GDAŃSK UNIVERSITY OF TECHNOLOGY

# Eksploracja danych

## KLASYFIKACJA I REGRESJA cz. 2


***Wojciech Waloszek***

*wowal@eti.pg.gda.pl*

***Teresa Zawadzka***

*tegra@eti.pg.gda.pl*

*Katedra Inżynierii Oprogramowania*
*Wydział Elektroniki, Telekomunikacji i Informatyki*
*Politechnika Gdańska*


GDAŃSK UNIVERSITY OF TECHNOLOGY

# Budowa reguł decyzyjnych

- Reguły decyzyjne są bardzo popularną formą wyrażania zasad klasyfikacji,

- Budowane są różnymi metodami, jedną z nich jest metoda oparta na podejściu separate-and-conquer

GDAŃSK UNIVERSITY OF TECHNOLOGY

# Pokrycie i poprawność reguły

| S.C. | D.O.R. | Wiek | Wykształcenie | Sam. | Z.K. |
|---|---|---|---|---|---|
| S | 800 | 32 | wyższe | tak | tak |
| S | 1200 | 35 | średnie | tak | tak |
| S | 700 | 26 | podstawowe | nie | nie |
| M | 600 | 45 | wyższe | nie | tak |
| M | 650 | 38 | średnie | tak | tak |
| S | 900 | 28 | wyższe | nie | nie |
| S | 1100 | 65 | średnie | tak | nie |
| M | 500 | 22 | średnie | nie | nie |
| S | 800 | 43 | podstawowe | tak | nie |

**`if Sam.=nie then Z.K.=nie`**

**Pokrycie: *s* = 4** **Poprawność: *a* = 3 / 4**

GDAŃSK UNIVERSITY OF TECHNOLOGY

# Pokrycie i poprawność reguły (2)

- Pokrycie (ang. *coverage*, *support*) to liczba przykładów, dla których zadziała reguła (które spełniają część testową reguły).

- Poprawność (ang. *accuracy*, *confidence*) to liczba przykładów poprawnie klasyfikowanych przez regułę.

GDAŃSK UNIVERSITY OF TECHNOLOGY

# Budowa reguł

| S.C. | D.O.R. | Wiek | Wykształcenie | Sam. | Z.K. |
|---|---|---|---|---|---|
| S | 800 | 32 | wyższe | tak | tak |
| S | 1200 | 35 | średnie | tak | tak |
| S | 700 | 26 | podstawowe | nie | nie |
| M | 600 | 45 | wyższe | nie | tak |
| M | 650 | 38 | średnie | tak | tak |
| S | 900 | 28 | wyższe | nie | nie |
| S | 1100 | 65 | średnie | tak | nie |
| M | 500 | 22 | średnie | nie | nie |
| S | 800 | 43 | podstawowe | tak | nie |

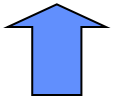
1. **Mamy wyróżniony atrybut decyzyjny, wyznaczający *klasy***
2. **Budujemy regułę pod kątem najlepszej poprawności i, w drugim rzędzie, największego pokrycia.**
3. **Rozpoczynamy od pustej reguły wyznaczając kolejne testy.**

GDAŃSK UNIVERSITY OF TECHNOLOGY

# Budowa reguły decyzyjnej

Do reguły pustej:

```
if ? then Z.K.=?
```



wstawiamy „na próbę” wszystkie możliwe testy proste (z rezultatem):

```
S.C.=S
S.C.=M
Wykształcenie=podstawowe
Wykształcenie=średnie
Wykształcenie=wyższe
Sam.=tak
Sam.=nie
```

GDAŃSK UNIVERSITY OF TECHNOLOGY

# Budowa reguły decyzyjnej (2)

| S.C. | D.O.R. | Wiek | Wykształcenie | Sam. | Z.K. |
|---|---|---|---|---|---|
| S | 800 | 32 | wyższe | tak | tak |
| S | 1200 | 35 | średnie | tak | tak |
| S | 700 | 26 | podstawowe | nie | nie |
| M | 600 | 45 | wyższe | nie | tak |
| M | 650 | 38 | średnie | tak | tak |
| S | 900 | 28 | wyższe | nie | nie |
| S | 1100 | 65 | średnie | tak | nie |
| M | 500 | 22 | średnie | nie | nie |
| S | 800 | 43 | podstawowe | tak | nie |

`if S.C.=S then Z.K.=?`

`if S.C.=S then Z.K.=nie`

**Pokrycie: *s* = 6**

**Poprawność: *a* = 4 / 6**

GDAŃSK UNIVERSITY OF TECHNOLOGY

# Budowa reguły decyzyjnej (3)

Po zbadaniu wszystkich testów otrzymujemy:

| | *a* | *s* |
|---|---|---|
| S.C.=S | 4/6 | 6 |
| S.C.=M | 2/3 | 3 |
| Wykształcenie=podstawowe | 2/2 | 2 |
| Wykształcenie=średnie | 2/4 | 4 |
| Wykształcenie=wyższe | 2/3 | 3 |
| Sam.=tak | 3/5 | 5 |
| Sam.=nie | 3/4 | 4 |

Wybieramy test o największej poprawności

GDAŃSK UNIVERSITY OF TECHNOLOGY

# Budowa reguły decyzyjnej (4)

- Ponieważ osiągnięta poprawność wynosi 100%, pierwsza reguła jest gotowa:

```
if Wykształcenie=podstawowe then Z.K.=nie
```

- Następne reguły budowane są analogicznie po zastosowaniu zasady separate-and-conquer

GDAŃSK UNIVERSITY OF TECHNOLOGY

# Separate-and-conquer

| S.C. | D.O.R. | Wiek | Wykształcenie | Sam. | Z.K. |
|---|---|---|---|---|---|
| S | 800 | 32 | wyższe | tak | tak |
| S | 1200 | 35 | średnie | tak | tak |
| S | 700 | 26 | podstawowe | nie | nie |
| M | 600 | 45 | wyższe | nie | tak |
| M | 650 | 38 | średnie | tak | tak |
| S | 900 | 28 | wyższe | nie | nie |
| S | 1100 | 65 | średnie | tak | nie |
| M | 500 | 22 | średnie | nie | nie |
| S | 800 | 43 | podstawowe | tak | nie |

1. **W zbiorze przykładów wyróżniliśmy dwa pokrywane przez regułę**
2. **Usuwamy te przykłady z naszego zbioru trenującego, a dla pozostałej jego części stosujemy ponownie procedurę budowy najlepszej reguły**

GDAŃSK UNIVERSITY OF TECHNOLOGY

# Budowa kolejnych reguł

- Ponownie rozpoczynamy od reguły pustej, rozważając pozostałe testy:

```
S.C.=S
S.C.=M
Wykształcenie=średnie
Wykształcenie=wyższe
Sam.=tak
Sam.=nie
```

GDAŃSK UNIVERSITY OF TECHNOLOGY

# Budowa kolejnych reguł (2)

Po zbadaniu wszystkich testów otrzymujemy:

| | *a* | *s* |
|---|---|---|
| S.C.=S | 2/4 | 4 |
| S.C.=M | 2/3 | 3 |
| Wykształcenie=średnie | 2/4 | 4 |
| Wykształcenie=wyższe | 2/3 | 3 |
| Sam.=tak | 3/4 | 4 |
| Sam.=nie | 2/3 | 3 |

Wybieramy test o największej poprawności

GDAŃSK UNIVERSITY OF TECHNOLOGY

# Budowa kolejnych reguł (3)

- Ponieważ osiągnięta poprawność wynosi 75%, regułę można jeszcze poprawić:

```
if Sam.=tak then Z.K.=tak
```

- Regułę możemy rozbudowywać, wybierając kolejne testy:

```
if Sam.=tak and ? then Z.K.=?
```

GDAŃSK UNIVERSITY OF TECHNOLOGY

# Rozbudowa reguły

| S.C. | D.O.R. | Wiek | Wykształcenie | Sam. | Z.K. |
|---|---|---|---|---|---|
| S | 800 | 32 | wyższe | tak | tak |
| S | 1200 | 35 | średnie | tak | tak |
| S | 700 | 26 | podstawowe | nie | nie |
| M | 600 | 45 | wyższe | nie | tak |
| M | 650 | 38 | średnie | tak | tak |
| S | 900 | 28 | wyższe | nie | nie |
| S | 1100 | 65 | średnie | tak | nie |
| M | 500 | 22 | średnie | nie | nie |
| S | 800 | 43 | podstawowe | tak | nie |

**Przy rozbudowywaniu reguły ograniczmy się do rozpatrywania przykładów pokrywanych przez pierwszy test niepokrywanych przez poprzednie reguły.**

```
if Sam.=tak and ? then Z.K.=?
```

GDAŃSK UNIVERSITY OF TECHNOLOGY

# Rozbudowa reguły (2)

- W miejsce „?” moglibyśmy wstawić jeden z testów wartości atrybutów nominalnych:

```
S.C.=S
S.C.=M
Wykształcenie=średnie
Wykształcenie=wyższe
```

- Pokrycie i poprawność tak wygenerowanej reguły sprawdzamy analogicznie jak poprzednio

GDAŃSK UNIVERSITY OF TECHNOLOGY

# Rozbudowa reguły (3)

| S.C. | D.O.R. | Wiek | Wykształcenie | Sam. | Z.K. |
|---|---|---|---|---|---|
| S | 800 | 32 | wyższe | tak | tak |
| S | 1200 | 35 | średnie | tak | tak |
| S | 700 | 26 | podstawowe | nie | nie |
| M | 600 | 45 | wyższe | nie | tak |
| M | 650 | 38 | średnie | tak | tak |
| S | 900 | 28 | wyższe | nie | nie |
| S | 1100 | 65 | średnie | tak | nie |
| M | 500 | 22 | średnie | nie | nie |
| S | 800 | 43 | podstawowe | tak | nie |

**`if Sam.=tak and S.C.=S then Z.K.=?`**

**Pokrycie: *s* = 3** **Poprawność: *a* = 2 / 3**

GDAŃSK UNIVERSITY OF TECHNOLOGY

# Rozbudowa reguły (4)

Po zbadaniu testów otrzymujemy:

| | *a* | *s* |
|---|---|---|
| **S.C.=S** | **2/3** | **3** |
| **S.C.=M** | **1/1** | **1** |
| **Wykształcenie=średnie** | **2/3** | **3** |
| **Wykształcenie=wyższe** | **1/1** | **1** |

Moglibyśmy na tym poprzestać, wybierając test o największej poprawności, ale tutaj włączymy do rozważań atrybuty numeryczne

GDAŃSK UNIVERSITY OF TECHNOLOGY

# Obsługa atrybutów numerycznych

Testy dla atrybutów numerycznych mogą być wyznaczane na różne sposoby – również omówioną metodą podziału binarnego minimalizującego entropię:

| | | | | |
|---|---|---|---|---|
| **Wiek:** | **32** | **35** | **38** | **65** |
| **Z.K.:** | **tak** | **tak** | **tak** | **nie** |

Co daje nam dwa dodatkowe testy do rozpatrzenia:

**`Wiek<=45`**
**`Wiek>45`**

GDAŃSK UNIVERSITY OF TECHNOLOGY

# Rola pokrycia

- Tym razem otrzymujemy wynik:

| | *a* | *s* |
|---|---|---|
| S.C.=S | 2/3 | 3 |
| S.C.=M | 1/1 | 1 |
| Wykształcenie=średnie | 2/3 | 3 |
| Wykształcenie=wyższe | 1/1 | 1 |
| Wiek<=45 | 3/3 | 3 |
| Wiek>45 | 1/1 | 1 |
| D.O.R.<=900 | 2/2 | 2 |
| D.O.R.>900 | 1/2 | 2 |

Wybieramy test o największej poprawności i największym pokryciu

GDAŃSK UNIVERSITY OF TECHNOLOGY

# Kolejne kroki

- Ponieważ osiągnięta poprawność wynosi 100%, druga reguła jest gotowa:

```
if Sam.=tak and Wiek<=45 then Z.K.=tak
```

- Cały proces jest powtarzany (następna iteracja) po kolejnym zastosowaniu zasady separate-and-conquer

GDAŃSK UNIVERSITY OF TECHNOLOGY

# Kolejne kroki (2)

| S.C. | D.O.R. | Wiek | Wykształcenie | Sam. | Z.K. |
|---|---|---|---|---|---|
| S | 800 | 32 | wyższe | tak | tak |
| S | 1200 | 35 | średnie | tak | tak |
| S | 700 | 26 | podstawowe | nie | nie |
| M | 600 | 45 | wyższe | nie | tak |
| M | 650 | 38 | średnie | tak | tak |
| S | 900 | 28 | wyższe | nie | nie |
| S | 1100 | 65 | średnie | tak | nie |
| M | 500 | 22 | średnie | nie | nie |
| S | 800 | 43 | podstawowe | tak | nie |

```
if Wykształcenie=podstawowe then Z.K.=nie
if Sam.=tak and Wiek<=45 then Z.K.=tak
if S.C.=S then Z.K.=nie
if D.O.R.<=500 then Z.K.=nie else Z.K.=tak
```

GDAŃSK UNIVERSITY OF TECHNOLOGY

# Budowa reguł – komentarz

- Zastosowanie zasady separate-and-conquer sprawia, że wygenerowane reguły muszą być rozpatrywane łącznie i w kolejności,
- Metoda w sposób naturalny radzi sobie z brakującymi wartościami atrybutów (są one „spychane” do następnej iteracji),
- Metodę można rozszerzyć o pewne mechanizmy uogólniające, przycinające reguły według pewnych zależności statystycznych

GDAŃSK UNIVERSITY OF TECHNOLOGY

# Budowa reguł – podsumowanie

## Algorytm budowy reguł decyzyjnych:

1. Zadanie: predykcja (klasyfikacja)
2. Struktura modelu: reguły rozpatrywane w kolejności
3. Funkcja oceny jakości: poprawność (1) i pokrycie (2)
4. Metody przeszukiwania: zachłanna, separate-and-conquer
5. Dodatkowe założenia:

   Obsługa atrybutów numerycznych metodą podziału binarnego minimalizującego entropię

GDAŃSK UNIVERSITY OF TECHNOLOGY

# Dziękujemy za uwagę

Zapraszamy na wykład:

KLASYFIKACJA I REGRESJA cz. 3