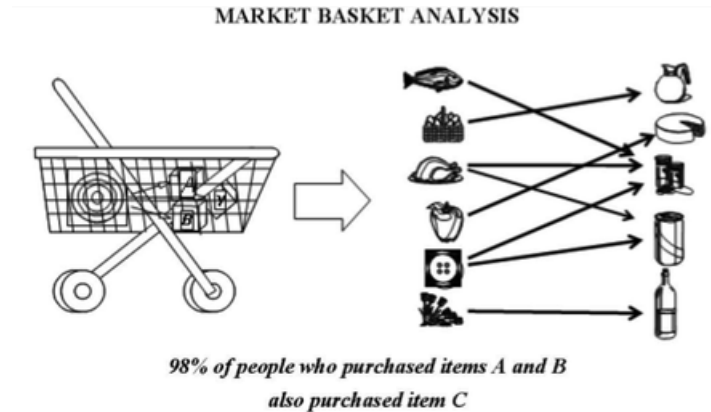


Ćwiczenie 4

Algorytm Apriori i Reguły Asocjacyjne



Zadanie do wykonania

- 1) Tworzymy na pulpicie katalog w formacie Imię_nazwisko, w którym umieszczamy wszystkie pliki związane z ćwiczeniem.
- 2) Czytamy teorię wyliczania zbiorów zdarzeń częstych oraz reguł asocjacyjnych, w razie problemów ze zrozumieniem, analizujemy przykłady na kartce.
- 3) Generujemy zbiór paragonów za pomocą programu paragon_generator.exe.
- 4) Do otrzymanego zbioru implementujemy w wybranym języku algorytm Apriori w celu znalezienia zdarzeń częstych, przyjmując próg częstości $\Phi = 2$.
- 5) Ze zbiorów zdarzeń częstych tworzymy wszystkie możliwe reguły Asocjacyjne spełniające warunki,

$$a) \text{ wsparcie reguły} * \text{ufność reguły} \geq \frac{1}{10}$$

$$b) \text{ wsparcie reguły} * \text{ufność reguły} \geq \frac{2}{10}$$

$$c) \text{ wsparcie reguły} * \text{ufność reguły} \geq \frac{3}{10}$$

$$d) \text{ wsparcie reguły} * \text{ufność reguły} \geq \frac{4}{10}$$

- 6) W przypadku programowania w C++, pomocny może być program demonstracyjny umieszczony na stronie <http://wmii.uwm.edu.pl/~artem> w zakładce Dydaktyka/Sztuczna Inteligencja.

Algorytm Apriori - Tworzenie częstych zbiorów zdarzeń - teoria

- Rozważając pewien zbiór transakcji D (np zbiór paragonów).
- Ustalamy pewien próg częstości Φ , który mówi nam, że zbiór jest częsty, gdy występuje co najmniej Φ razy w D .
- Znajdujemy zbiór F_1 , zawierający jednoelementowe zbiory częste.
- Z kombinacji bez powtórzeń elementów zbioru F_1 tworzymy zbiór C_2 , zawierający dwuelementowych kandydatów na zbiory częste.
- Elementy C_2 , które mieszczą się w progu Φ tworzą zbiór częstych par F_2 .
- Teraz aby znaleźć F_k , Algorytm Apriori tworzy zbiór C_k , k elementowych kandydatów, poprzez łączenie elementów zbioru F_{k-1} , które mają $k - 2$ pierwszych wspólnych pozycji.
- W kolejnym kroku przecinamy C_k własnością Apriori. Dla każdego elementu $c \in C_k$ są tworzone i sprawdzane podzbiory o rozmiarze $k - 1$, jeżeli dowolny z tych podzbiorów nie jest częsty, c nie może być zbiorem częstym i usuwamy go ze zbioru C_k .
- Spośród pozostałych w C_k kandydatów, sprawdzamy ich częstość i usuwamy te zbiory, które nie przekroczyły progu częstości.
- Kończymy szukanie częstych zbiorów zdarzeń w zbiorze D , gdy F_k zawiera 1 lub nie zawiera żadnego elementu.

Reguły Asocjacyjne - teoria

- Reguły Asocjacyjne wyliczamy ze zbiorów zdarzeń częstych F_k, F_{k-1}, \dots, F_2 zbioru transakcji D , wybieramy z każdego zbioru częstego $x \in F_k$ podzbiory wielkości $k - 1$ i tworzymy reguły postaci,

$$ss \Rightarrow s - ss$$

gdzie, ss jest poprzednikiem (wybrany $k - 1$ elementowy zbiorem)
 $s - ss$ jest jednoelementowym następnikiem

(Istnieje możliwość budowania reguł, których następnik zawiera więcej niż jeden element, ale najczęściej stosujemy następnik jednoelementowy)

- Reguły możemy wartościować parametrami,

$$\text{Wsparcie reguły} = \frac{\text{liczba obiektów ze zbioru } D, \text{ do których pasuje reguła w sensie poprzednik} \Rightarrow \text{następnik}}{\text{liczba obiektów zbioru } D}$$

$$\text{Ufność reguły} = \frac{\text{liczba obiektów ze zbioru } D, \text{ do których pasuje reguła w sensie poprzednik} \Rightarrow \text{następnik}}{\text{liczba obiektów ze zbioru } D, \text{ do których pasuje reguła w sensie poprzednika}}$$

- Jeżeli ustalimy próg jakości reguł $Wsparcie\ reguły * Ufność\ reguły$, akceptujemy tylko reguły, które go przekraczają.

Przykład budowania zbioru zdarzeń częstych (Algorytm Apriori)

Dla zbioru transakcji D postaci,
 $\{\text{kapusta, ogórki, pomidory, kabaczki}\}$
 $\{\text{ogórki, pomidory, kabaczki}\}$
 $\{\text{cytryny, pomidory, woda}\}$
 $\{\text{cytryny, woda, jajka}\}$
 $\{\text{ogórki, grzybki, żółądkowa}\}$
 $\{\text{żółądkowa, ogórki, pomidory}\}$

ustalamy próg częstości $\Phi = 2$

Budujemy zbiory zdarzeń częstych

$F_1 = \{\{\text{ogórki}\}, \{\text{pomidory}\}, \{\text{kabaczki}\}, \{\text{cytryny}\}, \{\text{woda}\}, \{\text{żółądkowa}\}\}$
 Sortujemy F_1 alfabetycznie,
 $F_1 = \{\{\text{cytryny}\}, \{\text{kabaczki}\}, \{\text{ogórki}\}, \{\text{pomidory}\}, \{\text{woda}\}, \{\text{żółądkowa}\}\}$

Teraz kombinacje bez powtórzeń elementów zbioru F_1 , tworzą zbiór kandydatów,
 $C_2 = \{\{\text{cytryny, kabaczki}\}, \{\text{cytryny, ogórki}\}, \{\text{cytryny, pomidory}\}, \{\text{cytryny, woda}\},$
 $\{\text{cytryny, żółądkowa}\}, \{\text{kabaczki, ogórki}\}, \{\text{kabaczki, pomidory}\}, \{\text{kabaczki, woda}\},$
 $\{\text{kabaczki, żółądkowa}\}, \{\text{ogórki, pomidory}\}, \{\text{ogórki, woda}\}, \{\text{ogórki, żółądkowa}\},$
 $\{\text{pomidory, woda}\}, \{\text{pomidory, żółądkowa}\}, \{\text{woda, żółądkowa}\}\}$

Do F_2 trafiają ci kandydaci, którzy mają częstość przynajmniej 2,
 $F_2 = \{\{\text{cytryny, woda}\}, \{\text{kabaczki, ogórki}\}, \{\text{kabaczki, pomidory}\},$
 $\{\text{ogórki, pomidory}\}, \{\text{ogórki, żółądkowa}\}\}$

Teraz łączymy elementy zbioru F_2 , o $k - 2$ pierwszych identycznych pozycjach, czyli w tym przypadku na pierwszej pozycji, tworząc zbiór kandydatów,
 $C_3 = \{\{\text{kabaczki, ogórki, pomidory}\}, \{\text{ogórki, pomidory, żółądkowa}\}\}$

Zbiory zawarte w C_3 , przecinamy własnością Apriori, czyli odrzucamy kandydatów, którzy zawierają podzbiór długości 2, nie będący elementem zbioru F_2

Sprawdzamy

Dla $\{\text{kabaczki, ogórki, pomidory}\}$, podzbiory długości 2 są postaci,
 $\{\text{kabaczki, ogórki}\}$ jest w F_2
 $\{\text{kabaczki, pomidory}\}$ jest w F_2
 $\{\text{ogórki, pomidory}\}$ jest w F_2

Sprawdzamy zbiór

$\{\text{ogórki}, \text{pomidory}, \text{żółądkowa}\}$
 podzbiory długości 2 są postaci,
 $\{\text{ogórki}, \text{pomidory}\}$ jest w F_2
 $\{\text{ogórki}, \text{żółądkowa}\}$ nie jest w F_2 , czyli nie jest zbiorem częstym w sensie $\Phi = 2$
 $\{\text{pomidory}, \text{żółądkowa}\}$ jest w F_2
 stąd zbiór $\{\text{ogórki}, \text{pomidory}, \text{żółądkowa}\}$, nie jest częsty usuwamy go ze zbioru C_3

$$C_3 = \{\{\text{kabaczki}, \text{ogórki}, \text{pomidory}\}\}$$

Teraz sprawdzamy, czy pozostali w C_3 kandydaci mają w D częstość 2,

Zbiór $\{\text{kabaczki}, \text{ogórki}, \text{pomidory}\}$ ma w D częstość 2, stąd,

$F_3 = \{\{\text{kabaczki}, \text{ogórki}, \text{pomidory}\}\}$, F_3 zawiera jeden element, warunek stopu został spełniony, kończymy algorytm Apriori.

Liczmy reguły Asocjacyjne

Naszym progiem jakości będzie $wsp * ufn \geq \frac{1}{3}$
 gdzie, $ufn = ufn\acute{o}ść\ reguły$, $wsp = wsparcie\ reguły$

ze zbioru F_3 mamy

$$\begin{aligned}
 \text{kabaczki} \wedge \text{ogórki} &\Rightarrow \text{pomidory} \quad wsp = \frac{1}{3}, \quad ufn = \frac{2}{2} = 1.0, \quad wsp * ufn = \frac{1}{3} \\
 \text{kabaczki} \wedge \text{pomidory} &\Rightarrow \text{ogórki} \quad wsp = \frac{1}{3}, \quad ufn = 1.0, \quad wsp * ufn = \frac{1}{3} \\
 \text{ogórki} \wedge \text{pomidory} &\Rightarrow \text{kabaczki} \quad wsp = \frac{1}{3}, \quad ufn = \frac{2}{3}, \quad wsp * ufn = \frac{2}{9}
 \end{aligned}$$

ze zbioru F_2 mamy

$$\begin{aligned}
 \text{cytryny} &\Rightarrow \text{woda} \quad wsp = \frac{1}{3}, \quad ufn = 1.0, \quad wsp * ufn = \frac{1}{3} \\
 \text{woda} &\Rightarrow \text{cytryny} \quad wsp = \frac{1}{3}, \quad ufn = 1.0, \quad wsp * ufn = \frac{1}{3} \\
 \text{kabaczki} &\Rightarrow \text{ogórki} \quad wsp = \frac{1}{3}, \quad ufn = 1.0, \quad wsp * ufn = \frac{1}{3} \\
 \text{ogórki} &\Rightarrow \text{kabaczki} \quad wsp = \frac{1}{3}, \quad ufn = \frac{1}{2}, \quad wsp * ufn = \frac{1}{6} \\
 \text{kabaczki} &\Rightarrow \text{pomidory} \quad wsp = \frac{1}{3}, \quad ufn = 1.0, \quad wsp * ufn = \frac{1}{3} \\
 \text{pomidory} &\Rightarrow \text{kabaczki} \quad wsp = \frac{1}{3}, \quad ufn = \frac{1}{2}, \quad wsp * ufn = \frac{1}{6} \\
 \text{ogórki} &\Rightarrow \text{pomidory} \quad wsp = \frac{1}{2}, \quad ufn = \frac{3}{4}, \quad wsp * ufn = \frac{3}{8} \\
 \text{pomidory} &\Rightarrow \text{ogórki} \quad wsp = \frac{1}{2}, \quad ufn = \frac{3}{4}, \quad wsp * ufn = \frac{3}{8} \\
 \text{ogórki} &\Rightarrow \text{żółądkowa} \quad wsp = \frac{1}{3}, \quad ufn = \frac{1}{2}, \quad wsp * ufn = \frac{1}{6} \\
 \text{żółądkowa} &\Rightarrow \text{ogórki} \quad wsp = \frac{1}{3}, \quad ufn = 1.0, \quad wsp * ufn = \frac{1}{3}
 \end{aligned}$$

Po uwzględnieniu progu jakości $wsp * ufn \geq \frac{1}{3}$, reguły o częstości przynajmniej $\Phi = 2$ mają postać,

$$\begin{aligned}
 \text{kabaczki} \wedge \text{ogórki} &\Rightarrow \text{pomidory} \quad wsp * ufn = \frac{1}{3} \\
 \text{kabaczki} \wedge \text{pomidory} &\Rightarrow \text{ogórki} \quad wsp * ufn = \frac{1}{3} \\
 \text{cytryny} &\Rightarrow \text{woda} \quad wsp * ufn = \frac{1}{3} \\
 \text{woda} &\Rightarrow \text{cytryny} \quad wsp * ufn = \frac{1}{3} \\
 \text{kabaczki} &\Rightarrow \text{ogórki} \quad wsp * ufn = \frac{1}{3} \\
 \text{kabaczki} &\Rightarrow \text{pomidory} \quad wsp * ufn = \frac{1}{3} \\
 \text{ogórki} &\Rightarrow \text{pomidory} \quad wsp * ufn = \frac{3}{8} \\
 \text{pomidory} &\Rightarrow \text{ogórki} \quad wsp * ufn = \frac{3}{8}, \quad \text{żółądkowa} \Rightarrow \text{ogórki} \quad wsp * ufn = \frac{1}{3}
 \end{aligned}$$