dr inż. Adam Deptuła
prof. dr hab. Marian A. Partyka
Politechnika Opolska, Wydział Inżynierii Produkcji i Logistyki
E-mail: a.deptula@po.opole.pl

Znaczenie kolejności atrybutów dla zbiorów przykładów klasyfikowanych z wykorzystaniem indukcyjnych drzew decyzyjnych

W indukcyjnym drzewie decyzyjnym węzły przechowują testy sprawdzające wartości atrybutów przykładów, a liście przypisywane im kategorie. Dla każdego z możliwych wyników testu z węzła prowadzi odpowiadająca mu gałąź do pewnego poddrzewa. W ten sposób mogą być reprezentowane dowolne dopuszczalne dla danego zestawu atrybutu hipotezy [1].

W klasyfikacji z wykorzystaniem drzew indukcyjnych zakłada się, że dana jest dziedzina X, na której są określone atrybuty $a_1, a_2, \ldots, a_n$, klasa pojęć C o zbiorze kategorii C oraz:

1. Liść zawierający dowolną etykietę kategorii $d \in C$ jest drzewem decyzyjnym;

2. $t : X \to R_t$ jest testem przeprowadzanym na wartościach atrybutów przykładów o zbiorze możliwych wyników $R_t = \{r_1, r_2, \ldots, r_m\}$ [1, 2, 3].

Każdemu ze skończonej liczby możliwych wyników testu odpowiada gałąź prowadząca z węzła do poddrzewa. Jeśli węzeł zawiera test t o zbiorze wyników $R_t = \{r_1, r_2, \ldots, r_m\}$, a odpowiadające im gałęzie prowadzą do poddrzew $T_1, T_2, \ldots, T_m$, to hipotezę reprezentowaną przez ten węzeł można dla każdego przykładu $x \in X$ zapisać następująco [1, 2]:

$$h(x) = \begin{cases} h_1(x), & \text{jeśli } t(x) = r_1, \\ h_2(x), & \text{jeśli } t(x) = r_2, \\ \ldots\ldots & \\ h_m(x), & \text{jeśli } t(x) = r_m. \end{cases} \tag{1}$$

przy czym $h_1, h_2, \ldots, h_m$ są odpowiednio hipotezami reprezentowanymi przez drzewa $T_1, T_2, \ldots, T_m$.

Dla dowolnego zbioru przykładów P i testu t odwołujemy się do podzbiorów, na jakie zbiór P jest dzielony ze względu na różne wyniki testu t. Oznacza to, że ważna jest kolejność deklaracji atrybutów zmiennych decyzyjnych.

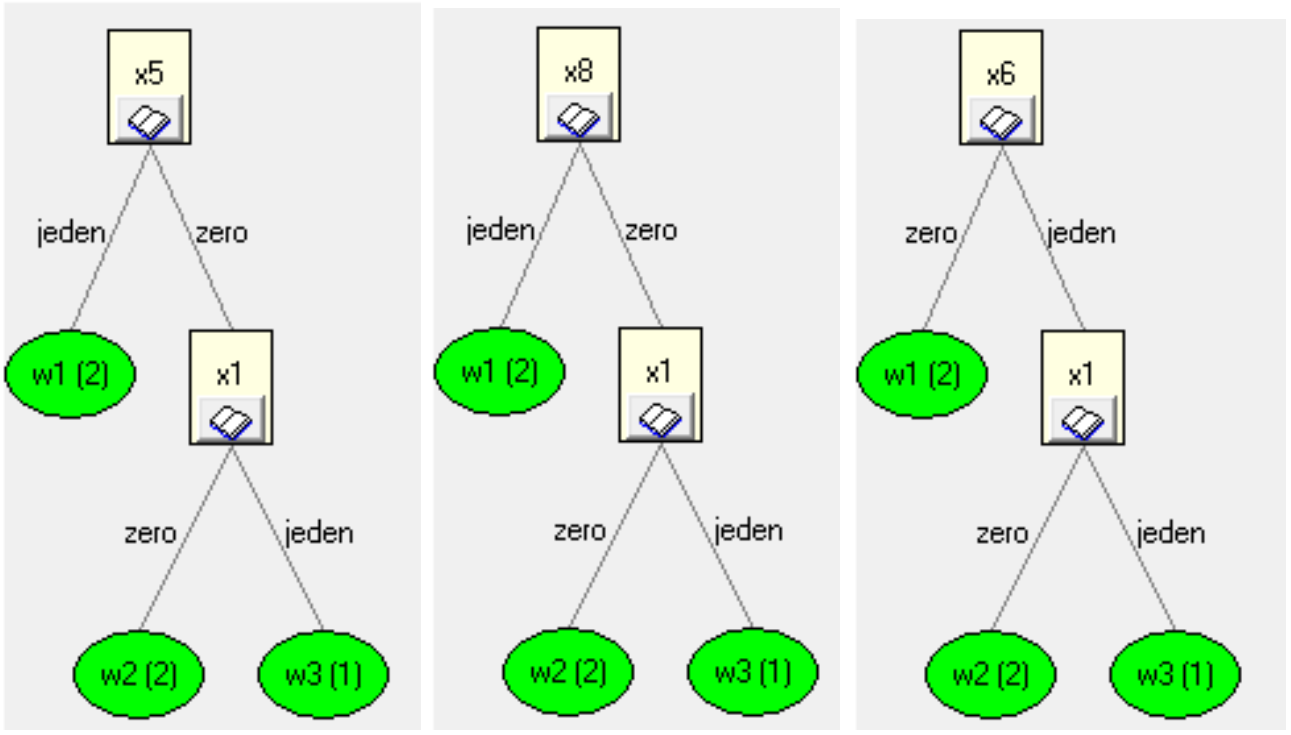
Przykład. Klasyfikację pacjentów z punktu widzenia choroby tarczycy opisano atrybutowo i logicznie w sensie zerojedynkowym [4, 5, 6], zakładając, że istnieje 8 atrybutów (objawów) wejściowych (we), wg kolejności: $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8$, dla wyjściowych wskaźników jakościowych (wy): (t_1, t_2) — nie ma choroby tarczycy; (t_4, t_5) — istnieje na pewno choroba tarczycy; t_3 — nie można stwierdzić ani wykluczyć z całą pewnością choroby tarczycy (Tab. 1).

Tab. 1

wy; we	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
t_1	0	1	0	0	1	0	0	1
t_2	0	0	0	1	1	0	0	1
t_4	0	1	0	1	0	1	1	0
t_3	1	1	0	1	0	1	1	0
t_5	0	1	1	1	0	1	1	0

Indukcyjny system przybliżony DeTreex [4] na podstawie obliczeń entropii sklasyfikował rangę ważności zmiennych decyzyjnych na drzewie indukcyjnym jako $x_5 - x_1$ (Rys. 1a), pomijając zmienną x_8, o takiej samej randze ważności jak x_5. Dla kolejności zmiennych decyzyjnych: $x_1, x_2, x_3, x_4, x_8, x_6, x_7, x_5$ istnieje drzewo indukcyjne o układzie pięter $x_8 - x_1$ (Rys.1b), a dla kolejności zmiennych decyzyjnych: $x_1, x_2, x_3, x_4, x_6, x_7, x_8, x_5$ istnieje drzewo indukcyjne o układzie pięter $x_6 - x_1$ (Rys. 1c).

a) b) c)



Rys. 1. Indukcyjne drzewa decyzyjne o układach pięter: a) $x_5 - x_1$, b) $x_8 - x_1$, c) $x_6 - x_1$.

W omawianym przykładzie zamiana kolejności zmiennych decyzyjnych mało ważnych x_1, x_2, x_3, x_4 nie ma znaczenia dla procesu klasyfikacji, pod warunkiem wcześniejszego prawidłowego wyznaczenia rangi ważności za pomocą wielowartościowych logicznych drzew decyzyjnych. Drzewo indukcyjne zawiera tylko pierwszy sklasyfikowany atrybut, pomimo istnienia w zbiorze przykładów większej liczby atrybutów o takich samych wynikach testu t, tzn. takiej samej randze ważności.

Dodatkowo należy zaznaczyć, że system przybliżony DeTreex [4] nie rozpoznał najmniej ważnych atrybutów x_1 oraz x_3 w przeciwieństwie do decyzyjnych dwuwartościowych drzew logicznych [6]. Wynika to z faktu, że atrybuty x_1, x_3 są klasyfikowane jedynie jako trochę mniej ważne wobec najważniejszych x_5 [4], x_8, x_6, ... według systemu DeTreex.

Literatura

[1] J. R. Quinlan, *Induction of Decision Trees*, Machine Learning 1 (1986), 81–100.

[2] A. Deptuła, *Indukcyjne drzewa decyzyjne* (*entropia*) *jako odpowiednik zmodyfikowanych drzew logicznych w wyznaczaniu rangi ważności zmiennych decyzyjnych projektowanego układu*, XLIII Konf. Zast. Mat., Zakopane 2014, Inst. Mat. PAN, Warszawa 2014.

[3] A. Deptuła, M. A. Partyka, *Application of dependence graphs and game trees for decision decomposition for machine systems*, Journal of Automation, Mobile Robotics & Intelligent Systems 5 (2011), No. 3, 17–26.

[4] A. Deptuła, M. A. Partyka, *Analiza porównawcza klasyfikacyjnych metod informacyjnych*, XLIV Konf. Zast. Mat., Zakopane 2015, Inst. Mat. PAN, Warszawa 2015.

[5] Z. Pawlak, *Systemy informacyjne*, WNT, Warszawa 1983.

[6] M. A. Partyka, *Podobieństwa i różnice przybliżonej klasyfikacji obiektów w ujęciu logiki i systemów informacyjnych dla CAD procesów decyzyjnych*, XXIV Konf. Zast. Matem., Zakopane 1995, Inst. Mat. PAN, Warszawa 1995.