

Metody klasyfikacji

Opracowano na podstawie:

Morzy T. i inni: Eksploracja danych.

http://wazniak.mimuw.edu.pl/index.php?title=Eksploracja_danych

Dane wejściowe:

Treningowy zbiór krotek (przykładów, obserwacji, próbek), będących listą wartości atrybutów opisowych (tzw. deskryptorów) i wybranego *atrybutu decyzyjnego*

Dane wyjściowe:

Model (klasyfikator), przydziela każdej krotce wartość atrybutu decyzyjnego w oparciu o wartości pozostałych atrybutów (deskryptorów)

Klasyfikacja danych jest dwu-etapowym procesem:

- Etap 1:

Budowa modelu (klasyfikatora) opisującego predefiniowany zbiór klas danych lub zbiór pojęć

- Etap 2:

Zastosowanie opracowanego modelu do klasyfikacji nowych danych

I etap: Uczenie

Dane treningowe

Wiek_kierowcy	Typ samochodu	Ryzyko
18	Bus	Duże
20	Sport	Duże
35	Combi	Małe
32	Sedan	Małe
40	Sport	Małe
60	Combi	Małe
25	Sport	Duże
30	Bus	Małe

Atrybut decyzyjny

Reguły decyzyjne:

If Wiek <=35 or Samochód = „Sport” then Ryzyko = „Duże”

II etap: Testowanie

Wiek_kierowcy	Typ samochodu	Ryzyko
28	Bus	Duże
31	Sport	Duże
44	Combi	Małe
65	Sedan	Małe
57	Bus	Duże

Wynik klasyfikacji

Dokładność: $4/5 = 80\%$

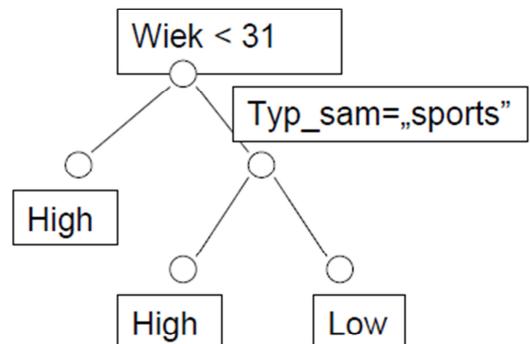
Ryzyko
Duże
Duże
Małe
Małe
Małe

III etap: Predykcja

Wiek_kierowcy	Typ samochodu	Ryzyko
63	Bus	?
34	Sport	?
22	Combi	?
80	Sedan	?
55	Bus	?

If Wiek <=35 or Samochód = „Sport” then Ryzyko = „Duże”

Wiek	Typ_sam	Ryzyko
20	Combi	High
18	Sports	High
40	Sports	High
50	Family	Low
35	Minivan	Low
30	Combi	High
32	Family	Low
40	Combi	Low



Tworzenie klasyfikatora

Dla danych trenigowych jak wyznaczyć reguły klasyfikacji w postaci drzewa decyzyjnego.

Problem-1:

Dla danego atrybutu deskryptora jak wyznaczyć punkty podziału?

Problem-2:

Jak wybrać pierwszy i kolejny atrybut do klasyfikatora?

Problem-1:

Dla danego atrybutu deskryptora jak wyznaczyć punkty podziału?

Metoda:

- Zbudować drzewo decyzyjne (algorytm SPRINT) z użyciem indeksu Gini

```
Partition(Data S) {  
    if (all points in S are of the same class) then  
        return;  
    for each attribute A do  
        evaluate splits on attribute A;  
        Use best split found to partition S  
        into S1 and S2  
        Partition(S1);  
        Partition(S2);  
}  
Initial call: Partition(Training Data)
```

Typy atrybutów:

- Numeryczny
- Kategoryczny-wyliczeniowy
- Kategoryczny-niewyliczeniowy

- **Definicja:**

gdzie:

$$\text{gini}(S) = 1 - \sum p_j^2$$

- **S** – zbiór przykładów należących do n klas
- p_j – względna częstość występowania klasy j w S
- Przykładowo:
dwie klasy, Pos i Neg, oraz zbiór przykładów S zawierający p elementów należących do klasy Pos i n elementów należących do klasy Neg

$$p_{\text{pos}} = p/(p+n) \quad p_{\text{neg}} = n/(n+p)$$

$$\text{gini}(S) = 1 - p_{\text{pos}}^2 - p_{\text{neg}}^2$$

- Punkt podziału dzieli zbiór S na dwie partie S_1 i S_2 – indeks podziału Gini jest zdefiniowany następująco:

$$\text{gini}_{\text{SPLIT}}(S) = (p_1 + n_1)/(p+n) * \text{gini}(S_1) + \\ (p_2 + n_2)/(p+n) * \text{gini}(S_2)$$

gdzie p_1, n_1 (p_2, n_2) oznaczają, odpowiednio,

- p_1 - elementów w S_1 należących do klasy Pos,
- n_1 - liczba elementów w S_1 należących do klasy Neg,
- p_2 - elementów w S_2 należących do klasy Pos,
- n_2 - liczba elementów w S_2 należących do klasy Neg

Sposób wyboru punktu podziału:

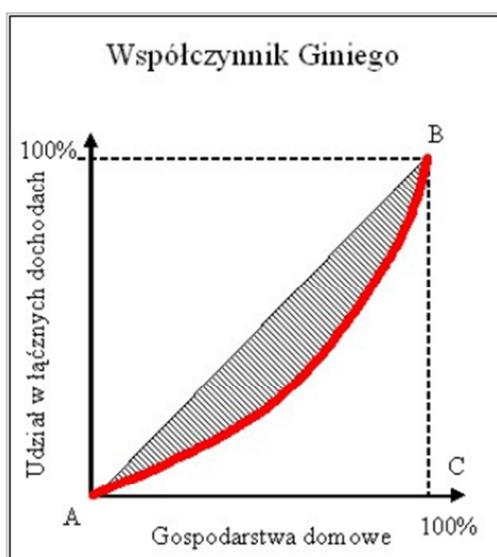
- „Najlepszym” punktem podziału zbioru S jest punkt podziału, który charakteryzuje się najmniejszą wartością indeksu podziału Gini $gini_{SPLIT}$
- Dla każdego atrybutu, dla wszystkich możliwych punktów podziału, oblicz wartość indeksu podziału Gini – wybierz punkt podziału o najmniejszej wartości $gini_{SPLIT}$
- Wybrany punkt podziału włącz do drzewa decyzyjnego
- Punkt podziału dzieli zbiór S na dwie partycje S_1 i S_2 .
- Powtórz procedurę obliczania indeksu podziału dla partycji S_1 i S_2 – znalezione punkty podziału włącz do drzewa decyzyjnego.
- Powtarzaj procedurę dla kolejnych partycji aż do osiągnięcia warunku stopu

WSPÓŁCZYNNIK GINIEGO



Aby zmierzyć stan nierówności dochodowych, wykorzystuje się zazwyczaj tzw. współczynnik Giniego, którego wartość zawiera się w przedziale od 0 do 1, gdzie wyższa wartość współczynnika oznacza większą skalę nierówności. Idea wyliczenia współczynnika Giniego jest dość prosta i może być łatwo przedstawiona na rysunku.

Na wykresie odkładamy krzywą, obrazującą rozkład dochodów (udział w łącznych dochodach skumulowanych grup gospodarstw domowych, liczących od najuboższego do najbogatszego). Krzywa ta, zaznaczona czerwoną linią, zaczyna się w punkcie A w środku wykresu (0% gospodarstw domowych ma 0% dochodu) i dochodzi do punktu B przecięcia linii kreskowanych (100% gospodarstw domowych ma 100% dochodu). Im bardziej równomierny rozkład dochodów w gospodarce, tym krzywa ta byłaby bliższa prostej łączącej oba te punkty. Zakreskowane pole pokazuje więc, jak duża jest skala nierówności dochodowych.



Odnosząc powierzchnię zakreskowanego pola do całego trójkąta ABC, uzyskujemy wartość współczynnika Giniego: równą 0 w przypadku kraju o idealnej równości dochodów (w praktyce taką sytuację nigdy nie występuje), a rosnącą do 1 w przypadku kraju o skrajnych nierównościach dochodowych (również w praktyce taką wartość jest niemożliwa). We współczesnych gospodarkach współczynnik Giniego waha się od 0.25 (kraje skandynawskie) do 0.65 (kraje Ameryki Łacińskiej).

(źródło: NBP)

Id	Wiek_kierowcy	Typ samochodu	Ryzyko
1	28	Bus	Duże
2	31	Sport	Duże
3	44	Combi	Małe
4	65	Sedan	Małe
5	57	Bus	Duże

Możliwe przedziały:

- Wiek ≤ 28
- Wiek ≤ 31
- Wiek ≤ 44
- Wiek ≤ 57
- Wiek ≤ 65

Liczba krotek	Duże	Małe
Wiek ≤ 28	1	0
Wiek > 28	2	2

$$\text{Gini}(\text{Wiek} \leq 28) = 1 - (1^2 + 0^2) = 0$$

$$\text{Gini}(\text{Wiek} > 28) = 1 - ((2/4)^2 + (2/4)^2) = 1/2$$

$$\text{Gini}_{\text{SPLIT}} = (1/5)*0 + (4/5)*1/2 = \mathbf{2/5}$$

$$\text{Gini}(\text{Wiek} \leq 31) = 1 - (1^2 + 0^2) = 0$$

$$\text{Gini}(\text{Wiek} > 31) = 1 - ((1/3)^2 + (2/3)^2) = 4/9$$

$$\text{Gini}_{\text{SPLIT}} = (2/5)*0 + (3/5)*4/9 = \mathbf{4/15}$$

$$\text{Gini}(\text{Wiek} \leq 44) = 1 - ((2/3)^2 + (1/3)^2) = 4/9$$

$$\text{Gini}(\text{Wiek} > 44) = 1 - ((1/2)^2 + (1/2)^2) = 1/2$$

$$\text{Gini}_{\text{SPLIT}} = (3/5)*(4/9) + (2/5)*(1/2) = \mathbf{7/15}$$

$$\text{Gini}(\text{Wiek} \leq 57) = 1 - ((3/4)^2 + (1/4)^2) = 3/8$$

$$\text{Gini}(\text{Wiek} > 57) = 1 - (0^2 + (1/1)^2) = 0$$

$$\text{Gini}_{\text{SPLIT}} = (4/5)*(3/8) + (1/5)*0 = \textcolor{red}{3/10}$$

$$\text{Gini}(\text{Wiek} \leq 65) = 1 - ((3/5)^2 + (2/5)^2) = 12/25$$

$$\text{Gini}(\text{Wiek} > 65) = 1 - (0^2 + 0^2) = 1$$

$$\text{Gini}_{\text{SPLIT}} = (5/5)*(12/25) + 0*1 = \textcolor{red}{12/15}$$

Punkt podziału: **37.5**

Dla Wiek <= 37.5

Id	Wiek	Typ	Ryzyko
1	28	Bus	Duże
2	31	Sport	Duże

Dla Wiek > 37.5:

Id	Wiek	Typ	Ryzyko
3	44	Combi	Małe
4	65	Sedan	Małe
5	57	Bus	Duże

Typ	Id	Ryzyko
Combi	3	Małe
Sedan	4	Małe
Bus	5	Duże

$$\text{Gini}(\text{Typ} \in \{\text{Combi}\}) = 1 - (1^2 + 0^2) = 0$$

$$\text{Gini}(\text{Typ} \in \{\text{Sedan}\}) = 1 - (1^2 + 0^2) = 0$$

$$\text{Gini}(\text{Typ} \in \{\text{Bus}\}) = 1 - (0^2 + 1^2) = 0$$

$$\text{Gini}(\text{Typ} \in \{\text{Combi, Sedan}\}) = 1 - (1^2 + 0^2) = 0$$

$$\text{Gini}(\text{Typ} \in \{\text{Combi, Bus}\}) = 1 - ((1/2)^2 + (1/2)^2) = \frac{1}{2}$$

$$\text{Gini}(\text{Typ} \in \{\text{Bus, Sedan}\}) = 1 - ((1/2)^2 + (1/2)^2) = \frac{1}{2}$$

$$\text{Gini}_{\text{SPLIT}}(\text{Typ} \in \{\text{Combi}\}) = (1/3)*0 + (2/3)*(1/2) = 1/3$$

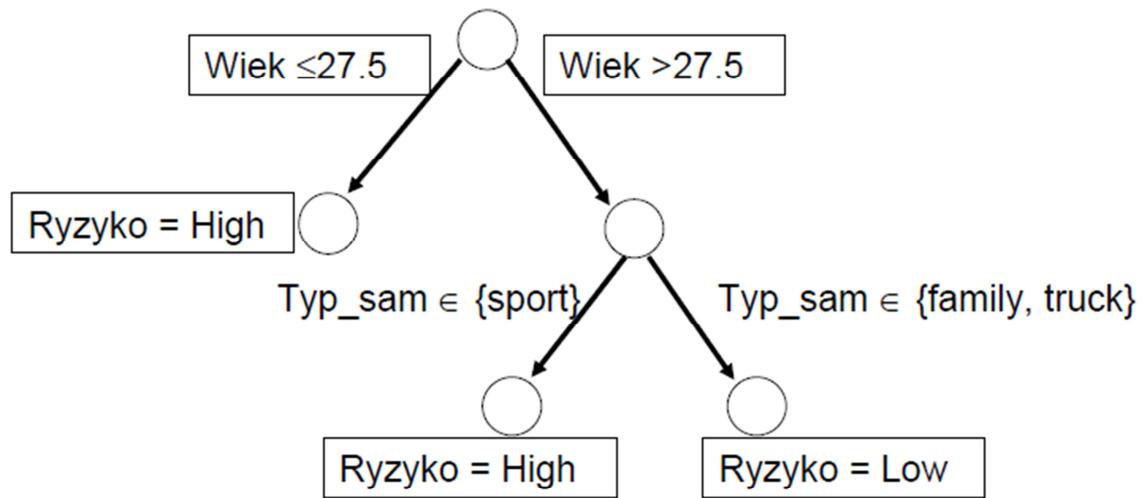
$$\text{Gini}_{\text{SPLIT}}(\text{Typ} \in \{\text{Sedan}\}) = (1/3)*0 + (2/3)*(1/2) = 1/3$$

$$\text{Gini}_{\text{SPLIT}}(\text{Typ} \in \{\text{Bus}\}) = (1/3)*0 + (2/3)*0 = 0$$

$$\text{Gini}_{\text{SPLIT}}(\text{Typ} \in \{\text{Combi, Sedan}\}) = (2/3)*0 + (2/3)*0 = 0$$

$$\text{Gini}_{\text{SPLIT}}(\text{Typ} \in \{\text{Combi, Bus}\}) = (2/3)*(1/2) + (1/3)*0 = 1/3$$

Klasyfikator:



Problem-2:

Jak wybrać pierwszy i kolejny atrybut do klasyfikatora?

Przykład: Dane treningowe na temat zakupu komputerów
(z książki J Hana)

ID	wiek	dochód	student	status	kupi_komputer
1	<=30	wysoki	nie	kawaler	nie
2	<=30	wysoki	nie	żonaty	nie
3	31..40	wysoki	nie	kawaler	tak
4	>40	średni	nie	kawaler	tak
5	>40	niski	tak	kawaler	tak
6	>40	niski	tak	żonaty	nie
7	31..40	niski	tak	żonaty	tak
8	<=30	średni	nie	kawaler	nie
9	<=30	niski	tak	kawaler	tak
10	>40	średni	tak	kawaler	tak
11	<=30	średni	tak	żonaty	tak
12	31..40	średni	nie	żonaty	tak
13	31..40	wysoki	tak	kawaler	tak
14	>40	średni	nie	żonaty	nie

Procedura:

- Do wyboru atrybutu testowego w wierzchołku drzewa decyzyjnego wykorzystujemy miarę **zysku informacyjnego**
- Jako atrybut testowy (aktualny wierzchołek drzewa decyzyjnego) wybieramy atrybut o największym zysku informacyjnym (lub największej redukcji entropii)
- Atrybut testowy minimalizuje ilość informacji niezbędnej do klasyfikacji przykładów w partycjach uzyskanych w wyniku podziału
- Niech S oznacza zbiór s przykładów. Założymy, że atrybut decyzyjny posiada m różnych wartości definiujących m klas, C_i (dla $i=1, \dots, m$)
- Niech s_i oznacza liczbę przykładów zbioru S należących do klasy C_i
- Oczekiwana ilość informacji niezbędna do zaklasyfikowania danego przykładu:

$$I(s_1, \dots, s_m) = - \sum_{i=1}^m p_i \log_2(p_i)$$

- Niech s_{ij} oznacza liczbę przykładów z klasy C_i w partycji S_j . Entropię podziału zbioru S na partycje, według atrybutu A definiujemy następująco:

$$E(A_1, A_2, \dots, A_v) = \sum_{j=1}^v \frac{(s_{1j} + s_{2j} + \dots + s_{mj})}{S} I(s_{1j}, s_{2j}, \dots, s_{mj})$$

Im mniejsza wartość entropii, tym większa „czystość” podziału zbioru S na partycje

Sposób obliczenia :

- Współczynnik $(s_{1j} + s_{2j} + \dots + s_{mj})/s$ stanowi wagę j-tej partycji i zdefiniowany jest jako iloraz liczby przykładów w j-tej partycji (i.e. krotek posiadających wartość a_j atrybutu A) do całkowitej liczby przykładów w zbiorze S. Zauważmy, że dla danej partycji S_j ,

$$I(s_{1j}, s_{2j}, \dots, s_{mj}) = -\sum_{i=1}^m p_{ij} \log_2(p_{ij})$$

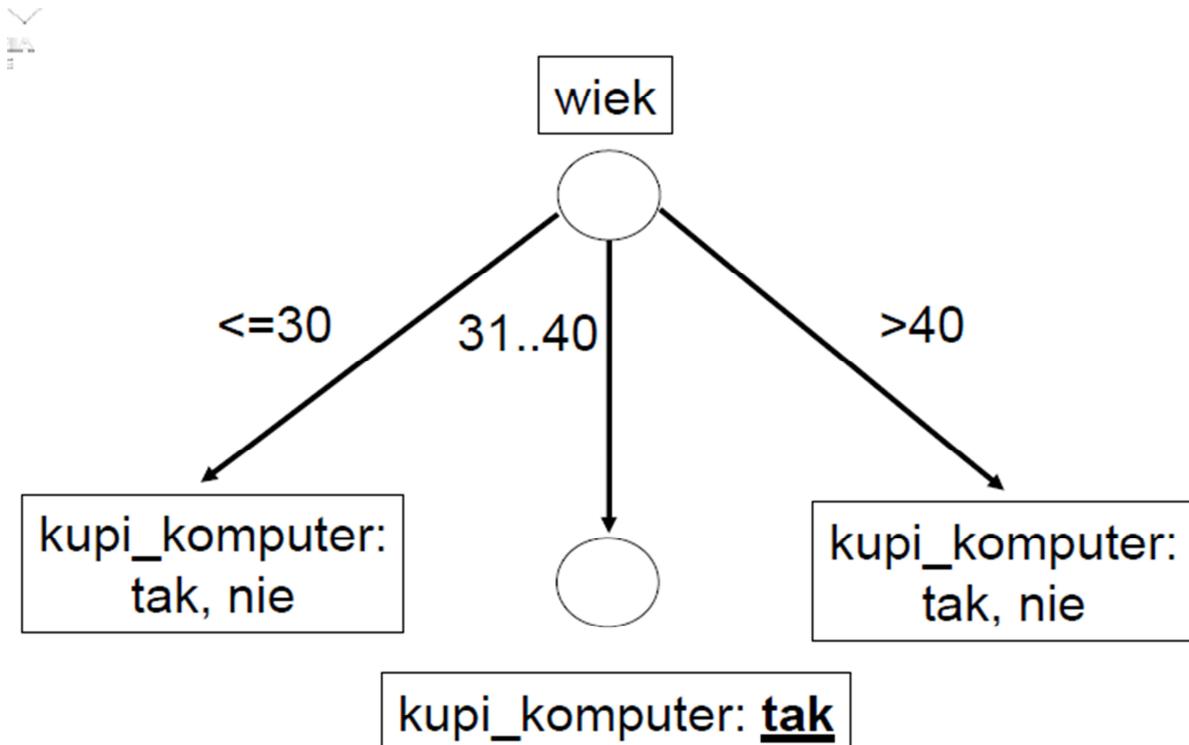
gdzie $p_{ij} = s_{ij}/|S_j|$ i określa prawdopodobieństwo, że przykład z S_j należy do klasy C_i

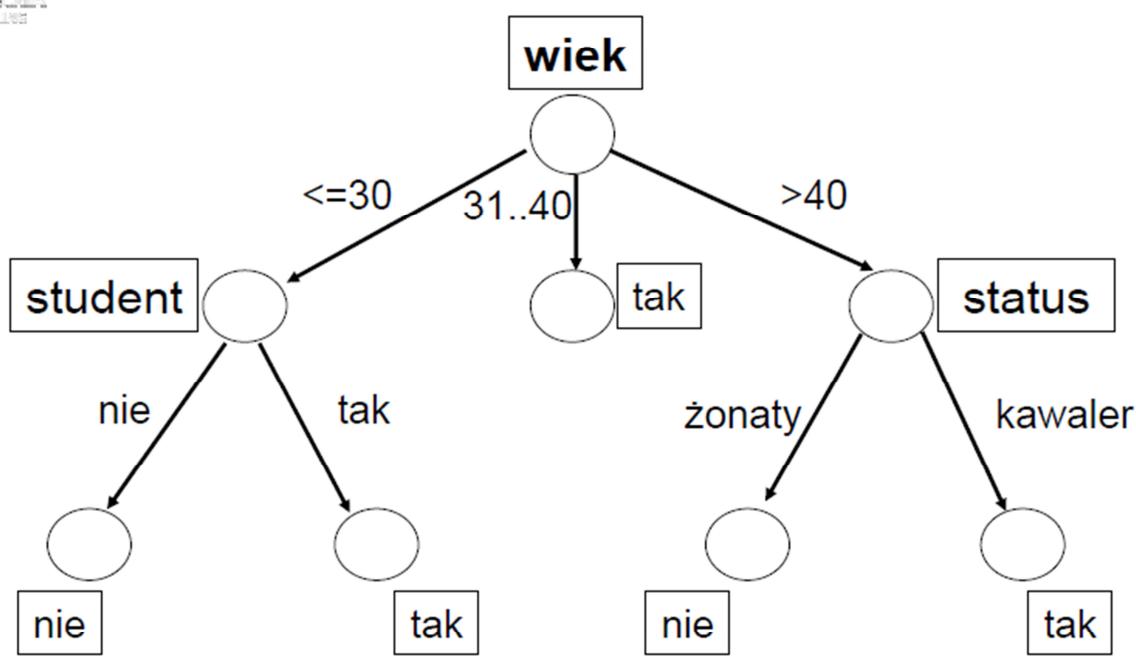
- Zysk informacyjny, wynikający z podziału zbioru S na partycje według atrybutu A, definiujemy następująco:

$$Gain(A) = I(s_1, s_2, \dots, s_m) - E(A)$$

- Gain(A) oznacza oczekiwana redukcję entropii (nieuporządkowania) spowodowaną znajomością wartości atrybutu A

Po obliczeniu, atrybut Wiek jest najlepszy, więc jest pierwszy:





Ostateczna postać drzewa decyzyjnego