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

Małgorzata Misztal
Uniwersytet Łódzki

ZAGREGOWANE I HYBRYDOWE MODELE DYSKRYMINACYJNE. PRÓBA PORÓWNANIA WYBRANYCH ALGORYTMÓW

1. Wstęp

Drzewo klasyfikacyjne jest graficzną prezentacją metody rekurencyjnego podziału *p*-wymiarowej przestrzeni cech na podzbiory rozłączne jednorodne z punktu widzenia wyróżnionej cechy *y*. W prowadzonych rozważaniach *y* jest zmienną nominalną reprezentującą klasy, do których należą obiekty zbioru uczącego, mamy zatem do czynienia z nieparametryczną analizą dyskryminacji (czyli z modelami dyskryminacyjnymi). Podstawową wadą drzew klasyfikacyjnych jest brak stabilności. Oznacza to, że mała zmiana wartości cech obiektów w zbiorze uczącym może prowadzić do powstania zupełnie innego modelu, co z kolei może wpływać na trafność prognostyczną uzyskanego modelu w przypadku analizy zbioru rozpoznawanego. Poprawę stabilności oraz dokładności predykcji można uzyskać, stosując modele złożone – zagregowane lub hybrydowe (por. np. [Stefanowski 2001]).

Model zagregowany to zbiór pojedynczych klasyfikatorów, których odpowiedzi są zagregowane do jednej odpowiedzi całego systemu, przy czym klasyfikatory składowe mogą być homogeniczne lub różnorodne. Model hybrydowy natomiast integruje w fazie uczenia przynajmniej dwa różne modele. Motywacja budowy takich modeli opiera się na twierdzeniu NFL (*No Free Lunch*) [Wolpert, Macready 1997], z którego wynika, że dla każdego algorytmu uczącego istnieje pewna klasa problemów, dla których jest skuteczny. W odniesieniu do pozostałych problemów algorytm może być mniej skuteczny od innych algorytmów (por. [Stefanowski 2001]).

Celem pracy jest próba porównania i oceny przydatności w zastosowaniach praktycznych zagregowanych modeli drzew klasyfikacyjnych: *Bagging* [Breiman 1996], *Boosting* [Freund, Schapire 1997], *Random Forests* [Breiman 2001], oraz wybranych modeli hybrydowych: CRUISE [Kim, Loh 2003], LOTUS [Chan, Loh 2004], PLUS [Lim 2000], k-NN Tree [Buttrey, Karo 2002].

2. Modele złożone

Agregacja modeli, najogólniej ujmując, polega na wyodrębnieniu ze zbioru uczącego V prób uczących, na podstawie których budowane są modele drzew $D_1(\mathbf{x})$, …, $D_v(\mathbf{x})$ łączone w kolejnym kroku w jeden model zagregowany $D^*(\mathbf{x})$. Określenie wartości zmiennej zależnej y odbywa się najczęściej na podstawie zasady majoryzacji:

$$\hat{D}^*(\mathbf{x}) = \arg\max_y \left\{ \sum_{v=1}^{V} I(\hat{D}_v(\mathbf{x}) = y) \right\}. \tag{1}$$

Metody agregacji mogą być oparte na losowaniu ze zbioru uczącego kolejnych prób uczących (*bagging* i *boosting*) lub na losowym wyborze zmiennych do każdego z V modeli składowych (*Random Forests*).

W metodzie *bagging* (*Bootstrap Aggregation*) dokonywane jest losowanie ze zwracaniem obiektów ze zbioru uczącego do N-elementowych prób uczących U_1,…, U_v, przy czym waga każdego obiektu jest jednakowa i równa: $w_i = \frac{1}{N}$. Na podstawie każdej próby uczącej budowane jest drzewo klasyfikacyjne i zapamiętywana jest dla każdego obiektu ustalona przez model wartość zmiennej zależnej $\hat{y}_n$. Wyniki uzyskane na podstawie V prób uczących są następnie agregowane w jeden model, a obiekt przydzielany jest do klasy najczęściej wskazywanej przez modele składowe.

W metodzie *boosting* również losuje się ze zwracaniem obiekty ze zbioru uczącego do N-elementowych prób uczących, przy czym wagi obiektów ulegają zmianie. Wagi określają prawdopodobieństwo wyboru obiektu do próby uczącej U_v i zmieniają się w zależności od wyników klasyfikacji dla poprzedniej próby U_{v-1}. Innymi słowy – obiekty źle zaklasyfikowane przez model w kolejnej próbie uczącej dostają wyższą wagę. Wyniki klasyfikacji uzyskane dla V prób uczących są agregowane z wykorzystaniem ważonego głosowania.

Metoda *Random Forests* także jest oparta na losowaniu ze zwracaniem obiektów ze zbioru uczącego, jednak w każdym węźle budowanego drzewa losowane są niezależnie zmienne opisujące obiekty, z których najlepsza jest wybierana do podziału. Drzewo budowane jest bez przycinania. Dla każdego obiektu zapamiętuje się ustaloną przez model wartość zmiennej zależnej $\hat{y}_n$, a każdy obiekt przydzielany jest w końcowym kroku do klasy najczęściej wskazywanej przez V modeli składowych.

Szczegółowy opis metod agregacji modeli klasyfikacyjnych znaleźć można w pracach Gatnara i Walesiaka [*Metody statystycznej*… 2004], Gatnara [2008] oraz Misztal [2008].

W wyniku wykorzystania metod *bagging* i *Random Forests* uzyskuje się dużą liczbę niezależnych drzew, a redukcja błędu predykcji następuje przez uśrednianie błędów predykcji pojedynczych klasyfikatorów. W metodzie *boosting* tworzona jest sekwencja pojedynczych drzew, a każde drzewo wyjaśnia zmienność niewyjaśnioną przez wcześniejsze drzewa.

Modele hybrydowe budowane są, najogólniej biorąc, w dwóch krokach. W pierwszym kroku wykorzystuje się metodę rekurencyjnego podziału, tworząc drzewo klasyfikacyjne. Wybór zmiennych do podziału oraz najlepszego podziału może się odbywać metodą przeszukiwania wszystkich możliwych podziałów lub z wykorzystaniem testów statystycznych, przy czym nie jest konieczne uzyskanie w wyniku podziału jednorodnych podzbiorów. W drugim kroku w wyodrębnionych podzbiorach stosowany jest kolejny algorytm klasyfikacji – model regresji logistycznej, liniowa funkcja dyskryminacyjna, jeden z algorytmów minimalnoodległościowych itp.

W przypadku algorytmów LOTUS − *Logistic Regression Trees with Unbiased Selection* [Chan, Loh 2004] − i PLUS − *Polytomous Logistic Regression Trees with Unbiased Split* [Lim 2000] − w podzbiorach wyodrębnionych metodą rekurencyjnego podziału budowane są modele regresji logistycznej jednej lub wielu zmiennych:

$$P(Y=1)=\frac{1}{1+e^{-(\alpha_0+\alpha_1 X_1+\ldots+\alpha_k X_k)}}. \tag{2}$$

Algorytm CRUISE − *Classification Rule with Unbiased Interaction Selection and Estimation* [Kim, Loh 2003] − z kolei w każdym węźle uzyskanym za pomocą podziałów jednowymiarowych próbuje dopasować liniową funkcję dyskryminacyjną dla dwóch najlepszych zmiennych. Dobór tych zmiennych odbywa się z wykorzystaniem wielowymiarowej analizy wariancji (MANOVA) lub analizy dyskryminacji (LDF). W rezultacie w każdym węźle dostajemy oszacowania parametrów funkcji klasyfikujących dwóch zmiennych:

$$\tilde{u}_i(\mathbf{x}) = a_{i0} + a_{i1}X_1 + a_{i2}X_2 . \tag{3}$$

Porównując parami funkcje klasyfikujące, otrzymuje się równania powierzchni decyzyjnych (funkcji dyskryminacyjnych) o postaci:

$$\tilde{u}_i(\mathbf{x}) - \tilde{u}_j(\mathbf{x}) = 0 . \tag{4}$$

Wykorzystanie w modelu dyskryminacyjnym jedynie dwóch najlepszych zmiennych predykcyjnych pozwala dodatkowo przedstawić graficznie uzyskane podziały.

W algorytmie k-NN Tree [Buttrey, Karo 2002] obiekty ze zbioru testowego przypisuje się zgodnie z regułami klasyfikacyjnymi do odpowiedniego liścia utworzonego w pierwszym kroku drzewa klasyfikacyjnego, a następnie określa się ich przynależność do klasy za pomocą reguły *k* najbliższych sąsiadów, obliczając odległości obiektu od tych obiektów uczących, które znalazły się w danym liściu.

Szczegółowy opis wymienionych algorytmów hybrydowych znajduje się w pracach Lima [2000], Chana i Loha [2004], Buttreya i Karo [2002] oraz Kima i Loha [2003]. Komputerowe implementacje wymienionych algorytmów można zaś znaleźć na stronach internetowych: www.stat.wisc.edu/~kinyee/lotus; www.recursive-partitioning.com/plus; www.stat.wisc.edu/~loh/cruise oraz w środowisku R (pakiety: `ada`, `randomForest`, `ipred`, `kkNN`).

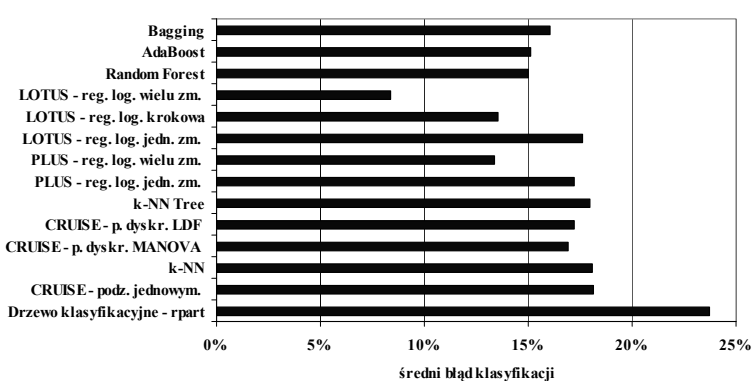
Ocena działania wspomnianych modeli zagregowanych i hybrydowych została przez ich autorów dokonana z wykorzystaniem zbiorów danych z *UCI Repository of Machine Learning Datasets* [Blake, Keogh, Merz 1988].

W wyniku przeprowadzonych eksperymentów Breiman [1996] wykazał, że błąd modelu zagregowanego metodą *bagging* jest niższy od przeciętnego błędu modeli bazowych. Metoda *boosting* daje lepsze modele dyskryminacyjne niż *bagging* (por. [Gatnar 2008]), ale jest wrażliwa na występowanie szumów w zbiorze uczącym. *Random Forests* z kolei pozwala uzyskać dokładność predykcji porównywalną z metodą *boosting*, a przy tym działa szybciej niż obie wymienione wcześniej metody agregacji (por. [Breiman 2001]).

Algorytm CRUISE z podziałami dyskryminacyjnymi daje zwykle mniejszy błąd klasyfikacji niż pojedyncze drzewo uzyskane w wyniku podziałów jednowymiarowych (por. [Kim, Loh 2003]).

Chan i Loh [2004] pokazali, że algorytm LOTUS poprawia dokładność predykcji o kilkanaście procent w stosunku do modelu regresji logistycznej. Lim [2000] wykazał zaś, że PLUS daje wyniki porównywalne z modelami zagregowanymi w przypadku zbiorów danych bez brakujących wartości. k-NN Tree pozwala natomiast w większości przypadków uzyskać wyniki nie gorsze niż algorytmy minimalnoodległościowe oraz pojedyncze drzewa z podziałami jednowymiarowymi (por. [Buttrey, Karo 2002]).

Porównanie dokładności predykcji wszystkich wymienionych algorytmów wymaga wyboru do badań tych zbiorów danych, w których występują tylko dwie klasy (ze względu na algorytm LOTUS) oraz zmienne objaśniające są mierzone na skali co najmniej porządkowej (ze względu na wybór miary odległości w algorytmie k-NN Tree). Do wstępnych analiz wybrano trzy zbiory z *UCI Repository of Machine Learning Datasets* [Blake, Keogh, Merz 1988]: *Breast Cancer Wisconsin* (n = 699), *Pima Indians Diabetes* (n = 768) oraz *Sonar, Mines vs. Rocks* (n = 208). Średni błąd klasyfikacji (dla trzech analizowanych zbiorów) przedstawia rys. 1.



Rys. 1. Średni błąd klasyfikacji badanych algorytmów

Źródło: obliczenia własne.

Łatwo zauważyć, że zdecydowanie najgorsze wyniki w sensie dokładności predykcji uzyskujemy dla pojedynczego drzewa klasyfikacyjnego wygenerowanego za pomocą procedury `rpart`. Pozostałe indywidualne klasyfikatory (k-NN, CRUISE z podziałami jednowymiarowymi) również nie dają zadowalających wyników.

Najlepiej w przypadku analizowanych trzech zbiorów wypada algorytm LOTUS z modelami regresji wielu zmiennych w liściach drzewa. Oczywiście wyciąganie wiążących wniosków wymaga dokonania obliczeń dla znacznie większej liczby zbiorów danych.

3. Przykład zastosowań

Przedstawione modele złożone wykorzystano do wspomagania procesów decyzyjnych w diagnostyce medycznej.

Materiał badawczy stanowi 300 pacjentów po zabiegu wymiany zastawki aortalnej opisanych zestawem 9 cech przedoperacyjnych: BMI (wskaźnik masy ciała w kg/m^2), płeć, wiek (w latach), EF (frakcja wyrzutowa lewej komory w %), wymiar lewej komory w skurczu i w rozkurczu (w cm), grubość przegrody w skurczu i w rozkurczu (w cm) oraz wymiar lewego przedsionka (w cm); por. [Misztal, Banach 2008].

Najczęściej występującym powikłaniem po zabiegach kardiologicznych jest migotanie przedsionków – AF (*Atrial Fibrillation*), zatem zmienną zależną jest zmienna binarna określająca przynależność pacjenta do jednej z dwóch klas: 0 – (bez AF) oraz 1 (z AF).

Grupę 300 pacjentów w sposób losowy podzielono na próbę uczącą ($n_u = 150$) oraz próbę testową ($n_t = 150$). Modele budowano na podstawie próby uczącej, a ich zdolność predykcyjną oceniano z wykorzystaniem obiektów z próby testowej. Uzyskane wyniki przedstawia tab. 1. Jak widać, problem klasyfikacji badanych pacjentów nie należy do łatwych.

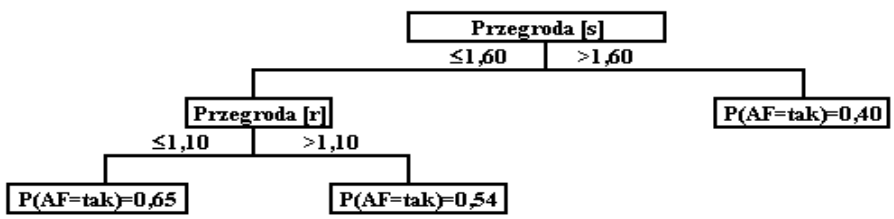
Tabela 1. Odsetek błędnych klasyfikacji w próbie testowej

Klasyfikator	% błędnych klasyfikacji (próba testowa)		
	ogółem	AF = tak	AF = nie
Regresja logistyczna	39,33	56,16	23,38
Drzewo klasyfikacyjne `rpart`	46,00	49,32	42,86
Drzewo klasyfikacyjne CRUISE	29,33	16,44	41,56
k-NN Tree (NN z odległością Euklidesa)	28,00	19,18	36,36
PLUS − regresja jednej zmiennej	38,00	56,16	20,78
LOTUS − regresja jednej zmiennej	34,00	42,47	25,97
LOTUS − regresja logistyczna krokowa	38,00	56,16	20,78
LOTUS − regresja wielu zmiennych	36,67	57,53	16,88
CRUISE − podziały dyskryminacyjne (MANOVA)	33,33	28,77	37,66
CRUISE − podziały dyskryminacyjne (LDF)	29,33	27,40	31,17
Bagging (100 drzew)	26,67	27,40	25,97
Boosting (100 drzew)	30,67	36,99	24,68
Random Forest (100 drzew)	28,00	30,14	25,97

Źródło: obliczenia własne.

Drzewo klasyfikacyjne utworzone za pomocą procedury `rpart` ma 8 liści i jest łatwe w interpretacji, ale daje zdecydowanie największy błąd klasyfikacji dla zbioru testowego. Zmniejszenie odsetka błędnych klasyfikacji otrzymuje się dla algorytmu CRUISE z podziałami jednowymiarowymi, ale w tym przypadku drzewo klasyfikacyjne ma aż 28 liści. Dla modeli zagregowanych najmniejszy błąd klasyfikacji daje zaś metoda *bagging*.

Algorytm LOTUS z modelem regresji logistycznej jednej zmiennej generuje drzewo przedstawione na rys. 2. Oszacowania parametrów w poszczególnych liściach (por. tab. 2) są najczęściej istotne statystycznie.

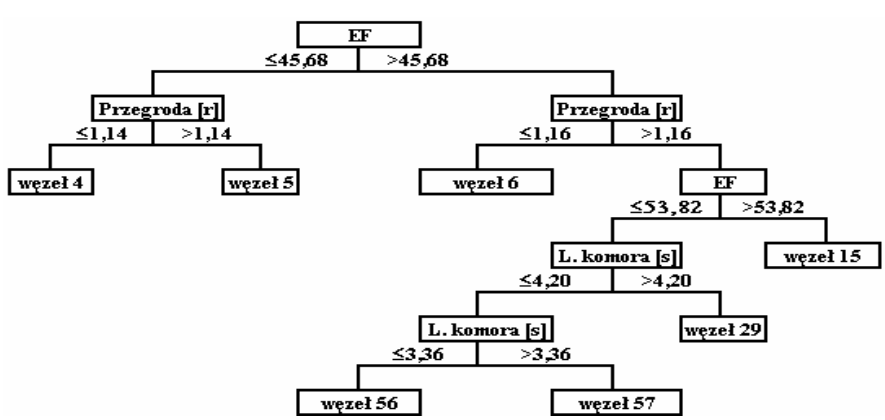


Rys. 2. Drzewo regresji logistycznej (LOTUS)
Źródło: opracowanie własne.

Tabela 2. Oszacowania parametrów modeli regresji logistycznej w liściach drzewa

Nr węzła	$P(Y = 1)$	Zmienna	Ocena współczynnika	Wartość *p*
3	0,40	wyraz wolny	2,411	0,0548
		EF	–0,056	0,0237
4	0,65	wyraz wolny	0,619	0,2024
		x	x	x
5	0,54	wyraz wolny	–31,946	0,0338
		przegroda [*r*]	24,985	0,0325

Źródło: obliczenia własne.



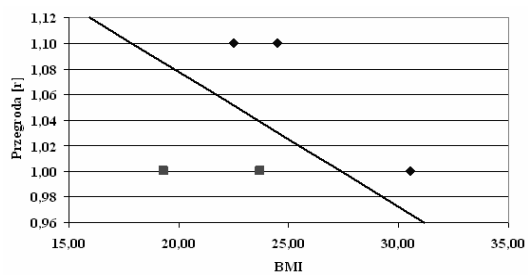
Rys. 3. Drzewo klasyfikacyjne – algorytm CRUISE z podziałami dyskryminacyjnymi (LDF)
Źródło: opracowanie własne.

Tabela 3. Oszacowania parametrów funkcji klasyfikacyjnych – węzeł 4

Grupa	Nazwa zmiennej		
	BMI	Przegroda [rozkurcz]	Wyraz wolny
AF = nie	36,249	3518,3	–2344,9
AF = tak	33,491	3256,0	–2007,1

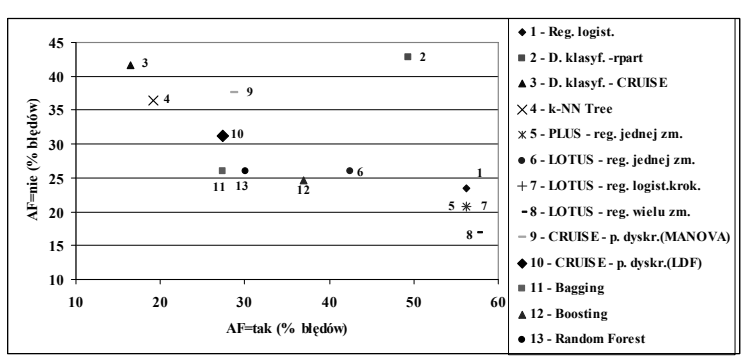
Źródło: obliczenia własne.

Drzewo uzyskane z wykorzystaniem algorytmu CRUISE z podziałami dyskryminacyjnymi przedstawia rys. 3. W każdym z węzłów końcowych oszacowano parametry liniowych funkcji klasyfikacyjnych. Przykładowe oszacowania (dla węzła nr 4) przedstawia tab. 3, a graficzną prezentację podziału dyskryminacyjnego – rys. 4.



Rys. 4. Powierzchnia decyzyjna – węzeł 4
Źródło: opracowanie własne.

W diagnostyce medycznej przydatność modeli złożonych należy oceniać nie tylko z punktu widzenia błędnych klasyfikacji ogółem, ale również ze względu na błędy klasyfikacji w grupach. W rozważanym przykładzie bardziej „kosztowne" jest błędne rozpoznanie pacjenta zagrożonego wystąpieniem AF. Odsetki błędnych klasyfikacji w analizowanych grupach pacjentów przedstawia rys. 5.



Rys. 5. Odsetek błędnych klasyfikacji w próbie testowej w analizowanych grupach pacjentów
Źródło: opracowanie własne na podstawie wyników przedstawionych w tab. 1

Jak widać, drzewo klasyfikacyjne `rpart` źle rozpoznaje pacjentów z obu grup; żaden inny algorytm nie daje podobnych wyników. Bardzo wysokie odsetki błędnych klasyfikacji wśród pacjentów zagrożonych migotaniem przedsionków uzyskujemy w odniesieniu do modelu regresji logistycznej oraz drzew regresji logistycznej (PLUS i LOTUS). Algorytmy te jednak poprawnie klasyfikują pacjentów z grupy bez AF.

Drzewo klasyfikacyjne CRUISE z podziałami jednowymiarowymi oraz algorytm k-NN Tree wykazują stosunkowo niskie odsetki błędnych klasyfikacji w grupie pacjentów z AF, natomiast źle rozpoznają pacjentów niezagrożonych wystąpieniem AF.

Mniej więcej podobne i stosunkowo niewielkie odsetki błędów w obu grupach uzyskujemy dla modeli zagregowanych metodą *bagging* oraz *Random Forests*.

4. Uwagi końcowe

W większości zadań klasyfikacyjnych lepsze wyniki (w sensie dokładności predykcji) otrzymujemy dla modeli zagregowanych i hybrydowych niż dla pojedynczych drzew. Jednakże pojedyncze drzewa są intuicyjnie zrozumiałe i łatwo na ich podstawie zapisać reguły klasyfikacyjne. W modelach hybrydowych należy się liczyć z możliwością wystąpienia problemów z prostym zapisem reguł decyzyjnych oraz z ich interpretacją (np. podziały dyskryminacyjne w algorytmie CRUISE). Model zagregowany jest natomiast swego rodzaju czarną skrzynką − nie ma możliwości zapisu reguł klasyfikacyjnych.

W przypadku takich samych błędów klasyfikacji dla modelu hybrydowego i pojedynczego drzewa przewaga algorytmu hybrydowego zazwyczaj wyraża się w redukcji wielkości drzewa. Na przykład drzewo klasyfikacyjne CRUISE dla pacjentów z AF ma 28 liści, a drzewo klasyfikacyjne CRUISE z podziałami dyskryminacyjnymi − tylko 7.

W odniesieniu do modeli hybrydowych obserwuje się poprawę dokładności predykcji dla niejednorodnych zbiorowości − za pomocą podziałów jednowymiarowych dokonuje się podziału zbioru na bardziej jednorodne podzbiory, w których następnie stosowany jest inny klasyfikator. Dodatkowo do podziałów mogą być wykorzystane zmienne (np. jakościowe), których nie da się wykorzystać np. w modelu regresji logistycznej czy algorytmie NN.

Oceniając przydatność różnych algorytmów klasyfikacji, należy pamiętać o twierdzeniu NFL [Wolpert, Macready 1997]. Nie istnieje uniwersalny, lepszy od innych algorytm do wszystkich zadań klasyfikacyjnych, a średnie zachowanie algorytmu dla wszystkich zadań jest takie samo. Zatem jeśli klasyfikator X jest lepszy od klasyfikatora Y w przypadku zadania A, to w przypadku zadania B sytuacja może być odwrotna. Warto więc pamiętać, że optymalne wyniki klasyfikacji można uzyskać, mając wiedzę o modelowanym zjawisku i znając jego strukturę [Gatnar 2008].

Literatura

Blake C., Keogh E., Merz C.J. (1988), *UCI repository of machine learning datasets*, Department of Information and Computer Science, University of California, Irvine.

Breiman L. (1996), *Bagging predictors*, "Machine Learning", 24, s. 123-140.
Breiman L. (2001), *Random forests*, "Machine Learning", 45, s. 5-32.
Buttrey S.E., Karo C. (2002), *Using k-nearest-neighbor classification in the leaves of a tree*, "Computational Statistics & Data Analysis", 40, s. 27-37.
Chan K.-Y., Loh W.-Y. (2004), *LOTUS: an algorithm for building accurate and comprehensible logistic regression trees,* "Journal of Computational and Graphical Statistics", 13 (4), s. 826-852.
Freund Y., Schapire R.E. (1997), *A decision-theoretic generalization of on-line learning and an application to boosting*, "Journal of Computer and System Sciences", 55, s. 119-139.
Gatnar E. (2008), *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*, PWN, Warszawa.
Kim H., Loh W.-Y. (2003), *Classification trees with bivariate linear discriminant node models*, "Journal of Computational and Graphical Statistics", 12, s. 512-530.
Lim T.-S. (2000), *Polytomous logistic regression trees*, PhD Thesis, Department of Statistics, University of Wisconsin, Madison.
Metody statystycznej analizy wielowymiarowej w badaniach marketingowych (2004), red. E. Gatnar, M. Walesiak, AE, Wrocław.
Misztal M. (2008), *Zagregowane modele dyskryminacyjne i regresyjne w prognozowaniu czasu pobytu na OIOM pacjentów z chorobą wieńcową*, [w:] Taksonomia 15, *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 7 (1207), UE, Wrocław, s. 316-322.
Misztal M., Banach M. (2008), *On distance-based algorithms in medical applications*, Acta Uniwersitatis Lodziensis, Folia Oeconomica, Wydawnictwo Uniwersytetu Łódzkiego, Łódź, w druku.
Stefanowski J. (2001), *Algorytmy indukcji reguł decyzyjnych w odkrywaniu wiedzy*, Wydawnictwo Politechniki Poznańskiej, Rozprawy nr 361, Poznań.
Wolpert D.H., Macready W.G. (1997), *No free lunch theorems for optimization*, IEEE Transactions on Evolutionary Computation, 1(1), s. 62-68.

AGGREGATED AND HYBRID DISCRIMINANT MODELS. AN ATTEMPT TO COMPARE SELECTED ALGORITHMS

Summary

To improve the stability and prediction accuracy of classification trees we can use ensembles of classifiers or hybrid models, combining recursive partitioning with some others algorithms (i.e. linear discriminant functions, logistic regression, distance-based algorithms, etc.).

The aim of the paper is to compare the performances of classifier combination methods (*Bagging* [Breiman 1996], *Boosting* [Freund, Shapire 1997], *Random forests* [Breiman 2001]) and hybrid models (CRUISE [Kim, Loh 2003], LOTUS [Chan, Loh 2004], PLUS [Lim 2000], k-NN Tree [Buttrey, Karo 2002]). A medical diagnosis example is used to demonstrate the advantages and disadvantages of the algorithms examined.