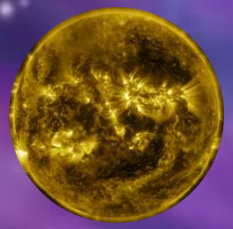
METODY INŻYNIERII WIEDZY

KNOWLEDGE ENGINEERING AND DATA MINING



EKSPLORACJA DANYCH

Ćwiczenia

Adrian Horzyk

Akademia Górniczo-Hutnicza

Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej

Katedra Automatyki i Inżynierii Biomedycznej, Laboratorium Biocybernetyki

30-059 Kraków, al. Mickiewicza 30, paw. C3/205

horzyk@agh.edu.pl, Google: Adrian Horzyk

DOKONAJ BEZPOŚREDNIEJ EKSPLORACJI

1. **Załóżmy, że minimalne wsparcie smin=30% oraz minimala pewność cmin = 20%**
2. **Wykorzystaj zbiór Iris i znajdź:**
   - **Wszystkie wzorce częste**
   - **Wzorce zamknięte**
   - **Wzorce maksymalne**
3. **Weź dowolny tekst (np. Baśnie Grimm) zawierający minimalnie kilkaset zdań i znajdź w nim częste i maksymalne wzorce sekwencyjne, traktując zdania jak transakcje, zaś słowa jak elementy tych transakcji.**
4. **W celu realizacji zadań wykorzystaj metodę Apriori oraz inne poznane na wykładzie**

POLICZ WSPARCIE POSZCZEGÓLNYCH WZORCÓW I POSORTUJ JE WZGLĘDEM CZĘSTOŚCI

1. **Wczytaj zbiór transakcji ze strony:** http://home.agh.edu.pl/~horzyk/lectures/ahdydmiw.php
2. **Oblicz wsparcie (*support*) s – to częstotliwość lub ilość wystąpień wzorca lub zbioru elementów X w analizowanej encji lub transakcji.**
3. **Wyznacz próg na poziomie σ = 50% i określ, które wzorce są częste (*frequent*), tzn. gdy jego wsparcie (*support*) jest nie mniejszy niż ustalony próg σ (*minsup*)**

Przykład:

CZĘSTE > 50%

- ✓ Cukier (80%)
- ✓ Kawa (60%)
- ✓ Jajka (60%)
- ✓ Mleko (40%)
- ✓ Orzeszki (40%)
- ✓ Masło (40%)
- ✓ Chleb (20%)
- ✓ Miód (20%)

ID TRANSAKCJI	ELEMENTY TRANSAKCJI
1	kawa, mleko, cukier, orzeszki
2	kawa, cukier, jajka
3	kawa, chleb, cukier, masło
4	orzeszki, cukier, miód, jajka
5	masło, mleko, jajka

OKREŚL REGUŁY ASOCJACYJNE (wsparcie i pewność) DLA TRANSAKCJI Z POPRZEDNIEGO ĆWICZENIA

- ✓ **Reguły asocjacyjne (*association rules*) elementów transakcji/wzorców: X → Y (s, c).**
- ✓ **Wsparcie (support) s to prawdopodobieństwo, że określona transakcja zawiera X ∪ Y liczone względem wszystkich możliwych transakcji.**
- ✓ **Pewność (*confidence*) c – to prawdopodobieństwo warunkowe, że transakcja zawierająca X zawiera również Y.**
- ✓ **Eksploracja reguł asocjacyjnych polega na odnalezieniu wszystkich reguł X → Y o określonym minimalnym wsparciu s oraz o określonej minimalnej pewności c: np. s ≥ 50%, c ≥ 50%.**
- ✓ **Wielowymiarowe reguły asocjacyjne, np.:**
  wiek (X, „18-24”) ∧ zawód (X, „student”) ⇒ kupuje (X, „cola”)
  wiek (X, „18-24”) ∧ kupuje (X, „pop-corn”) ⇒ kupuje (X, „cola”)

PRZYKŁADY REGUŁ ASOCJACYJNYCH:

- ✓ **Kawa → Cukier (80%, 100%)**
- ✓ **Cukier → Kawa (80%, 75%)**
- ✓ **Cukier → Jajka (100%, 50%)**
- ✓ **Jajka → Cukier (100%, 67%)**

NIE SĄ NIMI dla s ≥ 50%, c ≥ 50%:

- ❖ **Kawa → Jajka (100%, 33%)**
- ❖ **Jajka → Kawa (100%, 33%)**

ID TRANSAKCJI	ELEMENTY TRANSAKCJI
1	kawa, mleko, cukier, orzeszki
2	kawa, cukier, jajka
3	kawa, chleb, cukier, masło
4	orzeszki, cukier, miód, jajka
5	masło, mleko, jajka

ZASTOSUJ REGUŁĘ OCZYSZCZANIA APRIORI

Zastosuj regułę oczyszczania Apriori (*pruning principle*) do usunięcia rzadkich podzbiorów i odfiltrowania częstych.

Reguła Apriori:

Każdy podzbiór zbioru częstego (*frequent itemset*) jest częsty (frequent).

Wniosek:

Jeśli jakikolwiek podzbiór zbioru S jest rzadki (*infrequent*), wtedy S również jest rzadki (*infrequent*).

Powyższy wniosek umożliwia odfiltrowanie wszystkich większych wzorców (*super-patterns*), które zawierają rzadkie (*infrequent*) podzbiory (*itemsubsets*), w celu podniesienia efektywności przeszukiwania wzorców w trakcie ich eksploracji.

Reguła oczyszczania Apriori (*pruning principle*) mówi, iż jeśli istnieje jakikolwiek podzbiór (*itemsubset*), który jest rzadki (*infrequent*), wtedy jego dowolny zawierający go zbiór (*superset*) nie powinien być uwzględniany/generowany w procesie eksploracji.

DOKONAJ EKWIWALENTNEJ TRANSFORMACJI KLAS

Ekwiwalentna Transformacja Klas ECLAT (*Equivalence Class Transformation*) to algorytm przeszukiwania w głąb (depth-first search) wykorzystujący przecięcie zbiorów. Służy do eksploracji częstych wzorców poprzez badanie ich wertykalnego (kolumnowego) formatu:

$t(B) = \{T_2, T_3\}$; $t(C) = \{T_1, T_3\}$ ➔ $t(BC) = \{T_3\}$

$t(E) = \{T_1, T_2, T_3\}$ ➔ diffset (BE, E) = $\{T_1\}$ – zbiór różnic

HORYZONTALNY FORMAT DANYCH	
TRANSAKCJE	ELEMENTY ZBIORU
1	A, C, D, E
2	A, B, E
3	B, C, E

TRANSFORMACJA

WERTYKALNY FORMAT DANYCH	
ELEMENT	LISTA TRANSAKCJI
A	1, 2
B	2, 3
C	1, 3
D	1
E	1, 2, 3

tablica asocjacji

Częsty wzorzec to taki podzbiór elementów, który często występuje w transakcjach. Należy więc w tablicy asocjacji odszukać takie elementy, które równocześnie występują w kilku transakcjach, a więc policzyć przecięcie zbiorów transakcji dla poszczególnych elementów, np. dla C i E otrzymamy podzbiór transakcji {1, 3}.

EKSPLORACJA WZORCÓW SEKWENCYJNYCH

Wzorce sekwencyjne (*sequential patterns*) składają się z sekwencji zbiorów elementów (*sets of items*), zwanych też zdarzeniami (*events*), np.:

<EF(AB)(ABC)D(CF)G>

Elementy zbiorów tworzących sekwencje nie są porządkowane, tzn. ich kolejność nie ma znaczenia: np. (ABC) = (CBA) = (ACB) – zapisujemy je w nawiasach.

Dla poniższej bazy sekwencji i minimalnego progu wsparcia minsup = 3 otrzymamy sekwencyjny wzorzec (*sequential pattern*) <(AB)CA>

TRANSAKCJE	WZORCE SEKWENCYJNE
1	<D(ABC)(BC)A(DF)>
2	<(AE)C(BC)(AE)>
3	<(CF)(AB)(DC)CBA>
4	<EH(AF)CBC>
5	<C(AB)DF(CA)DA>

Wzorce sekwencyjne mają liczne zastosowania, np. w: inżynierii oprogramowania, analizy i porównywania łańcuchów DNA, protein, sekwencji czasowych i zmian w czasie (np. na giełdzie kursów walut, akcji), procedur leczniczych w medycynie, analizie i przewidywaniu pogody, analizy, indywidualnego dostosowania ofert i optymalizacji akcji promocyjnych oraz reklamowych…

EKSPLORACJA APRIORI WZORCÓW SEKWENCYJNYCH

Eksploracja Apriori wzorców sekwencyjnych (apriori-based sequential pattern mining) polega na określeniu częstotliwości wystąpień (wsparcia/*support*) sekwencji jedno, następnie dwu, … elementowych:

<A>, <B>, <C>, <D>, <E>, <F>, <H>

Dla których minimalna częstotliwość czyli wsparcie (*minsup*) jest powyżej pewnego ustalonego progu, np. ≥ 5.

TRANSAKCJE	WZORCE SEKWENCYJNE
1	<D(ABC)(BC)A(DF)>
2	<(AE)C(BC)(AE)>
3	<(CF)(AB)(DC)CBA>
4	<EH(AF)CBC>
5	<C(AB)DF(CA)DA>

KANDYDAT	WSPARCIE
<A>	10
<B>	7
<C>	11
<D>	5
<E>	3
<F>	4
<H>	1

Stopniowo generujemy kandydatów o długości k+1 na podstawie wcześniej wygenerowanych kandydatów o długości k, przy czym zawsze bierzemy pod uwagę tylko tych kandydatów, których wsparcie jest powyżej pewnego ustalonego progu. Postępujemy tak dopóki istnieją dłuższi kandydaci spełniający to kryterium (APRIORI).

Apriori pozwala badać tylko ograniczoną ilość kandydatów, a nie wszystkie podciągi.

KANDYDACI	<A>	<B>	<C>	<D>
<A>	<AA>	<AB>	<AC>	<AD>
<B>	<BA>	<BB>	<BC>	<BD>
<C>	<CA>	<CB>	<CC>	<CD>
<D>	<DA>	<DB>	<DC>	<DD>

KANDYDACI	<A>	<B>	<C>	<D>
<A>		<(AB)>	<(AC)>	<(AD)>
<B>			<(BC)>	<(BD)>
<C>				<(CD)>
<D>				

Eksploracja wzorców wygenerowanych i oczyszczonych na podstawie reguły Apriori nazywana jest algorytmem *Generalized Sequential Pattern (GSP) algorithm for Mining and Pruning*.