

**Mariusz Kubus**  
Politechnika Opolska

## **ZASTOSOWANIE FUNKCJI LOGICZNYCH DO PORZĄDKOWANIA CECH JAKOŚCIOWYCH**

### **1. Wstęp**

Jednym z celów porządkowania cech jest redukcja opisu. W przypadku cech jakościowych można w tym celu wykorzystać drzewa logiczne, które są graficzną ilustracją skończonej przestrzeni kombinacji wartości cech. Minimalne drzewo logiczne wyznacza porządek zmiennych ze względu na ogólność opisu zbioru obiektów. Tak otrzymane uporządkowania wyznaczają kierunek przeszukiwania przestrzeni opisów klas w klasyfikacji wzorcowej.

Metody konstruowania reguł klasyfikacji obiektów symbolicznych można podzielić na dwie grupy [Gatnar 1998]. Do pierwszej należą te, które znajdują charakterystyki klas na podstawie wartości cech należących do nich obiektów. Wymienić tu można najczęściej stosowane implementacje w postaci algorytmów z rodziny AQ [Michalski 1983] oraz CN2 [Clark, Niblett 1989]. Drugie podejście polega na rekurencyjnym podziale zbioru uczącego na coraz bardziej jednorodne podzbiory w sensie przynależności do klas. Rekurencyjny podział jest przedstawiany za pomocą drzew klasyfikacyjnych.

Celem referatu jest prezentacja metody porządkowania cech jakościowych za pomocą drzew logicznych oraz zastosowanie otrzymanych uporządkowań zmiennych do tworzenia charakterystyki klas. Zmienne ostatnie w uporządkowaniu mają mniejsze znaczenie w tworzeniu reguł klasyfikacyjnych i na ogół można je pomijać. Otrzymane wyniki zostaną porównane z powszechnie używanym algorytmem AQ11.

### **2. Porządkowanie cech jakościowych metodą drzew logicznych**

Macierz danych jakościowych może być traktowana jako zapis funkcji logicznej w postaci kanonicznej. Przedstawienie takie jest jednoznaczne. Każdą funkcję

logiczną można zilustrować na drzewie logicznym [Jaczewski 1978]. Poziomy drzewa (piętra hierarchiczne) reprezentują zmienne. Należy podkreślić, że każde piętro reprezentuje jedną zmienną, tzn. we wszystkich wierzchołkach jednego poziomu znajdują się testy wartości tej samej zmiennej. Z wierzchołków wewnętrznych wychodzą krawędzie kodowane wartościami zmiennych. Umownie przyjęto, że wartości kodowane 0, 1, 2, ... są przypisywane krawędziom od strony lewej do prawej. Liczba liści jest równa liczbie obiektów, a krawędzie na ścieżkach od korzenia do liści reprezentują odpowiednie współrzędne wektorów obserwacji.

Na drzewie logicznym dopuszcza się przestawienia piętrowe zmiennych, co oznacza jedynie zamianę kolejności kolumn w macierzy danych. Efektem takich przestawień jest na ogół zmienna liczba krawędzi drzewa. Przez minimalne drzewo logiczne rozumie się drzewo z najmniejszą liczbą krawędzi (a tym samym wierzchołków). Zagadnienie minimalizacji drzewa logicznego można zatem sformułować jako wyznaczenie takiego uporządkowania zmiennych, które dałoby drzewo z najmniejszą liczbą krawędzi. Ponieważ istnieje  $n!$  piętrowych rozmieszczeń zmiennych, metody konstrukcji minimalnego drzewa logicznego formułuje się na podstawie założeń heurystyczne [Kubus 2003]. Metody te wykorzystują strategię wspinaczki, dokonując w każdym kroku optymalnego wyboru zmiennej według ustalonego kryterium.

Liczbę krawędzi ( $B$ ) na drzewie logicznym określa się wzorem (1):

$$B = mn - (n-1)R_n - (n-2)R_{n-1} - (n-3)R_{n-2} - \dots - 2R_3 - R_2, \quad (1)$$

gdzie:  $m$  – liczba obiektów,

$n$  – liczba zmiennych,

$R_i$  – liczba rozgałęzień na  $i$ -tym piętrze drzewa.

Współczynniki mnożące liczby rozgałęzień maleją wraz z obniżaniem się poziomu drzewa. Obserwacja ta prowadzi do sformułowania metody minimalizacji drzewa logicznego, polegającej na maksymalizacji liczby rozgałęzień w kolejnych krokach od liści do korzenia. Zmienną o największej krotności tworzenia rozgałęzień ustala się na najwyższym poziomie drzewa i redukuje zagadnienie do  $n - 1$  zmiennych. Gdy wiele zmiennych daje maksymalne liczby rozgałęzień, można rozważać wiele wariantów lub losowo wybrać spośród nich zmienną optymalną. Całą procedurę wykonuje się do wyczerpania zestawu zmiennych. Liczby rozgałęzień wylicza się na podstawie wybranych metryk. Można wykorzystać metrykę miejską z współczynnikiem mnożącym (2) lub odległość syntaktyczną między obiektami dla cech nominalnych (3).

Niech  $x_1^{p_1} x_2^{p_2} \dots x_n^{p_n}$  będzie układem zmiennych o skończonych zbiorach wartości mocy  $p_i$  odpowiednio. Ponadto niech  $p = \max \{p_1, \dots, p_n\}$ .

$$T_{\frac{1}{p}}(w_i, w_j) = \sum_{l=1}^n \frac{1}{p^l} |w_{il} - w_{jl}|, \quad (2)$$

$$s(w_i, w_j) = \sum_{l=1}^n d(w_{il}, w_{jl}), \quad \text{gdzie } d(a, b) = \begin{cases} 1 & \text{dla } a \neq b \\ 0 & \text{dla } a = b \end{cases}. \quad (3)$$

Druga metoda minimalizacji drzew logicznych przyjmuje odwrotny kierunek porządkowania zmiennych – od korzenia do liści. Polega na ocenie kombinacji zmiennych (najpierw dwóch, później trzech itd.) według kryterium najmniejszej liczby krawędzi. Optymalne zestawy  $k$ -elementowe muszą zawierać optymalne zestawy  $k - 1$ -elementowe.

Obie metody korespondują z sobą ze względu na wzór (4), opisujący liczbę krawędzi  $i$ -tego piętra:

$$B_i = m - R_n - R_{n-1} - \dots - R_{i+1}. \quad (4)$$

Ponieważ liczba krawędzi na  $i$ -tym piętrze jest zależna od sumy rozgałęzień na wyższych piętrach drzewa, metody te często prowadzą do tych samych uporządkowań zmiennych. Niekiedy dają one jednak różne wyniki, dlatego jest uzasadnione poszukiwanie modyfikacji opisanych metod.

### 3. Przeszukiwanie przestrzeni opisu klas

Obiekty symboliczne (jakościowe) oraz klasy opisuje się za pomocą wyrażeń koniunkcyjnych, nazywanych w logice wielowartościowej  $VL_1$  (ang. *variable-valued logic*) Michalskiego kompleksami. Na przykład kompleks

[silnik = 1.9TD] ∧ [nadwozie = kombi] ∧ [kolor = biały]

opisuje wszystkie białe samochody kombi, które mają silnik 1.9TD. Pojedynczy składnik takiej koniunkcji jest nazywany selektorem. Selektor może także zawierać inny operator relacyjny np.  $\neq$ ,  $\geq$ ,  $<$ . Opis klasy jako alternatywa kompleksów jest nazywany opisem dysjunkcyjnym, a operację przydzielania obiektu spełniającego opis  $D$  do klasy  $K$  zapisuje się jako:  $D ::> K$ .

Metody znajdowania charakterystyki klas oparte na analizie wartości cech należących do klas obiektów przeszukują przestrzeń opisów od najbardziej ogólnych do szczegółowych. Ze zbioru przykładów pozytywnych (należących do opisywanej klasy) losowo wybiera się obiekt i tworzy się opisy dyskryminujące względem zbioru przykładów negatywnych (nie należących do opisywanej klasy). Otrzymane kompleksy są oceniane funkcją kryterialną LEF (*lexicographical evaluation functional*) o charakterze heurystycznym (AQ [Michalski 1983]) lub miarami statystycznymi czy funkcją entropii (CN2 [Clark, Niblett 1989]). Następnie usuwa się ze zbioru przykładów pozytywnych obiekty spełniające wszystkie testy w selektorach wybranego kompleksu i całą procedurę powtarza się, aż do wyczerpania zbioru przykładów pozytywnych.

Funkcjonał LEF jest definiowany jako uporządkowana lista par:  $(c_1, t_1), (c_2, t_2), \dots$ , gdzie  $c_i$  oznacza elementarne kryterium heurystyczne (np. uniwersalność opisu mierzoną liczbą obiektów opisywanych przez kompleks i nienależących do zbioru uczącego),  $t_i$  jest progiem tolerancji. Kryteria elementarne są dobierane adekwatnie do problemu przez prowadzącego klasyfikację, co oznacza nadanie im pewnego rodzaju wagi. W procesie oceny kompleksy testowane są parą  $(c_1, t_1)$ . Te, które

spełnią ustalone kryterium, są poddawane kolejnemu itd. Proces jest kontynuowany do momentu, gdy zostanie jeden „najlepszy” kompleks lub zbiór kryteriów będzie wyczerpany. W tym drugim przypadku wyboru kompleksu dokonuje się arbitralnie.

W algorytmie CN2 ocena kompleksów jest sterowana statystycznie. Jakość kompleksów jest mierzona funkcją entropii:

$$E = - \sum_{i=1}^k p_i \log_2 p_i,$$

gdzie  $p_i$  to prawdopodobieństwo występowania klasy  $K_i$  w zbiorze obiektów opisywanych przez badany kompleks,  $k$  to liczba klas. Im mniejsza jej wartość, tym lepszy opis w postaci kompleksu. Ocena statystycznej istotności kompleksu polega na sprawdzeniu, czy związek między wartościami cech a przynależnością do klasy nie jest przypadkowy. W CN2 stosuje się statystykę  $I$  opartą na ilorazie wiarygodności:

$$I = 2 \sum_{i=1}^k f_i \log_2 \left( \frac{f_i}{e_i} \right),$$

gdzie  $(f_1, \dots, f_k)$  to rozkład liczebności obiektów opisywanych przez badany kompleks należących do klas  $K_1, \dots, K_k$ ,  $(e_1, \dots, e_k)$  spodziewany rozkład tych liczebności w wyborze losowym. Statystyka  $I$  ma w przybliżeniu rozkład  $\chi^2$  o  $k - 1$  stopniach swobody, stąd wartości krytyczne testu istotności są znajduwane w tablicach rozkładu  $\chi^2$ . Im mniejsza wartość  $I$ , tym większe prawdopodobieństwo, że jakość opisu klasy, jaki daje kompleks, ma charakter losowy.

Minimalne drzewo logiczne dla zbioru obiektów wybranej klasy wytycza kierunek przeszukiwania przestrzeni opisu tej klasy. Proponuje się następującą procedurę.

1. Znajdź minimalne drzewo logiczne dla zbioru obiektów pierwszej klasy.
2. Przedstaw na drzewie obiekty pozostałych klas.
3. Odczytaj maksymalnie ogólny, odróżniający opis danej klasy w postaci dysjunkcyjnej.
4. Otrzymany opis uprość stosując prawa logiki.
5. Procedurę powtarzaj dla każdej klasy.

W odróżnieniu od wcześniej opisanego podejścia, minimalne drzewo logiczne generuje opisy dyskryminujące dla podzbiorów obiektów pozytywnych, które są identyczne ze względu na pewien podzbiór cech charakteryzujących się największą ogólnością opisu (a więc pierwszych w uporządkowaniu). Otrzymane w ten sposób kompleksy nie są już poddawane dodatkowej ocenie, choć, rezygnując z warunku niespełniania przez nie niektórych kontrprzykładów, można oceniać je podobnie jak w algorytmie CN2. Operacja opisana w kroku 3 oznacza tworzenie liści według kryterium całkowitej homogeniczności podzbiorów.

## 4. Przykład

Do ilustracji metody minimalnego drzewa logicznego wykorzystano zbiór danych HAYES-ROTH, dostępny na stronie internetowej <ftp://ftp.ics.uci.edu/pub/machine-learning-databases/>. W zbiorze tym są dane 132 obiekty (ludzie) należące do trzech klas. Opisane są czterema zmiennymi nominalnymi:  $X_1$  – hobby,  $X_2$  – wiek,  $X_3$  – wykształcenie,  $X_4$  – stan cywilny. Wartości cech kodowane są numerycznie;  $X_1$  trzema wartościami, pozostałe czterema. Dane zawierają także 28-elementowy zbiór testowy.

Metodą minimalnego drzewa logicznego otrzymano następujące uporządkowania zmiennych:  $(X_2X_3X_4)X_1$  dla klas 1 i 2 oraz  $X_1(X_2X_3X_4)$  dla klasy 3. Nawias oznacza dowolną kolejność zmiennych w nim występujących.

Rysunek 1 przedstawia fragment minimalnego drzewa wyznaczonego dla klas 1 i 2. Obiekty poszczególnych klas oznaczono znaczkami: \* klasa 1, + klasa 2 oraz ^ klasa 3. Linia przerywana reprezentuje obiekty spoza zbioru uczącego. Odczytać można kompleksy:

$$C_1: [X_2=1] \wedge [X_3=0] \wedge [X_4=0] ::> K_1, \quad C_2: [X_2=1] \wedge [X_3=0] \wedge [X_4=1] ::> K_2,$$

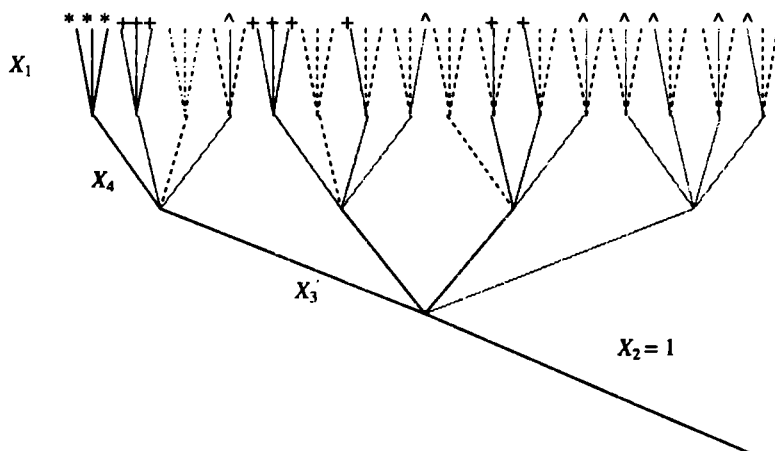
$$C_3: [X_2=1] \wedge [X_3=1] \wedge [X_4=0 \vee 2] ::> K_2, \quad C_4: [X_2=1] \wedge [X_3=2] \wedge [X_4=1 \vee 2] ::> K_2.$$

Poniżej przedstawiono zbiór reguł klasyfikacyjnych odczytanych z drzew logicznych.

$$\begin{aligned} & [X_2=0] \wedge [X_3=0] \wedge [X_4=1 \vee 2] \vee [X_2=0] \wedge [X_3=1 \vee 2] \wedge [X_4=0] \vee [X_2=1 \vee 2] \wedge [X_3=0] \wedge [X_4=0] \vee \\ & \vee [X_2=2] \wedge [X_3=0] \wedge [X_4=2] \vee [X_2=2] \wedge [X_3=2] \wedge [X_4=0] \quad ::> K_1 \\ & [X_2=0] \wedge [X_3=1] \wedge [X_4=1] \vee [X_2=1] \wedge [X_3=0] \wedge [X_4=1] \vee [X_2=1] \wedge [X_3=1] \wedge [X_4=0 \vee 2] \vee \\ & \vee [X_2=1] \wedge [X_3=2] \wedge [X_4=1 \vee 2] \vee [X_2=2] \wedge [X_3=1] \wedge [X_4=1 \vee 2] \vee [X_2=2] \wedge [X_3=2] \wedge [X_4=1] \quad ::> K_2 \\ & [X_1=0] \wedge [X_2=0 \vee 1] \wedge [X_3=3] \vee [X_1=1] \wedge [X_2=1 \vee 2] \wedge [X_3=3] \vee [X_1=2] \wedge [X_2=0 \vee 2] \wedge [X_3=3] \vee \\ & \vee [X_2=3] \quad ::> K_3 \end{aligned}$$

W celu porównania prezentuje się wyniki otrzymane algorytmem AQ11.

$$\begin{aligned} & [X_1=0] \wedge [X_2=0] \wedge [X_3=0] \vee [X_1=0] \wedge [X_2=0] \wedge [X_4=0] \vee [X_2=0] \wedge [X_3=2] \wedge [X_4=2] \vee \\ & \vee [X_2=1] \wedge [X_3=0] \wedge [X_4=0] \vee [X_2=0] \wedge [X_3=0] \wedge [X_4=1] \vee [X_1=1] \wedge [X_2=0] \wedge [X_4=0] \vee \\ & \vee [X_1=1] \wedge [X_2=2] \wedge [X_3=0] \vee [X_2=0] \wedge [X_3=1] \wedge [X_4=0] \vee [X_1=2] \wedge [X_2=2] \wedge [X_4=0] \quad ::> K_1 \\ & [X_1=0] \wedge [X_3=1] \wedge [X_4=1] \vee [X_2=1] \wedge [X_3=0] \wedge [X_4=1] \vee [X_1=0] \wedge [X_2=1] \wedge [X_3=1 \vee 2] \vee \\ & \vee [X_2=0] \wedge [X_3=1] \wedge [X_4=1] \vee [X_1=1] \wedge [X_2=1 \vee 2] \wedge [X_3=1] \vee [X_1=1] \wedge [X_3=2] \wedge [X_4=1] \vee \\ & \vee [X_2=1] \wedge [X_3=1] \wedge [X_4=0] \vee [X_1=2] \wedge [X_2=2] \wedge [X_3=1] \wedge [X_4 \neq 0] \vee [X_2=2] \wedge [X_3=2] \wedge [X_4=1] \quad ::> K_2 \\ & [X_2=3] \wedge \sim([X_1=0] \wedge [X_3=1] \wedge [X_4=1] \vee [X_1=1] \wedge [X_3=2] \wedge [X_4=1]) \vee \\ & \vee [X_3=3] \wedge \sim([X_1=0] \wedge [X_2=0] \wedge [X_4=0] \vee [X_1=1] \wedge [X_2=0] \wedge [X_4=0] \vee \\ & \vee [X_1=2] \wedge [X_2=2] \wedge [X_4=0]) \vee [X_4=3] \wedge \sim([X_1=0] \wedge [X_2=0] \wedge [X_3=0] \vee \\ & \vee [X_1=1] \wedge [X_2=2] \wedge [X_3=0] \vee [X_1=0] \wedge [X_2=1] \wedge [X_3=1 \vee 2] \vee [X_1=1] \wedge [X_2=1 \vee 2] \wedge [X_3=1] \\ & \vee \vee [X_1=2] \wedge [X_2=2] \wedge [X_3=1]) \quad ::> K_3 \end{aligned}$$



Rys.1. Fragment minimalnego drzewa logicznego dla klas 1 i 2  
Źródło: opracowane własne.

Tabela 1. Błąd klasyfikacji dla zbioru HAYES-ROTH

|      | Drzewa logiczne | AQ11   |
|------|-----------------|--------|
| Błąd | 0,3214          | 0,3571 |

Źródło: obliczenia własne.

Tabela 1 przedstawia porównanie błędu klasyfikacji mierzonego na zbiorze testowym jako frakcja obiektów błędnie sklasyfikowanych.

## 5. Wnioski

Drzewa logiczne dają inną od uogólnionych diagramów logicznych (ang. *generalized logical diagram*) ilustrację przestrzeni kombinacji wartości cech (zob. [Michalski 1977]). Minimalne drzewa logiczne porządkują zmienne jakościowe ze względu na ogólność opisu zbioru obiektów. Takie uporządkowania wytyczają kierunek przeszukiwania przestrzeni opisu klasy w klasyfikacji wzorcowej. Ponieważ zmienne ostatnie w uporządkowaniu mają mniejsze znaczenie w konstruowaniu charakterystyki klas (w przykładzie HAYES-ROTH zmienna  $X_1$ ) można je pomijać. Otrzymywane metodą minimalnego drzewa logicznego reguły klasyfikacji mogą się różnić od reguł algorytmu AQ11. W artykule pokazano, że może to prowadzić do otrzymania mniejszego błędu klasyfikacji. Są możliwości dalszych modyfikacji zaproponowanej metody.

## Literatura

- Clark P., Niblett T., *The CN2 Induction Algorithm*, „Machine Learning” 1989, t.3.
- Gatnar E., *Symboliczne metody klasyfikacji danych*, PWN, Warszawa 1998.
- Jaczewski J., *Układy logiczne i teoria automatów. Podstawy teoretyczne analizy i syntezy*, Wyd. Politechniki Gdańskiej, Gdańsk 1978.
- Kubus M., *Tworzenie logicznych drzew decyzyjnych w ujęciu analitycznym i syntetycznym*, XXXII Ogólnopolska Konferencja Zastosowań Matematyki PAN, Zakopane 2003.
- Michalski R.S., *Variable-Valued Logic and its Applications to Pattern Recognition and Machine Learning*, [w:] *Computer Science and Multiplevalued Logic*, red. D.C. Rine, North-Holland, Amsterdam 1977.
- Michalski R.S., *Theory and Methodology of Inductive Learning*, [w:] *Machine Learning: an Artificial Intelligence Approach*, red. R.S. Michalski, J.G. Carbonell, T. Mitchell, Palo Alto 1983.
- Pawlak Z., *Systemy informacyjne – podstawy teoretyczne*, WNT, Warszawa 1983.

## THE APPLICATION OF THE LOGICAL FUNCTIONS FOR NOMINAL ATTRIBUTES ORDERING

### Summary

This paper presents the method of nominal attributes ordering and its application to class description with the use of the logical trees. Minimal logical tree orders attributes according to generality of objects set description. Obtained order is used for description space search. The last variables in order have a little meaning in creating classification rules and often can be omitted. Proposed method is compared to well known algorithm AQ11.