PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1169 2007
TAKSONOMIA 14
Klasyfikacja i analiza danych – teoria i zastosowania

Mariusz Kubus
Politechnika Opolska

UPRASZCZANIE WSTĘPNE REGUŁ KLASYFIKACJI DRZEW LOGICZNYCH

1. Wstęp

Obszarem zainteresowań będą metody dyskryminacji, które można stosować dla zmiennych z mocnych i słabych skal pomiaru. W statystyce ugruntowaną już pozycję ma metoda drzew klasyfikacyjnych. Jednocześnie na gruncie nauki o indukcyjnych metodach uczenia rozwijano metody dyskryminacji prowadzące do modeli w postaci reguł klasyfikacji, których części warunkowe są koniunkcjami wartości cech. Znaczna ich liczba opiera się na ogólnym schemacie wypracowanym przez Michalskiego [1969]. Do konstrukcji takich modeli dyskryminacyjnych można zastosować także drzewa logiczne [Kubus 2006]. Zaproponowana przez autora metoda daje dokładności klasyfikacji porównywalne z powszechnie znanymi algorytmami CN2 (indukcja reguł) i C4.5 (drzewa klasyfikacyjne). Celem artykułu jest zastosowanie w metodzie drzew logicznych różnych technik upraszczania wstępnego modelu (*pre-pruning*) spotykanych w indukcji reguł i drzewach klasyfikacyjnych.

Ponieważ omawiane metody zostały opracowane głównie dla zmiennych niemetrycznych, postać klasyfikatora nie może zawierać operacji arytmetycznych. Naturalnym środkiem wyrazu w tej sytuacji jest posługiwanie się koniunkcjami wartości cech, co czynimy na co dzień. Na przykład zdanie: „jeśli będzie pochmurno i nie będzie wiatru, to pan X pójdzie na spacer" kryje regułę klasyfikacji. Można ją zapisać formalnie:

$$[\text{STAN NIEBA} = \text{pochmurno}] \wedge [\text{WIATR} = \text{nie}] \rightarrow \textit{spacer}.$$

Wszystkim obiektom opisanym przez koniunkcję warunków w części warunkowej reguły przypisywana jest klasa *spacer*. Wyrażenia postaci:

$$r : \textit{koniunkcja warunków} \rightarrow \textit{klasa}$$

nazywane będą dalej regułami klasyfikacji lub krótko – regułami. Model dyskryminacyjny w postaci zbioru reguł definiowany jest następująco:

$$K_1^1 \to \omega_1, \ldots, K_{r_1}^1 \to \omega_1$$
$$\ldots \qquad (1)$$
$$K_1^C \to \omega_C, \ldots, K_{r_C}^C \to \omega_C$$

gdzie: ω_j jest wartością zmiennej objaśnianej (klasą); K_i^j oznacza koniunkcję warunków w i-tej regule przypisującej klasyfikowany obiekt do j-tej klasy.

2. Konstrukcja reguł

W algorytmach indukcji reguł pojedyncza reguła powstaje przez przeszukiwanie przestrzeni opisów klas, a więc przestrzeni wszystkich możliwych koniunkcji warunków. Przestrzeń ta jest na ogół bardzo duża, a zadanie przeszukiwania kosztowne obliczeniowo. Dlatego algorytmy indukujące reguły przeważnie wykorzystują różne strategie przeszukiwania heurystycznego. Zachętą do stosowania tych technik są wyniki jakie uzyskali Quinlan i Cameron-Jones [1995]. Wykazali mianowicie, że ze wzrostem obszaru przeszukiwania dokładność predykcji uzyskiwanych modeli maleje. Znacząca część algorytmów indukcji reguł opiera się na schemacie wypracowanym przez Michalskiego [1969] (*covering algorithms*). Generowana jest pojedyncza reguła maksymalizująca wybrane kryterium jakości, a następnie ze zbioru uczącego usuwane są obiekty opisane przez część warunkową tej reguły. Na tak zmodyfikowanym zbiorze uczącym poszukuje się kolejnej optymalnej reguły itd. Proces kontynuowany jest do momentu opisania całego zbioru uczącego lub przyjmuje się ustalone kryteria stopu. Uzyskane segmenty w wielowymiarowej przestrzeni zmiennych (regiony decyzyjne) nie muszą być rozłączne. Nie muszą spełniać warunku zupełności, ani opisywać wszystkich obiektów zbioru uczącego. Taka struktura regionów decyzyjnych prowadzi do niejednoznaczności w etapie rozpoznawania, z którymi algorytmy te musiały się zmierzyć. Jednakże reguły niemające struktury drzewa dają nowe możliwości odkrywania regularności tkwiących w danych, innych niż wykryłyby drzewa klasyfikacyjne. To było główną motywacją zainteresowań tymi algorytmami.

Model dyskryminacyjny postaci (1) można też uzyskać z drzew logicznych (zob. [Kubus 2006]). Reguły klasyfikacji konstruuje się dla każdej klasy oddzielnie, wykorzystując minimalne drzewo logiczne oraz przeszukiwanie węzłów. Uzyskana struktura geometryczna regionów decyzyjnych również nie spełnia na ogół warunku rozłączności i zupełności.

Poniżej przedstawione zostaną niektóre miary stosowane do oceny jakości reguł. Szerszy ich przegląd można znaleźć w artykule Fürnkranza [1999]. Literami p, P oznaczono liczby obiektów klasy opisywanej, literami n, N liczby obiektów

pozostałych klas. Małe litery oznaczają liczebności obiektów w podzbiorach opisywanych przez części warunkowe ocenianych reguł r, natomiast duże litery, liczebności w całym zbiorze uczącym.

Precyzja 1: $$P_1(r) = \frac{p}{p+n}. \tag{2}$$

Precyzja 2: $$P_2(r) = \frac{p-n}{p+n}. \tag{3}$$

Entropia: $$E(r) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}. \tag{4}$$

Entropia uwzględniająca prawdopodobieństwa *a priori* (*cross entropy*):

$$CE(r) = -\frac{p}{p+n}\log_2\frac{\frac{p}{p+n}}{\frac{P}{P+N}} - \frac{n}{p+n}\log_2\frac{\frac{n}{p+n}}{\frac{N}{P+N}}. \tag{5}$$

J-miara: $$J(r) = pCE(r). \tag{6}$$

Estymator Laplace'a: $$LAP(r) = \frac{p+1}{p+n+C}. \tag{7}$$

m-estymator: $$M(r) = \frac{p+m\frac{P}{P+N}}{p+n+m}. \tag{8}$$

Zawartość: $$Z(r) = \frac{\frac{p+1}{P+1}}{\frac{n+1}{N+1}}. \tag{9}$$

Względna dokładność ważona (*weighted relative accuracy*):

$$WRA(r) = \frac{p+n}{P+N}\left(\frac{p}{p+n} - \frac{P}{P+N}\right). \tag{10}$$

3. Upraszczanie wstępne modeli

Podstawowym problemem budowania modeli dyskryminacyjnych (także regresyjnych) jest nadmierne dopasowanie do danych (*overfitting*). W celu uzyskania modelu o dokładnej predykcji na nowych obiektach spoza zbioru uczącego stosowane są różne techniki upraszczania (*pruning*). Możliwe są dwa podejścia. Upraszczanie wstępne (*pre-pruning*) nie dopuszcza do zbudowania modelu całkowicie separującego klasy. Można też model modyfikować po zbudowaniu i całkowitym dopasowaniu do danych (*post-pruning*). O ile drzewa klasyfikacyjne wykorzystują

na ogół upraszczanie końcowe, to w algorytmach indukcji reguł częściej stosuje się upraszczanie wstępne. Różne techniki upraszczania wstępnego formułowane są na ogół w postaci kryteriów stopu.

W przypadku drzew klasyfikacyjnych chodzi o zatrzymanie podziału. Kryteria stopu dla drzew klasyfikacyjnych odwołują się do miary jakości podziału lub do struktury drzewa (zob. [Gatnar 2001]).

W algorytmach indukcji reguł można wyróżnić trzy najczęściej stosowane sposoby upraszczania wstępnego (zob. [Furnkranz 1997]):

- testowanie statystycznej istotności reguł,
- kryterium stopu oparte na minimalnej jakości reguły (*cutoff stopping criterion*),
- kryterium stopu bazujące na zasadzie minimalnej długości opisu (*minimum description length principle*).

Test statystycznej istotności reguł nie jest kryterium stopu, a raczej rodzajem filtrowania reguł. Odrzucane są reguły, które mogłyby pojawić się losowo. W kryterium minimalnej jakości przyjmowana jest arbitralnie wartość progowa miary oceniającej jakość reguł. Generowanie pojedynczej reguły jest zatrzymane w momencie, gdy ocena przekroczy ustalony próg. Z kolei kryteria stopu bazujące na zasadzie minimalnej długości opisu zatrzymują budowę modelu w chwili pojawienia się reguły opisującej niewiele obiektów. Chodzi o to, by ilość informacji potrzebna do zakodowania reguły nie była większa od ilości informacji potrzebnej do zakodowania opisywanych przez nią obiektów. Dla drzew logicznych autor zastosował trzy, najbardziej adekwatne do metody, techniki upraszczania wstępnego:

- kryterium minimalnej jakości,
- ograniczenie głębokości przeszukiwania drzewa,
- ustalenie minimalnej liczebności podzbiorów.

Dwie pierwsze mają na celu szybsze tworzenie liści (a więc reguł) w trakcie przeszukiwania. Trzecia może być stosowana w dwóch odmianach. Węzeł reprezentujący niewiele obiektów można zamienić w liść (kryterium stopu) lub w ogóle z takiej reguły zrezygnować (filtrowanie).

4. Wyniki

Badania przeprowadzono na czterech powszechnie uznawanych zbiorach danych udostępnianych przez repozytorium Uniwersytetu Kalifornijskiego [Blake, Keogh, Mertz 1998]. Sprawdzono pięć miar oceniających homogeniczność podzbiorów opisywanych przez części warunkowe reguł: precyzję (3), entropię, J-miarę, estymator Laplace'a oraz m-estymator. W odniesieniu do każdej miary dobierano arbitralnie od pięciu do siedmiu wartości progowych w kryterium minimalnej jakości. W każdym wariancie sprawdzano wszystkie możliwe głębokości przeszukiwania. W odniesieniu do każdego zestawu parametrów dokładność klasyfikacji oszacowano metodą dziesięcioczęściowego sprawdzania krzyżowego. Za-

leżność dokładności klasyfikacji od głębokości przeszukiwania przedstawiono na rys. 1. Ze względu na dużą objętość opracowanego materiału przedstawiono tylko wybrane wykresy, najbardziej reprezentatywne dla poszczególnych zbiorów danych. Na każdym wykresie przedstawiono kilka linii odpowiadających różnym wartościom progowym miar. Niezależnie od stosowanych miar i przyjmowanych wartości progowych zauważono, że:

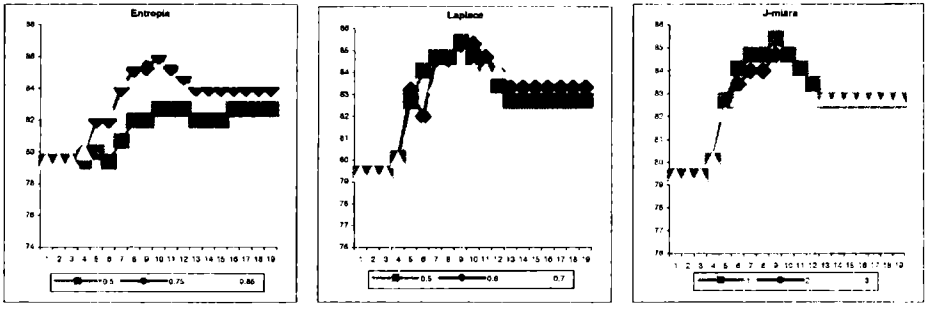
- dla zbioru *hepatitis* najwyższe dokładności klasyfikacji osiągano na głębokości przeszukiwania 10; głębsze przeszukiwanie powodowało spadek dokładności klasyfikacji,
- dla zbioru *voting-records* najwyższe dokładności klasyfikacji osiągano już na głębokości przeszukiwania 4; głębsze przeszukiwanie nie zmieniało osiągniętego maksimum,
- dla zbioru *echocardio* najwyższe dokładności klasyfikacji osiągano na ogół na głębokości przeszukiwania 3 lub 4,
- zbiór *heart disease H* był jedynym przykładem, gdzie maksymalną dokładność uzyskano dopiero na głębokości 12, co w praktyce oznaczało przeszukanie prawie całego drzewa; warto jednak zwrócić uwagę, że od głębokości 5 lub 6 dokładność klasyfikacji rosła już wolniej ze wzrostem głębokości. Zatem wszystko wskazuje na to, że nie warto przeszukiwać całego drzewa logicznego. Do korzyści płynących z tego wniosku należy zaliczyć skrócenie czasu obliczeń oraz uzyskanie prostszego, łatwiejszego w interpretacji modelu. Najwyższe dokładności klasyfikacji, jakie uzyskano w poszczególnych zbiorach danych, przedstawia tab. 1. Poprawiono więc wyniki przedstawione w pracy [Kubus 2006], w których zastosowano entropię z wartością progową 0,7 oraz przeszukiwanie kończące się w chwili opisania wszystkich obiektów.

Tabela 1. Najwyższe dokładności klasyfikacji (w %) dla wybranych pięciu miar oceniających jakość reguł w metodzie drzew logicznych

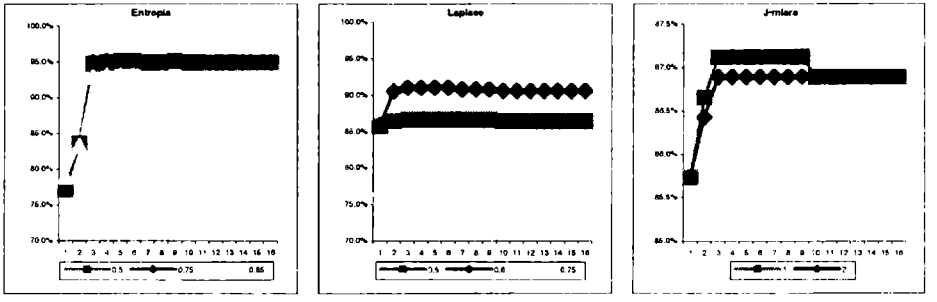
Wyszczególnienie	Precyzja	Entropia	J-miara	Estymator Laplace'a	M-estymator
Hepatitis	86,6	86,6	85,4	85,4	85,4
Voting-records	95,2	95,2	87,1	95	95
Echocardio	71,1	72,5	73,3	72,5	71,8
Heart-disease H	83	82,7	80,3	80,6	80,6

Źródło: obliczenia własne.

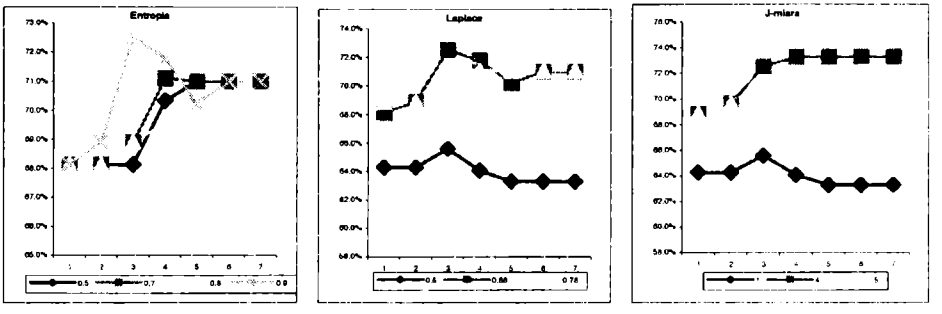
Hepatitis



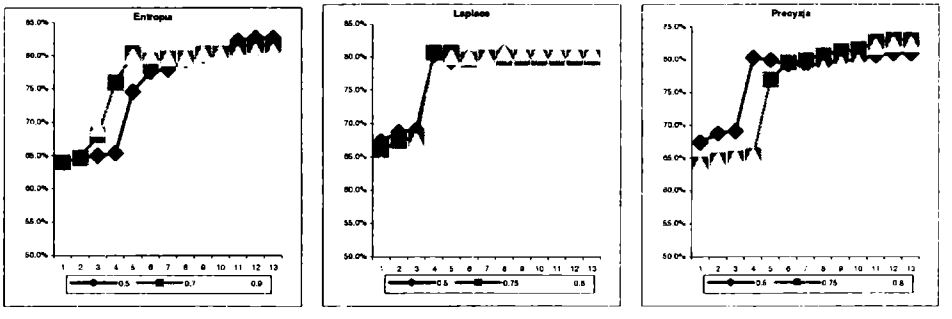
Voting-records



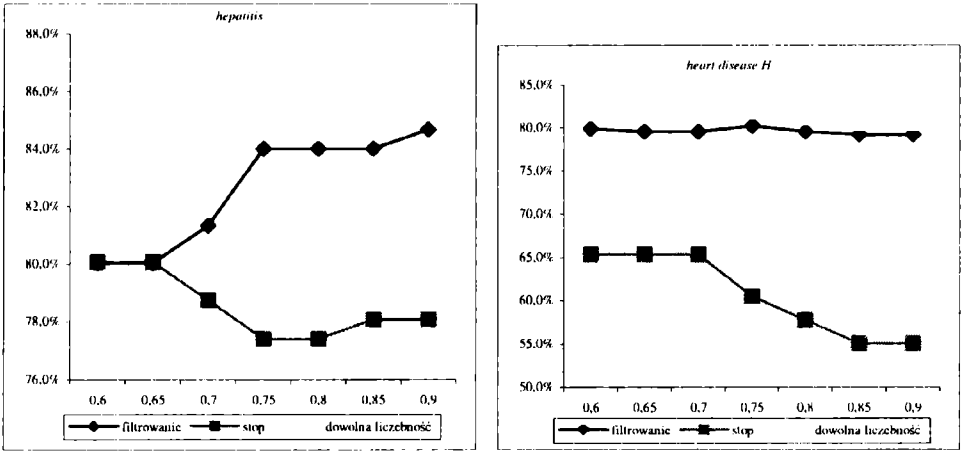
Echocardio



Heart disease H



Rys. 1. Zależność dokładności klasyfikacji od głębokości przeszukiwania drzewa logicznego
Źródło: obliczenia własne.



Rys. 2. Porównanie dokładności klasyfikacji dla filtrowania reguł i kryterium stopu ustalanych na podstawie minimalnej liczebności podzbiorów, którą ustalono na 5% zbioru uczącego
Źródło: obliczenia własne.

Kontrolę liczebności podzbiorów zbadano w dwóch odmianach: jako kryterium stopu oraz jako filtrowanie. Minimalną liczebność podzbiorów ustalano od 1 do 5% liczebności zbioru uczącego. Rysunek 2 przedstawia dokładności klasyfikacji w dwóch wybranych przykładach. Badanie przeprowadzono dla różnych wartości progowych entropii (oś X), a głębokość przeszukiwania ustalano optymalną dla poszczególnych zbiorów. Zastosowanie kryterium minimalnej liczebności podzbiorów nie wpłynęło na poprawę dokładności klasyfikacji w żadnym z badanych zbiorów, a nieraz znacznie ją obniżyło.

5. Podsumowanie

Najbardziej naturalną techniką upraszczania wstępnego dla drzew logicznych jest kryterium minimalnej jakości. Zapewnia ono uzyskanie reguł opisujących regiony decyzyjne o z góry ustalonym, minimalnym stopniu homogeniczności. Spośród zbadanych pięciu miar oceniających homogeniczność nieco lepsze wyniki uzyskano, stosując entropię lub precyzję. Trudnością stosowania tego kryterium jest określenie odpowiedniej wartości progowej miary. Na zbadanych zbiorach danych nie odkryto tu żadnej prawidłowości.

Dodatkowe zastosowanie ograniczenia głębokości przeszukiwania węzłów nie tylko skraca czas obliczeń, ale często (3 spośród 4 zbadanych zbiorów) poprawia dokładność klasyfikacji uzyskanego modelu. Również tu nie odkryto prawidłowości co do arbitralnego wyboru maksymalnej głębokości. Dla wyznaczenia optymalnej wartości progowej miary i głębokości przeszukiwania stosuje się sprawdzanie krzyżowe, co niestety zwiększa czas obliczeń. Warto odnotować, że stosowanie

wyłącznie ograniczenia głębokości przeszukiwania węzłów nie zapewnia dostatecznej homogeniczności regionów decyzyjnych i w efekcie prowadzi do mniej dokładnych modeli.

Z kolei ustalanie minimalnej liczebności podzbiorów opisywanych regułami we wszystkich zbadanych zbiorach prowadziło do obniżenia dokładności klasyfikacji.

Literatura

Blake C., Keogh E., Merz C.J. (1998), *UCI Repository of Machine Learning Databases*, Department of Information and Computer Science, University of California, Irvine, www.ics.uci.edu/~mlearn/MLRepository.html.

Fürnkranz J. (1997), *Pruning Algorithms for Rule Learning*, „Machine Learning" 27(2).

Fürnkranz J. (1999), *Separate-and-Conquer Rule Learning*, „Artif. Intelligence Review" 13(1).

Fürnkranz J., Flach P.A. (2005), *ROC 'n' Rule Learning – Towards a Better Understanding of Covering Algorithms*, „Machine Learning" 58, s. 39-77.

Gatnar E. (2001), *Nieparametryczna metoda dyskryminacji i regresji*, PWN, Warszawa.

Gatnar E. (1998), *Symboliczne metody klasyfikacji danych*, PWN, Warszawa.

Kubus M. (2006), *O pewnej metodzie dyskryminacji obiektów jakościowych*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1126, AE, Wrocław.

Michalski R.S. (1969), *On the Quasi-Minimal Solution of the Covering Problem*, [w:] Proceedings of the 5th International Symposium on Information Processing (FCIP-69) vol. A3 (Switching Circuits), Bled, Yugoslavia, s. 125-128 .

Quinlan J.R., Cameron-Jones R.M. (1995), *Oversearching and Layered Search in Empirical Learning*, [w:] C. Mellish (red.), Proceedings of the 14th International Joint Conference on Artificial Intelligence, Morgan Kaufman, s. 1019-1024.

PRE-PRUNING OF CLASSIFICATION RULES FOR LOGICAL TREES

Summary

Discriminant models that ideally separate classes do not perform well on the new observations, especialy in the presence of noise. This is the general problem of discriminant analysis. Several technics of pruning were proposed in the literature to avoide overfitting. In this paper, pre-pruning technics that deal with noise during model generation are considered.

The main goal of this paper is to adopt pre-pruning technics – used in classification trees and rules induction – for logical trees based discrimination method. The examples from UCI Repository of Machine Learning Databases are given.