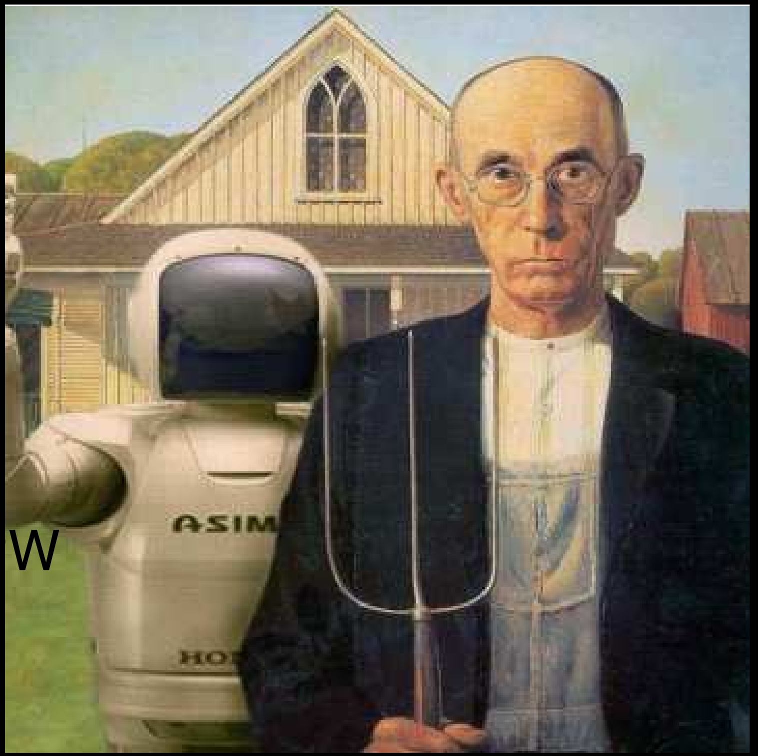
Sztuczna Inteligencja

Wykład 8

Piotr Wąsiewicz

Zakład Sztucznej Inteligencji - ISE PW

pwasiewi@elka.pw.edu.pl



Generowanie reguł asocjacyjnych

Odzyskiwanie wiedzy z tzw. koszyków

Klienci sklepu kupują różne zestawy produktów zwane koszykami ponumerowanymi od 1 do N (ich liczby) np.:

(1 2)

(1 2 3)

(1 2)

(1)

(1 2 3)

Jeśli dany koszyk występuje (też jako część składowa) więcej razy niż zakładana wartość progowa to jest istotny statystycznie np. $\frac{|P_{(1\ 2)}|}{N} > 20\%$, gdzie $|P_{(1\ 2)}|$ jest liczbą koszyków z produktami 1 i 2, a N liczbą wszystkich koszyków.

Odzyskiwanie wiedzy z tzw. koszyków c.d.

Najpierw wybiera się częste podgrupy produktów:
$\frac{|P_{(1)}|}{N} = 1 > 20\%$, $\frac{|P_{(1\ 2)}|}{N} = \frac{4}{5} > 20\%$, $\frac{|P_{(1\ 2\ 3)}|}{N} = \frac{2}{5} > 20\%$.
Następnie z częstych podgrup (1), (1 2), (1 2 3) można tworzyć reguły asocjacyjne czyli kojarzyć te podgrupy ze sobą np.:

$(1) \rightarrow (1\ 2) - (1)$

Do części warunkującej wybrano częsty podzbiór, a w części warunkowanej jest koszyk zawierający tamten podzbiór z części warunkującej (asocjacyjne skojarzenie), ale np. z reguły *modus ponens* $(\frac{a, a \Rightarrow b}{b})$ wynika, że konkluzja nie może zawierać przesłanki, stąd różnica w regule asocjacyjnej. W wyniku tych operacji powstaje reguła:

$(1) \rightarrow (2)$

Inna reguła uzyskana w ten sam sposób to:

$(1\ 2) \rightarrow (1\ 2\ 3) - (1\ 2)$

$(1\ 2) \rightarrow (3)$

W bazach relacyjnych wiedza jest zapisywana w tabelach. W kolumnach występują atrybuty mające zdyskretyzowane wartości np.: atrybut A_1 ma n_1 wartości od v_{11} do v_{1n_1}. W wybranej kolumnie atrybut C jest nazywany kategorią, a jej wartości d etykietami.

A_1 $(v_{11}, \ldots, v_{1n_1})$	$\ldots$ $\ldots$	A_i $(v_{11}, \ldots, v_{1n_i})$	C $(d_1, \ldots, d_{n_C})$
v_{11}	$\ldots$	v_{23}	d_3
v_{12}	$\ldots$	v_{25}	d_{n_C}
v_{1n_1}	$\ldots$	v_{2n_i}	d_1
$\ldots$	$\ldots$	$\ldots$	$\ldots$

Kompleksy

- Kompleks k składa się z selektorów.
- $k_1 = \{< \text{słoneczna} \vee \text{deszczowa}, \text{zimna} \vee \text{ciepła}, ?, ? >\}$
  $k_2 = \{< \text{słoneczna}, \text{ciepła}, ?, ? >\}$
  $k_2 \prec k_1$
  k_2 jest bardziej szczegółowe od k_1, k_1 jest bardziej ogólne od k_2
- $S \rhd k$ to dokładniej $(\exists k \in S) k \rhd x$ - zbiór wszystkich x pokrywanych przez $k \in S$
- $\{k_1 \rhd x\} = \{1, 2, 5, 6, 9\}$
- $\{k_2 \rhd x\} = \{1, 2\}$
- Kompleks tylko z jednym selektorem nieuniwersalnym zwany jest *kompleksem atomowym* .

Reguły asocjacyjne

- *Reguły asocjacyjne* składają się z kompleksów.
- $r = p \Rightarrow q$
  Reguła r składa się z kompleksu *warunkującego* r i kompleksu *warunkowanego* q.
- *Wsparcie reguły* r na zbiorze przykładów P jest określone jako stosunek liczby przykładów ze zbioru P pokrywanych jednocześnie przez dwa kompleksy p i q do liczby wszystkich przykładów:
  $s_r(P) = \frac{|P_{p \wedge q}|}{|P|}$
- *Wiarygodność reguły* r na zbiorze P jest określona jako stosunek liczby przykładów ze zbioru P pokrywanych jednocześnie przez dwa kompleksy p i q do liczby przykładów z P pokrywanych przez kompleks p:
  $f_r(P) = \frac{|P_{p \wedge q}|}{|P_p|}$
- *Wsparcie kompleksu* k analogicznie do porzednich wzorów:
  $s_k(P) = \frac{|P_k|}{|P|}$

Znajdowanie częstych kompleksów

$L_K = \prod_{i=1}^{n}(|A_i|+1)$ - jest liczbą wszystkich kompleksów zawierających tylko selektory pojedyńcze i uniwersalne. Aby znaleźć częste kompleksy należy stosować heurystyki np. założenie, że każdy kompleks zawierający się w pewnym częstym kompleksie jest także częstym kompleksem znajduje zastosowanie w algorytmie *Apriori* , który rozpoczynając od zbioru częstych kompleksów atomowych $\mathbb{S}$ generuje w pętli ich nadzbiory zawierające każdorazowo jeden dodatkowy selektor.

Algorytm *Apriori*

funkcja *częste-kompleksy*(T)

argumenty wejściowe:

- T - zbiór trenujący;

zwraca: zbiór częstych kompleksów dla zbioru trenującego T;

$S_1 := \{k \in \mathbb{S} | s_k(T) \geqslant \theta_s\}$;

dla wszystkich $i = 2, 3, \dots, n$ wykonaj

$S_i' :=$*połączenie* (S_{i-1});

$S_i'' :=$*przycięcie* (S_i', S_{i-1});

$S_i := \{k \in S_i'' | s_k(T) \geqslant \theta_s\}$;

koniec dla

zwróć $\bigcup_{i=1}^n S_i$

Algorytm *Apriori* - połączenie

Kandydatami do połączenia $k \in S'_i$ są dowolne dwa kompleksy $p, q \in S_{i-1}$, dla których spośród ich $i-1$ selektorów nieuniwersalnych $i-2$ są identyczne, a ich pozostałe selektory nieuniwersalne odpowiadają różnym atrybutom (czyli znajdują się na różnych pozycjach). Połączony kompleks $k = p \wedge q$ zawiera i selektorów nieuniwersalnych, z których pierwsze $i-2$ są wspólnymi selektorami kompleksów p i q, a ostatnie dwa są ich różnymi selektorami.

Algorytm Apriori - przycięcie

W związku z heurystyką algorytmu Apriori elementami zbioru S_i'' stają się takie i tylko takie kompleksy ze zbioru S_i', dla których wszystkie zawarte w nich kompleksy o $i-1$ selektorach są elementami zbioru S_{i-1}. Spełnione jest zatem założenie, że każdy kompleks zawierający się w pewnym częstym kompleksie jest także częstym kompleksem.

Generowanie reguł na podstawie częstych kompleksó

- Dla dowolnych kompleksów p i q dla tej samej przestrzeni atrybutów $p \subseteq q \Leftrightarrow p \wedge q = q$ (choć $p \succ q$, to zbiór selektorów nieuniwersalnych decyduje o zawieraniu się).
- Dla dowolnych kompleksów p i q, dla których $p \subseteq q$ oraz dla dowolnego kompleksu k dla tej samej przestrzeni atrybutów $k = q - p \Leftrightarrow q = p \wedge k$ i k jest maksymalnie ogólnym kompleksem spełniającym ten warunek czyli nie istnieje kompleks $k' \succ k$, dla którego $q = p \wedge k'$.
- Dla dowolnych dwóch kompleksów p i q mających wymagane minimalne wsparcie, dla których $p \subset q$ może być utworzona reguła $r = p \Rightarrow q - p$ dla $p \wedge (q - p) = q$ z następującym wsparciem i wiarygodnością:

$$s_r(P) = s_q(P), \quad f_r(P) = \frac{s_q(P)}{s_p(P)}.$$

Tablice kontyngencji

Dwa atrybuty $a_i : X \mapsto A_i$ i $a_j : X \mapsto A_j$ spośród atrybutów $a_1, a_2, \ldots, a_n$ i ich dziedziny $A_i = \{v_{i1}, v_{i2}, \ldots, v_{i|A_i|}\}$ oraz $A_j = \{v_{j1}, v_{j2}, \ldots, v_{j|A_j|}\}$ tworzą *tablicę kontyngencji* zawierającą $|A_i|$ wierszy i $|A_j|$ kolumn, przy czym wartość na przecięciu wiersza o numerze k i kolumny o numerze l równa się liczbie takich przykładów w zbiorze trenującym P dla których $a_i(x) = v_{ik}$ i ijednocześnie $a_j(x) = v_{jl}$ tzn.

$$N_P^{a_i a_j}[v_{ik}, v_{jl}] = |\{x \in P | a_i(x) = v_{ik} \wedge a_j(x) = v_{jl}\}|$$

przy czym $v_{ik} \in A_i$ i $v_{jl} \in A_j$.

Tablice kontyngencji c.d.

- Na dziedzinie X są określone pewne atrybuty:

  $a_1 : X \mapsto \{v_{11}, v_{12}, v_{13}, v_{14}\}$, $a_2 : X \mapsto \{v_{21}, v_{22}, v_{23}\}$

  $a_3 : X \mapsto \{v_{31}, v_{32}, v_{33}\}$, $a_4 : X \mapsto \{v_{41}, v_{42}, v_{43}\}$

  $a_5 : X \mapsto \{v_{51}, v_{52}, v_{53}, v_{54}\}$

- Dla dziedziny X i zbioru trenującego T liczącego 1728 przykładów otrzymujemy następujące przykładowe tablice kontyngencji:

$N_T^{a_1 a_5} =$

	v_{51}	v_{52}	v_{53}	v_{54}
v_{11}	326	81	15	10
v_{12}	300	99	18	15
v_{13}	292	102	18	20
v_{14}	292	102	18	20

$N_T^{a_2 a_5} =$

	v_{51}	v_{52}	v_{53}	v_{54}
v_{21}	576	0	0	0
v_{22}	312	198	36	30
v_{23}	322	186	33	35

$N_T^{a_3 a_5} =$

	v_{51}	v_{52}	v_{53}	v_{54}
v_{31}	450	105	21	0
v_{32}	392	135	24	25
v_{33}	368	144	24	40

$N_T^{a_4 a_5} =$

	v_{51}	v_{52}	v_{53}	v_{54}
v_{41}	576	0	0	0
v_{42}	357	180	39	0
v_{43}	277	204	30	65

- Przykład 1: Dla kompleksów $p_1 =<?, ?, ?, v_{41}, ? >$ i $q_1 =<?, ?, ?, v_{41}, v_{51} >$ ich wsparcie na zbiorze trenującym T wynosi odpowiednio:
  $s_{p_1}(T) = \frac{576}{1728} = 0,333, \quad s_{q_1}(T) = \frac{576}{1728} = 0,333$
  Jeśli $s_{p_1}(T) > \theta_s$ i $s_{q_1}(T) > \theta_s$ to powstaje reguła:
  $r_1 =<?, ?, ?, v_{41}, ? >\Rightarrow<?, ?, ?, ?, v_{51} >$
  $s_{r_1} = s_{q_1}(T) = 0,333, \quad f_{r_1}(T) = \frac{s_{q_1}(T)}{s_{p_1}(T)} = 1$
- Przykład 2: Dla kompleksów $p_2 =<?, ?, ?, v_{42}, ? >$ i $q_2 =<?, ?, ?, v_{42}, v_{51} >$ ich wsparcie na zbiorze trenującym T wynosi odpowiednio:
  $s_{p_2}(T) = \frac{576}{1728} = 0,333, \quad s_{q_2}(T) = \frac{357}{1728} = 0,207$
  Jeśli $s_{p_2}(T) > \theta_s$ i $s_{q_2}(T) > \theta_s$ to powstaje reguła:
  $r_2 =<?, ?, ?, v_{42}, ? >\Rightarrow<?, ?, ?, ?, v_{51} >$
  $s_{r_2} = s_{q_2}(T) = 0,207, \quad f_{r_2}(T) = \frac{s_{q_2}(T)}{s_{p_2}(T)} = 0,620$

Odkrywanie wiedzy - podsumowanie

- Odkrywanie wiedzy czyli zależności w danych (ang. Knowledge Mining, Data Mining) łączy techniki wywodzące się z uczenia się maszyn i statystyki w celu pozyskiwania wiedzy z dużych, rzeczywistych baz danych.
- Tradycyjne statystyczne metody analizy nie pozwalają na odkrywanie zależności o dostatecznej dokładności i ich symbolicznej reprezentacji.
- Hipotezy uzyskane za pomocą kilku algorytmów z dużych i zróżnicowanych zbiorów są na tyle dokładne, że można je łączyć zgodnie z koncepcją metauczenia się.
- Reguły asocjacyjne mówiące o częstym współwystępowaniu pewnych wartości atrybutów stosuje się tam, gdzie niemożliwa jest bardziej precyzyjna i szczegółowa klasyfikacja dla realnych zbiorów danych np. z powodu złożoności obliczeniowej.