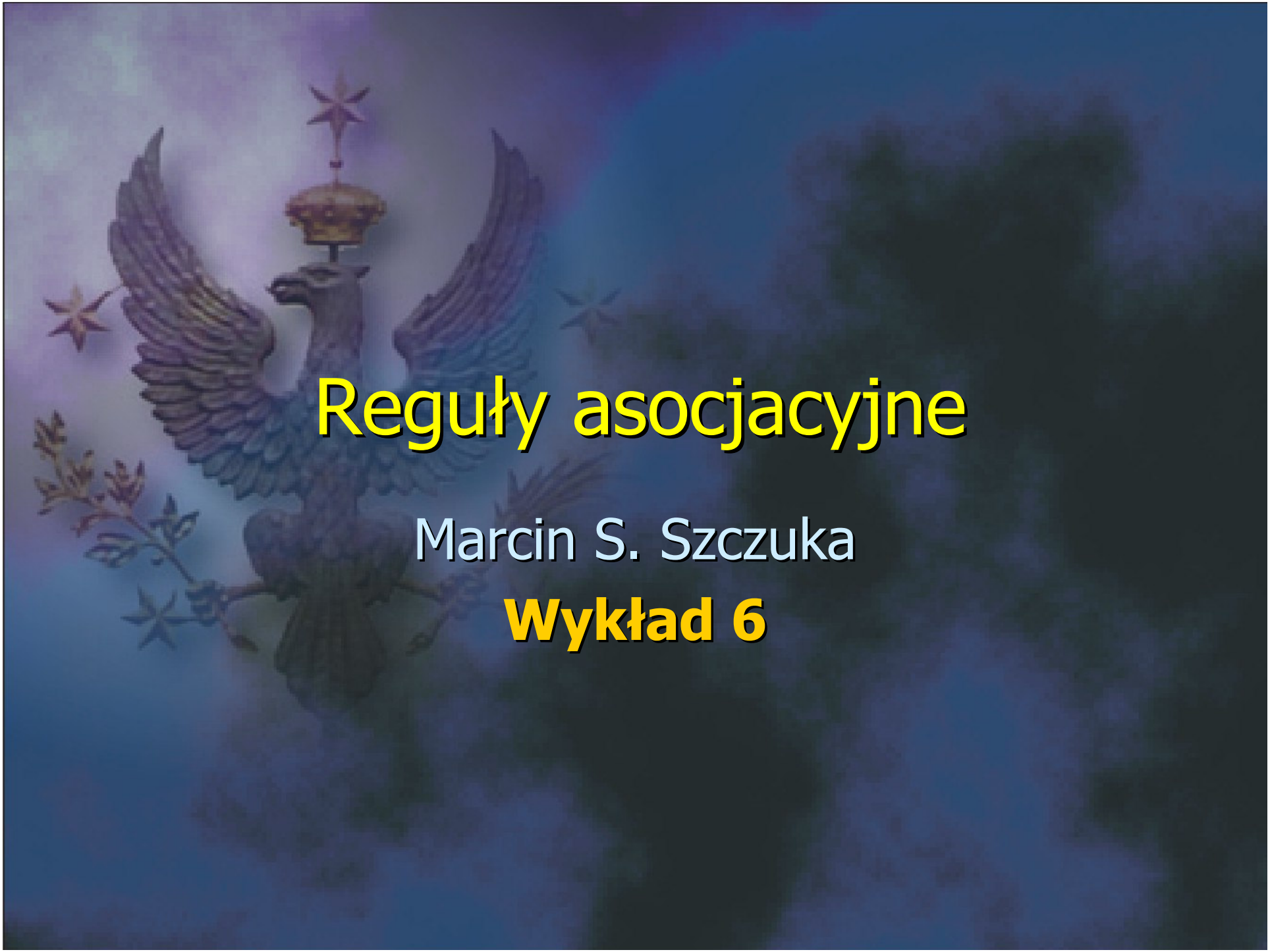

Reguły asocjacyjne

Marcin S. Szczuka

Wykład 6

Plan wykładu

- Terminologia dla reguł asocjacyjnych.
- Ogólny algorytm znajdowania reguł.
- Wyszukiwanie częstych zbiorów.
- Konstruowanie reguł - APRIORI.

Reguły asocjacyjne

Regułą asocjacyjną nazwiemy formułę postaci:

Jeśli ***warunki*** to ***efekty***

zwykle zapisywane jako:

$$\textit{warunki} \rightarrow \textit{efekty}$$

Przykłady asocjacji

(Status=Open) $\wedge$ (Gender=Male) $\wedge$ (Age=Young) $\Rightarrow$ (Activity=Active) $\wedge$ (ClientType=N)

Jeśli dochody przekraczają 50 tyś i miasto ma więcej niż 200 tyś. $\Rightarrow$ posiada samochód i wyjeżdżał za granicę w ostatnim roku.

Jeśli w piątkowy wieczór mężczyzna kupuje w supermarkecie piwo $\Rightarrow$ kupuje też pieluchy.

Motywacja

Chcemy wydobywać z danych związki (wzorce), które

- Mają intuicyjną interpretację.
- Sprawdzają się w licznych przypadkach.
- Mogą pokazać istotne, wcześniej nieuświadomione związki.
- Pozwalają uzupełnić brakującą informację.
- Mają przełożenie na zysk ☺

Oryginalnie problem formułowany jako badanie zawartości koszyka w sklepie.

Cel

Chcemy konstruować reguły które są:

- Zgodne z danymi na jak największym podzbiorze.
- Mocne i szczegółowe.
- Możliwe do wyliczenia (przy stosowaniu rozsądnych algorytmów) dla dużych zbiorów danych.
- Zrozumiałe

Oznaczenia

Dla atrybutu (cechy) $a \rightarrow V_a$ wprowadzamy pojęcie selektora (jak w zwykłych regułach).

Selektor s identyfikujemy z podzbiorem V_s zbioru wartości atrybutu.

Część warunkową i wynikową reguły przedstawiamy zwykle jako zestaw selektorów:

$$\{s_1, s_2, \ldots, s_m\}$$

Oznaczenia c.d.

Tradycyjnie (z analizy koszyka) rozważa się przede wszystkim selektory pojedyncze, które nazywa się przedmiotami (od ang. item).

O zestawach selektorów mówi się jako o zbiorach przedmiotów (od ang. itemset).

Interesują nas tzw. zbiory częste (przedmiotów)

(od ang. frequent itemsets)

Jeszcze troche oznaczeń

T – zbiór przykładów (transakcji)

(s=v) – pojedynczy przedmiot, selektor prosty

S – zbiór wszystkich selektorów prostych występujących w danych.

$T_{\mathbf{p}}$ – zbiór przykładów (transakcji) w których występuje zbiór przedmiotów **p**

p $\Subset$ **q** – wszystkie przedmioty z **p** występują w **q**

|p| - rozmiar zbioru przedmiotów - liczba występujących w nim pojedynczych selektorów.

Ważne miary dla reguł

Mamy regułę asocjacyjną $\mathbf{p} \Rightarrow \mathbf{q}$ dla zbiorów przedmiotów (warunków) **p** i **q**

Wsparcie (pokrycie) dla warunku **p**:

$$s_p(T) = |T_p|/|T|$$

Wsparcie (pokrycie) dla reguły $\mathbf{p} \Rightarrow \mathbf{q}$:

$$s_{p \Rightarrow q}(T) = |T_p \cap T_q|/|T|$$

Poziom zaufania (dokładność) dla reguły $\mathbf{p} \Rightarrow \mathbf{q}$:

$$c_{p \Rightarrow q}(T) = |T_p \cap T_q|/|T_p|$$

Generowanie reguł

Nie można po prostu wygenerować wszystkich reguł asocjacyjnych, a potem wybrać najlepszych.

W złośliwym przypadku reguł może być nawet

$$O(n \cdot 2^{n-1})$$

gdzie n – liczba przedmiotów (atrybutów).

Ponadto, ponieważ chcemy działać na dużych zbiorach danych, powinniśmy unikać wielokrotnego ich przeglądania.

Szukanie reguł

1. Ustalamy pożądany poziom wsparcia (pokrycia) θ i zaufania (dokładności) λ.
2. Tworzymy rodzinę zbiorów przedmiotów, które mają pokrycie powyżej ustalonego progu θ.
3. Z uzyskanych w 2 zbiorów tworzymy reguły o zaufaniu powyżej ustalonego progu λ.

Kluczowa obserwacja !!

Każdy podzbiór częstego zbioru przedmiotów jest zbiorem częstym.

Tylko zbiory zbudowane z częstych podzbiorów mają szansę być częste.

$$\forall_{p' \in p} \left(s_{p'}(T) > \theta\right) \Leftrightarrow \left(s_{p}(T) > \theta\right)$$

Znajdowanie częstych zbiorów

Algorytm APRIORI – przy ustalonym progu na wsparcie (pokrycie):

Wyszukaj wszystkie częste 1-zbiory.

Korzystając z kluczowej obserwacji spróbuj skonstruować z częstych 1-zbiorów, częste 2-zbiory – odrzuć te 2-zbiory, które są poniżej progu.

…

Korzystając z kluczowej obserwacji spróbuj skonstruować z częstych k-zbiorów, częste (k+1)-zbiory – odrzuć te (k+1)-zbiory, które są poniżej progu. Zwiększ k o jeden.

Zatrzymaj się gdy nie ma już większych częstych zbiorów lub gdy wszystkie przedmioty zostaną wykorzystane.

APRIORI(T,θ)

$S_1 := \{ \mathbf{p} \in \mathbf{S} \mid s_{\mathbf{p}}(T) > \theta \};$

for $k=1$ **to** n **do**

$S'_k := \text{join}(S_{k-1});$

$S''_k := \text{prune}(S'_k, S_{k-1});$

$S_k := \{ \mathbf{p} \in S''_k \mid s_{\mathbf{p}}(T) > \theta \};$

end;

return $S = \bigcup_{i=1}^{n} S_i;$

APRIORI - join

Z rodziny (k-1)-zbiorów S_{k-1} wybieramy pary takich zbiorów, które mają dokładnie k-2 wspólne elementy. Jeśli weźmiemy sumę takiej pary to dostaniemy k-zbiór. Dodatkowo stawiamy warunek, aby te zbiory różniły się na przedmiotach związanych z różnymi atrybutami. W ten sposób otrzymujemy kandydatów na częste k-zbiory.

Formalnie:

Niech **p**= { s_1, …, s_{k-2}, s_{k-1} } , **q**= { t_1, …, t_{k-2}, t_{k-1} } i $s_i=t_i$ dla i=1,… ,k-2.

Ponadto selektory s_{k-1} i t_{k-1} są zdefiniowane dla różnych atrybutów.

Wtedy rezultat oznaczony przez **p** $\vee$ **q** jest równy:

{ s_1, …, s_{k-2}, s_{k-1},t_{k-1} }.

APRIORI - `prune`

Zachowujemy w S''_k tylko te k-zbiory, których każdy podzbiór rozmiaru k-1 jest częsty, czyli występuje w S_{k-1}. Jest to dosłowne stosowanie kluczowej obserwacji.

$$(\mathbf{p} \in S_k) \Leftrightarrow (\forall_{\mathbf{q} \in \mathbf{p}} (|\mathbf{q}|=k\text{-}1) \Rightarrow (\mathbf{q} \in S_{k-1}))$$

Kolejna kluczowa obserwacja

Jeśli mamy dwie reguły asocjacyjne
p $\Rightarrow$ **q** i **p'** $\Rightarrow$ **q'** takie, że **q**=**q'** $\wedge$ **r** oraz **p'**=**p** $\wedge$ **r** dla pojedynczego selektora (1-zbioru) **r**=(s=v), to:

$$(c_{\mathbf{p} \Rightarrow \mathbf{q}}(T) > \lambda) \Rightarrow (c_{\mathbf{p} \Rightarrow \mathbf{q}}(T) > \lambda)$$

Oznacza to, że z punktu widzenia zaufania nie warto rozpatrywać reguły **p** $\Rightarrow$ **q** jeśli reguła powstała z tego samego zbioru (**p** $\wedge$ **q**) i mająca krótszą część decyzyjną (następnik) nie przekracza progu ustalonego dla dokładności.

Tworzenie reguł ze zbiorów

Mając k-zbiory częste zamieniamy je na reguły.

Chcielibyśmy dostać reguły które:

- Mają mały poprzednik i duży następnik
- Mają poziom zaufania powyżej progu λ.

Tworzenie reguł

Zaczynamy mając zbiory częste z $S = \bigcup_{i=1}^{n} S_i$ oraz ograniczenie λ.

- Tworzymy wszystkie reguły o jednym elemencie w następniku dla każdego ze zbiorów w S.
- Usuwamy reguły o dokładności mniejszej od λ z dalszych rozważań.
- W każdym następnym kroku staramy się dla każdej z jeszcze nie odrzuconych reguł przerzucić jeden warunek (przedmiot) ze strony poprzednika na stronę następnika. Odrzucamy te spośród tak otrzymanych reguł, które mają niedostateczną dokładność (poniżej λ).

Podsumowanie

- Przedstawiony algorytm został opracowany dla atrybutów binarnych, ale rozszerza się na ogólniejszy przypadek.
- Właściwy wybór ograniczeń (θ , λ) ma krytyczne znaczenie. Dlatego warto zaczynać od „bezpiecznych” wartości.
- W praktyce trzeba korzystać ze specjalnych struktur danych do przechowywania przykładów, zbiorów częstych i reguł.
- Stosowalność tego podejścia zależy od implementacji i doboru parametrów.