PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 47 2009
TAKSONOMIA 16
Klasyfikacja i analiza danych – teoria i zastosowania

Mariusz Kubus
Politechnika Opolska

PORÓWNANIE INDUKCJI REGUŁ Z WYBRANYMI METODAMI DYSKRYMINACJI

1. Wstęp

Klasyczne metody analizy dyskryminacyjnej, choć dobrze ugruntowane matematycznie, napotykają pewne ograniczenia i trudności w stosowaniu. Zostały one opracowane dla danych metrycznych, są wrażliwe na występowanie obserwacji oddalonych, wymagają kompletności danych. Pewnym ograniczeniem jest też *a priori* przyjmowany kształt dyskryminant (najczęściej liniowy). Powszechnie stosowana liniowa analiza dyskryminacyjna obliguje do poczynienia założeń, które w rzeczywistości mogą nie być spełnione, na przykład o rozkładach czy postaci macierzy kowariancji w klasach. Z kolei w przypadku kwadratowych (wielomianowych) funkcji dyskryminacyjnych liczba parametrów do oszacowania rośnie wykładniczo wraz z liczbą uwzględnianych w modelu zmiennych objaśniających (przekleństwo wymiarowości). Model w postaci nierówności nie jest też specjalnie łatwy w interpretacji, co nabiera znaczenia, gdy stosujemy analizę dyskryminacyjną w celu dokonania różnego rodzaju identyfikacji, na przykład w marketingu: cech różnicujących produkty czy firmy oraz klientów o różnych preferencjach.

Spowodowało to rozwój metod nieparametrycznych o charakterze adaptacyjnym, które przezwyciężają wymienione trudności, a ponadto generują modele wygodne w interpretacji, wykorzystujące język naturalny. Klasyfikacji dokonuje się na podstawie reguł w postaci: *jeśli spełnione są warunki, to przydziel obiekt do klasy*, gdzie warunki mają postać koniunkcji wartości cech. Największą chyba popularność w tej grupie metod zyskały drzewa klasyfikacyjne. Metoda ta polega na rekurencyjnym podziale przestrzeni cech na segmenty o coraz większym stopniu homogeniczności (ze względu na klasę). Podziały dokonywane są na podstawie wartości zmiennych wprowadzanych do modelu adaptacyjnie za pomocą strategii wspinaczki (zob. [Gatnar 2001]). Równolegle z drzewami klasyfikacyjnymi rozwijały się algorytmy indukcji reguł, które reprezentują alternatywne podejście do identyfikacji homogenicznych regionów (zob. [Fürnkranz, Widmer 1999]). Polega ono na analizie wartości cech obiektów, co realizowane jest przez

heurystyczne przeszukiwanie przestrzeni opisów klas (koniunkcje wartości cech). Stosuje się tu najczęściej strategię wspinaczki lub przeszukiwanie wiązką (*beam search*) oraz funkcję oceny jakości reguł (np. entropię). Przeszukanie całej przestrzeni (choć staje się już mniej problematyczne wobec mocy obliczeniowej komputerów) nie jest wskazane ze względu na zjawisko nadmiernego przeszukiwania (zob. [Quinlan, Cameron-Jones 1995]). Modele w postaci zbioru reguł nie muszą mieć graficznego przedstawienia drzewa i mogą prowadzić do klasyfikacji nierozłącznej. Jednym z najefektywniejszych algorytmów indukcji reguł jest RIPPER [Cohen 1995].

Celem artykułu jest empiryczne porównanie błędów klasyfikacji uzyskanych tym algorytmem z wynikami innych metod dyskryminacji. Wykorzystane będą zbiory danych rzeczywistych udostępniane przez repozytorium Uniwersytetu Kalifornijskiego [Blake, Keogh, Merz 1998].

2. Indukcja reguł

Metody indukcji reguł generują modele w postaci zbioru reguł $\{R_i\}_{i\in\{1,\ldots,s\}}$, gdzie pojedyncza reguła ma postać implikacji:

$$R_i : \textit{koniunkcja warunków} \rightarrow \textit{klasa}. \tag{1}$$

Warunki najczęściej mają postać $X_n = v$ w przypadku zmiennych nominalnych (v jest kategorią) lub $X_c \geq l$ $(X_c \leq l)$ dla zmiennych metrycznych (w algorytmie RIPPER l jest jedną z wartości przyjmowanych przez zmienną X_c). Odpowiadające regułom regiony decyzyjne nie muszą być rozłączne i nie muszą spełniać warunku zupełności, tzn. niektóre obiekty zbioru uczącego mogą pozostać nieopisane.

2.1. Schemat indukcji reguł *separuj-i-zwyciężaj*

Klasyczny już schemat indukcji reguł *separuj-i-zwyciężaj* zaproponował po raz pierwszy Michalski [1969]. Polega on na powtarzaniu dwóch kroków. W pierwszym generowana jest pojedyncza reguła. W przeciwieństwie do drzew klasyfikacyjnych dokonuje się tego przez heurystyczne przeszukiwanie przestrzeni opisów klas, a więc wszystkich możliwych koniunkcji warunków. Kandydujące opisy (koniunkcje wartości cech) oceniane są funkcją kryterium i wybierany jest najlepszy z nich. W drugim kroku ze zbioru uczącego usuwane są obiekty opisane przez wygenerowaną regułę. Na tak zmodyfikowanym zbiorze uczącym generowana jest kolejna reguła itd. Kroki powtarzane są do momentu opisania wszystkich obiektów zbioru uczącego lub do spełnienia wprowadzonego kryterium stopu.

Takie postępowanie, podobnie jak w drzewach klasyfikacyjnych, stwarza możliwość idealnego odseparowania klas. Zbudowany w ten sposób model dyskryminacyjny nie gwarantuje jednak wysokiej dokładności klasyfikacji dla nowo pojawiających się obiektów. W celu uniknięcia nadmiernego dopasowania do danych ze zbioru uczącego (*overfitting*) stosowane są różne techniki upraszczania modelu. Uprasz-

czanie końcowe (*post-pruning*) polega na usuwaniu warunków z reguł w modelu idealnie dopasowanym do danych tak, aby zwiększyć poprawność klasyfikacji nowych obiektów. Z kolei upraszczanie wstępne (*pre-pruning*) to techniki stosowane w trakcie budowy modelu, które mają na celu nie dopuścić do powstania modelu idealnie dopasowanego. Najczęściej realizuje się to przez rozmaite kryteria stopu.

Różne algorytmy indukcji reguł przebiegające według omówionego schematu różnią się stosowaną strategią przeszukiwania, funkcją oceny jakości reguł oraz technikami upraszczania modelu (zob. [Fürnkranz, Widmer 1999]).

2.2. Algorytm RIPPER

Algorytm RIPPER (*Repeated Incremental Pruning to Produce Error Reduction*) Cohena [1995] jest modyfikacją opracowanego przez Fürnkranza i Widmera [1994] algorytmu IREP. Ich skuteczność wynika z zaawansowanego mechanizmu upraszczania modelu, który łączy idee upraszczania wstępnego i końcowego.

W algorytmie RIPPER można uzyskać dwie postaci modelu, co ma ścisły związek z rozwiązaniem zadania dyskryminacji dla wielu klas. Pierwszy to nieuporządkowany zbiór reguł, w którym kolejność reguł nie wpływa na klasyfikację. Tu problem dyskryminacji wielu klas rozwiązywany jest przez klasyczny schemat przeciwstawiania każdej klasie obiektów pozostałych klas. Druga postać, uporządkowana lista opisów klas, jest charakterystyczna dla omawianego algorytmu. Najpierw klasy porządkowane są od najmniej licznej do najliczniejszej, a następnie opis klasy konstruowany jest przez przeciwstawienie jej klas następnych w uporządkowaniu. W ten sposób ostatnia (najliczniejsza) klasa nie jest opisywana, a obiekty są do niej klasyfikowane regułą domyślną wtedy, gdy nie są opisane żadną regułą w modelu. Gdy obiekt opisany jest więcej niż jedną regułą, klasyfikowany jest przez pierwszą regułę, która go opisze.

Opis wybranej klasy przebiega według następujących kroków. Najpierw zbiór uczący podzielony jest losowo na: *część uczącą* oraz *część testową*. Na podstawie *części uczącej* konstruowana jest pojedyncza reguła, która opisuje tylko obiekty klasy opisywanej. Konstrukcja zaczyna się od reguły ogólnej (opisującej cały zbiór uczący), do której systematycznie dołączane są warunki tak, by zoptymalizować funkcję oceniającą jakość reguły:

$$f(R) = -p(\log\frac{p'}{p'+n'} - \log\frac{p}{p+n}), \tag{2}$$

gdzie: p,n – liczby obiektów opisanych regułą R (p – klasy opisywanej, n − pozostałe); p',n' – liczby obiektów opisanych przed dołączeniem kolejnego warunku. Wykorzystuje się tu strategię wspinaczki. Natychmiast po skonstruowaniu pojedynczej reguły jest ona upraszczana na podstawie *części testowej*. Z reguły R usuwane są końcowe sekwencje warunków strategią wspinaczki, dopóki usunięcie kolejnej nie spowoduje obniżenia wartości funkcji oceniającej jakość reguły:

$$P(R) = \frac{p-n}{p+n}. \tag{3}$$

Jakość reguły oceniana jest na *części testowej*, która nie bierze udziału w konstrukcji reguł. W tym momencie RIPPER stosuje jeszcze kryterium stopu mające na celu zatrzymanie konstrukcji opisu klasy. Wykorzystuje ono zasadę minimalnej długości opisu MDL (*Minimum Description Lenght*) [Rissanen 1978], która w dyskryminacji stosowana jest do wyboru właściwej postaci modelu spośród wielu kandydujących. Zasada MDL wywodzi się z teorii informacji i polega na zakodowaniu modelu oraz obiektów błędnie klasyfikowanych przez model. Liczona jest liczba bitów potrzebna do transmisji takiego kodu i wybierany jest model, dla którego liczba ta jest najmniejsza. W omawianym algorytmie konstrukcja opisu klasy jest zatrzymywana, jeśli dołączenie ocenianej reguły do modelu spowodowałoby zwiększenie wartości funkcji MDL o więcej niż d bitów od najmniejszej dotychczas uzyskanej długości opisu. W implementacji Cohena $d = 64$. Jeśli kryterium to nie jest spełnione, reguła dodawana jest do opisu klasy, a opisane obiekty są usuwane ze zbioru uczącego. Następnie generowana jest kolejna reguła. Jeśli algorytm opisu klasy nie osiągnie wcześniej omówionego kryterium stopu, to kończy działanie po opisaniu obiektów danej klasy.

Uzyskany zbiór reguł $R_1,...,R_r$ jest jeszcze poddany modyfikacji. Zastosowane jest więc jeszcze w pewnym sensie upraszczanie końcowe. Rozważana jest każda reguła R_i z osobna w kolejności, w jakiej były dodawane do zbioru reguł, i konstruowane są dwie reguły alternatywne do niej: R_i^z, R_i^k. Reguła zastępcza R_i^z powstaje przez ponowne wygenerowanie reguły dyskryminującej (opisującej tylko obiekty klasy opisywanej) oraz uproszczenie tak, by zminimalizować błąd klasyfikacji na *części testowej* dla całego zbioru reguł $R_1,...,R_i^z,...,R_r$. Drugą korektę R_i^k reguły R_i buduje się podobnie, przy czym etap generowania nie zaczyna się od reguły ogólnej, lecz polega na uszczegóławianiu reguły R_i (dodawaniu do niej warunków). Ostateczna decyzja nad wyborem reguły oryginalnej R_i lub którejś z jej alternatyw: R_i^z, R_i^k podejmowana jest po zastosowaniu zasady MDL. Wariant rozważanej reguły wstawiany jest do zbioru reguł, a następnie usuwa się z tego zbioru reguły tak, by zminimalizować długość opisu.

Istotną poprawę dokładności klasyfikacji w stosunku do algorytmu IREP Cohen uzyskał nie tylko przez przedstawioną modyfikację zbioru reguł, ale też przez ponowne wywołanie omówionego wcześniej algorytmu w celu opisania obiektów dotąd nieopisanych (co może być powtarzane wielokrotnie).

3. Badania empiryczne

Badania porównawcze przeprowadzono na zbiorach danych udostępnianych przez repozytorium Uniwersytetu Kalifornijskiego [Blake, Keogh, Merz 1998] (zob. tab. 1). Dobrano je tak, by były zróżnicowane ze względu na: rodzaj zmiennych, liczbę klas, liczbę obiektów, zmiennych oraz kompletność danych. Zbiory,

które nie miały oryginalnie dołączonego zbioru testowego, podzielono losowo na próbę uczącą i testową (1/3 zbioru uczącego). Błąd klasyfikacji szacowano wszędzie na zbiorze testowym, który nie brał udziału w etapie uczenia.

W przeprowadzonych badaniach porównawczych, oprócz algorytmu RIPPER, zastosowano: drzewa klasyfikacyjne CART, wielowymiarowe drzewa klasyfikacyjne QUEST, liniową analizę dyskryminacyjną (LDA) oraz metodę *k* najbliższych sąsiadów (kNN). W metodzie CART zastosowano indeks Giniego oraz przycinanie na podstawie kosztu i złożoności. Do uzyskania mniejszego błędu klasyfikacji nie stosowano reguły jednego błędu standardowego (1*SE rule*). Nie wprowadzono też minimalnej liczebności podzbiorów. Wielowymiarowe drzewa klasyfikacyjne również przycinano na podstawie kosztu i złożoności oraz nie stosowano reguły 1SE. Minimalna liczebność podzbioru wynosiła 5 (opcja domyślna w programie Statistica 6.1). Liniową analizę dyskryminacyjną stosowano bez analizy krokowej. Obliczenia w zbiorze *ecoli* metodami LDA i QUEST były niewykonalne, ponieważ zbiór zawiera klasy jednoelementowe. W metodzie *k* najbliższych sąsiadów parametr *k* wybierano jako przybliżenie liczby $N^{2/8}$, gdzie *N* jest liczbą obiektów (zob. [Enas, Choi 1986]). Remisy rozstrzygano zasadą majoryzacji. Odległość liczona była metryką euklidesową. W metodach LDA i kNN brakujące dane zastępowano średnimi.

Tabela 1. Zbiory danych wykorzystane w badaniach

Zbiory	Liczba obiektów	Liczba zmiennych		Liczba klas	Braki danych
		nominalnych	metrycznych		
Adult	48842	8	6	2	6465
Breast cancer	569	-	30	2	-
Car	1728	6	-	4	-
Credit Australian	690	8	6	2	-
Credit German	1000	13	7	2	-
Echocardio	131	-	7	2	39
Ecoli	336	-	7	8	-
Glass	194	-	9	6	-
Heart-disease C	303	8	5	2	-
Heart-disease H	294	8	5	2	746
Hepatitis	155	13	6	2	167
Ionosphere	351	-	33	2	-
Lymphography	148	18	-	4	-
Pima	768	-	8	2	-
Satellite	6435	-	36	6	-
Thyroid	1960	22	7	3	3092
Vehicles	846	-	18	4	-
Voting-records	435	16	-	2	392
Vowel	990	-	10	11	-
Wine	178	-	13	3	-
Zoo	101	15	1	7	-

Źródło: [Blake, Keogh, Merz 1998].

Tabela 2. Błąd klasyfikacji (w %) dla zbiorów ze zmiennymi niemetrycznymi

Zbiory	RIPPER zbiór reguł	RIPPER lista reguł	CART
Adult	14,9	15,6	15,6
Car	6,1	14,9	2,4
Credit Australian	16,1	14,8	15,2
Credit German	27	27,6	24,6
Heart-disease C	21,7	19,4	27,4
Heart-disease H	18,5	21	22,4
Hepatitis	21,5	21,5	21,3
Lymphography	14,3	10,2	24,5
Thyroid	3,1	2,6	4,1
Voting-records	5	4,1	4,4
Zoo	5,9	11,8	5,9

Źródło: obliczenia własne.

Tabela 3. Błąd klasyfikacji (w %) dla zbiorów danych metrycznych

Zbiory	RIPPER zbiór reguł	RIPPER lista reguł	CART	QUEST	LDA	kNN
Breast cancer	3,7	4,2	6,3	2,6	3,2	6,8
Echocardio	35,9	32	32,8	32,8	27,3	31,8
Ecoli	20,5	24,1	19,6	-	-	17
Glass	45,3	34,4	42,2	45,3	39,1	45,3
Ionosphere	10,3	10,3	12	14,5	15,4	19,7
Pima	22,7	24,6	22,7	21,9	19,9	26,6
Satellite	14,9	14,6	13,9	14,7	17,2	10,4
Vehicles	29,1	31,6	34,8	19,5	19,1	41,5
Vowel	33,9	31,5	19,4	17,9	35,5	9,7
Wine	1,7	8,5	5,1	5,1	1,7	30,5

Źródło: obliczenia własne.

Tabela 4. Bilanse „wygrana – przegrana – remis”

	CART	QUEST	LDA	kNN
RIPPER zbiór reguł	9-10-2	3-6-1	4-5-1	5-4-1
RIPPER lista reguł	11-9-1	5-5-0	5-5-0	6-4-0

Źródło: obliczenia własne.

W tabeli 2 zestawiono wyniki dla danych, w których występowały zmienne niemetryczne, natomiast tab. 3 zawiera wyniki dla danych metrycznych. Żadna z metod nie wykazuje wyraźnej przewagi nad pozostałymi. Z kolei w tab. 4 zestawiono bilanse „wygrana – przegrana – remis”. Wygrana rozumiana jest tu jako liczba zbiorów danych, dla których metoda w wierszu tabeli dawała mniejszy błąd

klasyfikacji od metody w kolumnie tabeli itd. Każda zastosowana w badaniu metoda okazała się też najskuteczniejsza na kilku zbiorach danych. Bilanse „wygrana – przegrana – remis" przeprowadzono więc także osobno dla zbiorów z dużą liczbą zmiennych, z liczbą klas większą niż dwie itp. Wyniki okazały się porównywalne, z wyjątkiem przypadku wielu klas, kiedy RIPPER miał nieznaczną przewagę nad LDA (bilans 4-2-0), oraz w przypadku danych niekompletnych, gdy RIPPER okazał się nieco skuteczniejszy od CART (bilans 4-1-1).

4. Podsumowanie

Na podstawie porównania błędów klasyfikacji można stwierdzić, że algorytm RIPPER daje porównywalne wyniki z innymi metodami dyskryminacji, a w niektórych zadaniach wykazuje nieznaczną przewagę. W siedmiu z dwudziestu jeden zbadanych zbiorów dał najmniejszy błąd klasyfikacji. Można więc uznać, że indukcja reguł według schematu *separuj-i-zwyciężaj* jest cennym narzędziem dyskryminacji. Wśród zalet metody dodatkowo należy zwrócić uwagę na możliwość analizy danych mierzonych na różnych skalach pomiaru, danych niekompletnych, na wygodne w interpretacji modele wyrażone językiem naturalnym, a także możliwość klasyfikacji nierozłącznej, co może mieć szczególne znaczenie, na przykład na rynkach z dużą konkurencją.

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.

Cohen W.W. (1995), *Fast effective rule induction,* [w:] Proceedings of the 12th International Conference on Machine Learning, red. A. Prieditis, S. Russell, Morgan Kaufmann, Lake Tahoe, CA, s. 115-123.

Enas G.G., Choi S.C. (1986), *Choice of the smoothing parameter and efficiency of k-nearest neighbor classification,* "Computer and Mathematics with Applications", 12A(2), s. 235-244.

Fürnkranz J. (1999), *Separate-and-conquer rule learning,* "Artif. Intelligence Review", 13(1), s. 3-54.

Fürnkranz J., Widmer G. (1994), *Incremental reduced error pruning*, [w:] Proceedings of the Eleventh Conference on Machine Learning, red. W. Cohen, H. Hirsh, Morgan Kaufmann, New Brunswick, New Jersey, s. 70-77.

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

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

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:] Proceedings of the 14th International Joint Conference on Artificial Intelligence, red. C. Mellish, Morgan Kaufman, s. 1019-1024.

Rissanen J. (1978), *Modeling by shortest data description,* "Automatica", 14, s. 465-471.

Webb A.R. (2002), *Statistical pattern recognition*, 2nd edition, John Wiley & Sons.

THE COMPARISON OF RULES INDUCTION TO SOME DISCRIMINATION METHODS

Summary

Rules induction belongs to nonparametric and adaptive methods of discrimination. As in classification trees, it can deal with nonmetric variables and missing attribute values. The method is also robust in a presence of outliers. A model has the form of a set of "if-then" rules, where conditions are the conjunctions of attribute values, therefore it is easy for interpretation. Rules neither need be represented in the form of tree nor lead to disjoint classification.

The main goal of this paper is the comparison of error rates for rules induction and some discrimination methods. Over twenty real world datasets from UCI Repository of Machine Learning Databases were used. The RIPPER algorithm, which is considered as one of the most effective in rules induction, has been chosen.