

dr inż. Adam Deptuła

prof. dr hab. Marian A. Partyka

Politechnika Opolska, Wydział Inżynierii Produkcji i Logistyki

E-mail: a.deptula@po.opole.pl

## Analiza porównawcza klasyfikacyjnych metod informacyjnych

Przybliżona klasyfikacja obiektów w ujęciu systemów informacyjnych wykorzystuje m.in. pojęcia atrybutu oraz termu. Takie podejście jest analogiczne do zmiennej logicznej i wyrażenia logicznego typu iloczyn elementarny, o ile atrybut przyjmuje jedną z wartości 0, 1. Dla  $n$  atrybutów istnieje  $2^n$  termów elementarnych (w sensie teoretycznym), lecz tylko niektóre można uznać jako tzw. zbiory elementarne danego systemu [3, 4].

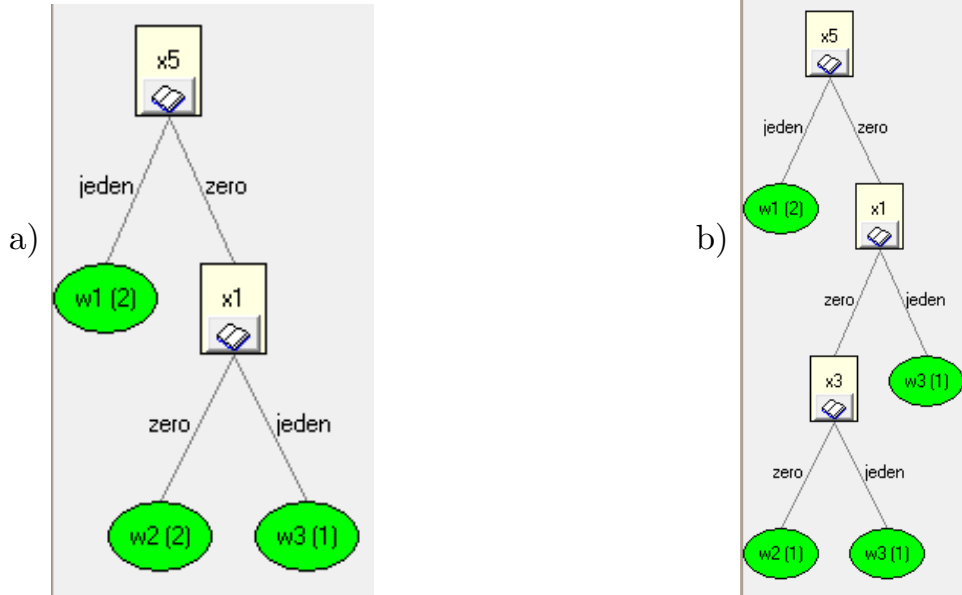
Systemy informacyjne korzystają z pojęć klasyfikacyjnych typu przybliżenie dolne i przybliżenie górne. Wykorzystanie analogiczne w logice nie jest możliwe, ale można wprowadzić klasyfikację formalną z uwzględnieniem maksimum upraszczania typu:  $Ax + A\bar{x} = A$  przy jednoczesnym zachowaniu ciągłości decyzyjnej struktury dendrytowo-drzewiastej, co oznacza m.in. brak gałęzi izolowanych. W ten sposób zmienna najbardziej upraszczająca się jest najmniej ważna w sensie informacyjnym i dlatego powinna być na najwyższym piętrze hierarchicznym drzewa. Upraszczenie pozostałych zmiennych wykonuje się wg tej samej zasady dla drzew o zmniejszającej się stopniowo liczbie pięter hierarchicznych, ale przy jednoczesnym założeniu, że takie upraszczenie jest możliwe tylko dla tych zmiennych i iloczynów, które powstały z wcześniejszego upraszczania [3].

**Przykład.** Gdyby klasyfikację pacjentów z punktu widzenia choroby tarczycy opisać atrybutowo i logicznie w sensie zerojedynkowym przy założeniu, że istnieje 8 atrybutów (objawów), to otrzymuje się 5 iloczynów logicznych dla 8 zmiennych:  $t_1, t_2$  — 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. W sensie systemów informacyjnych podane objawy dają: przybliżenie dolne zbioru chorych  $t_4 + t_5$  oraz górne  $t_3 + t_4 + t_5$ ; przybliżenie dolne ludzi zdrowych  $t_1 + t_2$  oraz górne  $t_1 + t_2 + t_3$ . Można ponadto stwierdzić, że dla grup  $t_1 + t_2$  oraz  $t_4 + t_5$  nie jest potrzebny atrybut pierwszy (przyjmujący tylko wartość 0). W przypadku klasyfikacji zbioru pacjentów w ujęciu logiki, dla danych 5 iloczynów elementarnych otrzymuje się wynik: atrybut  $x_1$  najmniej ważny, a potem (niżej) atrybut  $x_3$  lub na odwrót, co wynika z następujących obliczeń:

$x_1 : 5 - 1 \cdot 2 + 3 = 6$	$t_1$	0	1	0	0	1	0	0	1
$x_3 : 4 - 0 \cdot 2 + 3 = 7$	$t_2$	0	0	0	1	1	0	0	1
$x_2, x_4, x_5, x_6, x_7, x_8 : 4 - 0 \cdot 2 + 0 + 4 = 8$	$t_4$	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

Przy założeniu różnych wyjściowych wskaźników jakościowych dla  $(t_1 + t_2)$ ,  $(t_3)$ ,  $(t_4 + t_5)$ , indukcyjny system przybliżony DeTreex [5] na podstawie obliczeń entropii

odmiennie sklasyfikował rangę ważności zmiennych decyzyjnych. Inne systemy informatyczne oraz informacyjne nie wykorzystują dosłownie zagadnień logiki formalnej dla rangi ważności zmiennych decyzyjnych [1, 2] (rys. 1).



Rys. 1. Indukcyjne drzewa decyzyjne dla termów wyjściowych (wy):

a)  $w1, w1, w2, w3w2$ ; b)  $w1, w1, w2, w3w3$

### Literatura

- [1] Z. Banaszak, S. Kłos, J. Mleczko, *Zintegrowane systemy zarządzania*, PWE, Warszawa 2011.
- [2] W. Bojar, K. Rostek, L. Knopik, *Systemy wspomaganie decyzji*, PWE, Warszawa 2014.
- [3] M. A. Partyka, *Podobieństwa i różnice przybliżonej klasyfikacji obiektów w ujęciu logiki i systemów informacyjnych dla CAD procesów decyzyjnych*, XXIV Konfer. Zastos. Matem. PAN, Zakopane 1995.
- [4] Z. Pawlak, *Systemy informacyjne*, WNT, Warszawa 1983.
- [5] J. R. Quinlan, *Improved use of continuous attributes in C4.5*, Journal of Artificial Intelligence Research 4 (1996), 77–90.