, X nie jest generatorem, co przeczy założeniu.

18

Wnioskowanie o wsparciach zbiorów przy użyciu generatorów

- **Ważna własność generatorów:** Informacja o wsparciach wszystkich generatorów (kluczowych) jest wystarczająca do wyznaczenia wsparcia każdego zbioru *X* w 2^I, a mianowicie:

  $sup(X) = \min\{sup(Y) \mid Y \text{ jest generatorem } \wedge Y \subseteq X\}$.

19

Reprezentacja generatorowa (GR)

- *Reprezentacja generatorowa* (GR) składa się z:
  1) *Komponentu głównego*, który stanowią wszystkie częste generatory wraz z informacją o ich wsparciach,
  2) *Negatywnej granicy*, którą stanowią wszystkie minimalne rzadkie generatory.

20

Minimalny rzadki generator jest minimalnym rzadkim zbiorem

Twierdzenie. *X* jest minimalnym rzadkim generatorem ⇔ *X* jest minimalnym rzadkim zbiorem.

Proof (⇒). *X* jest minimalnym rzadkim generatorem

⇒ *X* jest minimalnym rzadkim generatorem oraz wszystkie właściwe podzbiory zbioru *X* są generatorami

⇒ *X* jest minimalnym rzadkim generatorem oraz wszystkie właściwe podzbiory zbioru *X* są częstymi generators

⇒ *X* jest rzadki i wszystkie podzbiory zbioru *X* są częste

⇒ *X* jest minimalnym rzadkim zbiorem.

21

Minimalny rzadki zbiór jest minimalnym rzadkim generatorem

Twierdzenie. *X* jest minimalnym rzadkim generatorem ⇔ *X* jest minimalnym rzadkim zbiorem.

Proof (⇐). *X* jest minimalnym rzadkim zbiorem

⇒ *X* jest rzadki i wszystkie właściwe podzbiory zbioru *X* są częste

⇒ *X* jest rzadki i wszystkie właściwe podzbiory zbioru *X* są częste oraz mają wsparcia różne od *sup*(*X*)

⇒ *X* jest rzadkim generatorem oraz wszystkie jego właściwe podzbiory są częste

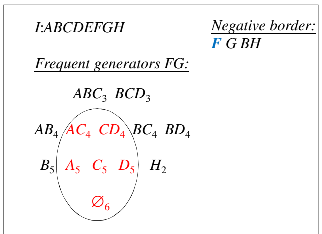
⇒ X jest rzadkim generatorem oraz wszystkie właściwe podzbiory zbioru *X* są częstymi generatorami

⇒ X jest minimalnym rzadkim generatorem.

22

Przykład: Wnioskowanie z użyciem GR

- GR for *minSup*=2.



- {*AF*} jest rzadki, ponieważ jest nadzbiorem zbioru ***F***, który jest (minimalnym) rzadkim zbiorem (generatorem).
- $sup(\{ACD\}) = \min(sup(\{AC\}), sup(\{CD\}), sup(\{A\}), sup(\{C\}), sup(\{D\}), sup(\emptyset)) = 4$.

23

Algorytm GR-Apriori - wyznaczanie GR

Id	Transaction
T1	{abcde}
T2	{abcdef}
T3	{abcdehi}
T4	{abe}
T5	{bcdehi}

- Algorytm GR-Apriori służy do wyznaczania reprezentacji generatorowej GR.
- Ponieważ GR składa się wyłącznie z częstych generatorów oraz minimalnych rzadkich generatorów (czyli minimalnych rzadkich zbiorów), w GR-Apriori wykorzystywane są następujące własności przy tworzeniu kandydatów na elementy reprezentacji GR:
  - Żaden właściwy nadzbiór zbioru nie będącego generatorem, nie jest generatorem, a zatem nie jest elementem reprezentacji GR;
  - Żaden właściwy nadzbiór zbioru rzadkiego, nie jest elementem reprezentacji GR.

24

Literatura

- Marzena Kryszkiewicz: Concise Representation of Frequent Patterns Based on Disjunction-Free Generators. ICDM 2001: 305-312
- Marzena Kryszkiewicz, Marcin Gajek: Concise Representation of Frequent Patterns Based on Generalized Disjunction-Free Generators. PAKDD 2002: 159-171
- Marzena Kryszkiewicz: Concise Representations of Frequent Patterns and Association Rules, Warsaw: Publishing House of Warsaw University of Technology (2002)
- Pasquier N.: Data mining: Algorithmes d'extraction et de Réduction des Règles d'association dans les Bases de Données. Thèse de Doctorat, Université Blaise Pascal – Clermont–Ferrand II (2000)
- Mohammed Javeed Zaki, Ching-Jui Hsiao: Efficient Algorithms for Mining Closed Itemsets and Their Lattice Structure. IEEE Trans. Knowl. Data Eng. 17(4): 462-478 (2005)

25

Literatura dodatkowa – reprezentowanie zbiorów częstych dopuszczających negację

- Marzena Kryszkiewicz: Generalized disjunction-free representation of frequent patterns with negation. J. Exp. Theor. Artif. Intell. 17(1-2): 63-82 (2005)
- Marzena Kryszkiewicz, Katarzyna Cichon: Support Oriented Discovery of Generalized Disjunction-Free Representation of Frequent Patterns with Negation. PAKDD 2005: 672-682
- Marzena Kryszkiewicz: Generalized Disjunction-Free Representation of Frequents Patterns with at Most k Negations. PAKDD 2006: 468-472
- Marzena Kryszkiewicz: Reasoning about Frequent Patterns with Negation. Encyclopedia of Data Warehousing and Mining 2009: 1667-1674
- Marzena Kryszkiewicz: Non-Derivable Item Set and Non-Derivable Literal Set Representations of Patterns Admitting Negation. DaWaK 2009: 138-150
- Marzena Kryszkiewicz, Henryk Rybinski, Katarzyna Cichon: On Concise Representations of Frequent Patterns Admitting Negation. Advances in Machine Learning II 2010: 259-289

26

Ćwiczenia…

1. Niech *sup*({*cd*}) = 20 oraz *sup*({*bcdefghij*}) = 20. Jakie jest wsparcie zbioru {*bcdej*}?
2. Niech *sup*({*ac*}) = 3 oraz *sup*({*abcde*}) = 3. Dla jakich innych zbiorów wiadomo, że ich wsparcie wynosi 3?
3. Rozważ zbiór danych ze slajdu 12. Jakie zbiory pozycji są domknięciami zbioru {*bcd*}?
4. Rozważ zbiór danych ze slajdu 24. Jakie zbiory pozycji są generatorami zbioru {*bcd*}?

27

Ćwiczenia…

5. Na podstawie reprezentacji opartej na częstych zbiorach zamkniętych (CR) ze slajdu 11, określ, czy {*BC*} jest zbiorem częstym. Jeśli tak, określ jego wsparcie. Jeśli nie, podaj optymistyczne oszacowanie wartości wsparcia zbioru {*BC*}.
6. Na podstawie reprezentacji opartej na częstych zbiorach zamkniętych (CR) ze slajdu 11, określ, czy {*BCH*} jest zbiorem częstym. Jeśli tak, określ jego wsparcie. Jeśli nie, podaj optymistyczne oszacowanie wartości wsparcia zbioru {*BCH*}.

28

Ćwiczenia

7. Na podstawie reprezentacji opartej na generatorach (GR) ze slajdu 23, określ, czy {*ABCDE*} jest zbiorem częstym. Jeśli tak, określ jego wsparcie. Jeśli nie, podaj optymistyczne oszacowanie wartości wsparcia zbioru {*ABCDE*}.
8. Na podstawie reprezentacji opartej na generatorach (GR) ze slajdu 23, określ, czy {*BCDH*} jest zbiorem częstym. Jeśli tak, określ jego wsparcie. Jeśli nie, podaj optymistyczne oszacowanie wartości wsparcia zbioru {*BCDH*}.

29