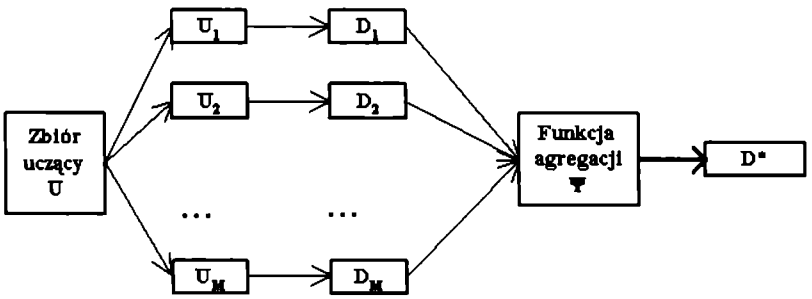
PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 7 (1207) 2008
TAKSONOMIA 15
Klasyfikacja i analiza danych – teoria i zastosowania

Eugeniusz Gatnar
Akademia Ekonomiczna w Katowicach

WPŁYW METODY ŁĄCZENIA MODELI NA WIELKOŚĆ BŁĘDU KLASYFIKACJI W PODEJŚCIU WIELOMODELOWYM

1. Wstęp

W podejściu wielomodelowym M modeli bazowych $D_1, ..., D_M$ jest łączonych w jeden model zagregowany D^*. Model ten, przy spełnieniu odpowiednich założeń, charakteryzuje się większą dokładnością klasyfikacji niż którykolwiek z modeli indywidualnych. Na rys. 1 przedstawiono kolejne etapy budowy modelu D^*.



Rys. 1. Etapy budowy modelu zagregowanego D^*
Źródło: opracowanie własne.

Budowa modeli bazowych jest oparta na zbiorze M prób uczących $U_1, ..., U_M$, będących podzbiorami oryginalnego zbioru uczącego U. Podzbiory te mogą zawierać albo wybrane obserwacje ze zbioru U, albo wszystkie obserwacje, lecz rzutowane na różne podprzestrzenie zmiennych.

Łączenie modeli bazowych realizuje funkcja Ψ określona na zbiorze ich wyników predykcji:

$$\hat{D}^*(\mathbf{x}_i) = \Psi\left(\hat{D}_1(\mathbf{x}_i), \ldots, \hat{D}_M(\mathbf{x}_i)\right), \quad (1)$$

przy czym jej rodzaj zależy od postaci wyników predykcji modeli bazowych $D_1, \ldots, D_M$. Funkcja ta nadaje wyższe wagi tym modelom, które charakteryzują się większą dokładnością, niż tym, które generują większy błąd predykcji.

Ponieważ w literaturze przedmiotu zaproponowano kilkanaście różnych funkcji Ψ, powstaje potrzeba zbadania wpływu postaci funkcji na błąd predykcji modelu zagregowanego (1). To zagadnienie jest właśnie podstawowym przedmiotem rozważań w niniejszym artykule.

2. Wyniki klasyfikacji

Modele bazowe, które podlegają agregacji, w przypadku analizy dyskryminacyjnej mogą generować wyniki predykcji w postaci:

- nazwy lub numeru klasy, do której należy obserwacja $\mathbf{x}_i$,
- wektora prawdopodobieństw *a posteriori* dla zbioru klas $C_1, \ldots, C_J$.

W pierwszym przypadku każdy z modeli $D_1, \ldots, D_M$ wskazuje dla każdej obserwacji pojedynczą klasę, tj. $\hat{D}_m(\mathbf{x}_i) = C_j$, co w sumie (dla wszystkich modeli) daje wektor klas w postaci $\mathbf{d} = [d_1, \ldots, d_M]$, gdzie $d_m = \hat{D}_m(\mathbf{x}_i)$. W programie **R** wartości d_m można uzyskać, stosując polecenie `predict` z parametrem `type="class"`.

Drugi rodzaj wyników predykcji modeli bazowych stanowi wektor $\hat{D}_m(\mathbf{x}_i) = [p_{m,1}(\mathbf{x}_i), \ldots, p_{m,J}(\mathbf{x}_i)]$, gdzie $p_{m,j}(\mathbf{x}_i)$ oznacza prawdopodobieństwo *a posteriori* $p_{m,j}(\mathbf{x}_i) = P(C_j \mid \mathbf{x}_i)$ wyznaczane przez model D_m. W programie **R** wartości te można uzyskać, stosując polecenie `predict` z parametrem `type="prob"`.

Jak widać, ten drugi rodzaj wyników predykcji można łatwo sprowadzić do tej pierwszej postaci, znajdując klasę, dla której prawdopodobieństwo *a posteriori* jest największe.

3. Funkcje agregacji

Jak już wspomniano we wstępie, postać funkcji agregacji Ψ zależy od rodzaju wyników predykcji modeli bazowych $D_1, \ldots, D_M$. Zostaną przedstawione dwa rodzaje tych wyników: dla pojedynczej klasy oraz dla wektora prawdopodobieństw *a posteriori*.

3.1. Funkcje dla pojedynczej klasy

W analizie dyskryminacyjnej modele bazowe najczęściej wskazują klasę dla obserwacji poprzez „głosowanie". Oznacza to, że model zagregowany $\hat{D}^*(\mathbf{x}_i)$ przydziela obserwację $\mathbf{x}_i$ do tej klasy, którą wskazała największa liczba modeli bazowych $D_1, ..., D_M$:

$$\hat{D}^*(\mathbf{x}_i) = \arg\max_j \sum_{m=1}^{M} I(\hat{D}_m(\mathbf{x}_i) = C_j). \tag{2}$$

Powstaje jednak problem remisów, np. w przypadku parzystej liczby modeli bazowych i parzystej liczby klas. Wtedy najczęściej klasę określa się losowo.

Pozostałe metody agregacji polegają na obliczeniu dla każdej klasy wartości tzw. wskaźnika dyskryminacji W_j, a następnie przydzieleniu obserwacji $\mathbf{x}_i$ do tej klasy C_j, dla której uzyskano największą wartość tego wskaźnika:

$$\hat{D}^*(\mathbf{x}_i) = \arg\max_j \left\{ W_j(\mathbf{x}_i) \right\}. \tag{3}$$

Gdy modele bazowe $D_1, ..., D_M$ są niezależne, to można wykorzystać klasyfikator bayesowski, w sposób zaproponowany przez Domingosa i Pazzaniego [1997]. Wtedy wskaźnik dyskryminacji wyznacza się jako:

$$W_j(\mathbf{x}_i) = p(C_j) \cdot p(d_1, ..., d_M \mid C_j), \tag{4}$$

co w istocie jest licznikiem wzoru na prawdopodobieństwo *a posteriori*:

$$p(C_j \mid d_1, ..., d_M) = \frac{p(C_j) \cdot p(d_1, ..., d_M \mid C_j)}{p(d_1, ..., d_M)}. \tag{5}$$

W formułach (4) oraz (5) d_m (dla $m = 1, ..., M$) oznacza wynik predykcji, czyli klasę, do której model bazowy D_m przydziela obserwację $\mathbf{x}_i$.

Wartość wskaźnika (4) jest najczęściej estymowana na podstawie wyników klasyfikacji modelu D_m, zamieszczonych w dwuwymiarowej tablicy kontyngencji $\mathbf{Z}^m = \left[Z^m_{j,d_m} \right]$. Na jej głównej przekątnej znajdują się liczebności obserwacji poprawnie sklasyfikowanych przez ten model, tj. faktycznie należących do kolejnych klas $C_1, ..., C_J$ i przydzielonych do tych klas przez model D_m. Poza przekątną są zaś liczebności obserwacji sklasyfikowanych błędnie.

Z kolei prawdopodobieństwo *a priori* $p(C_j)$ jest szacowane za pomocą frakcji $p(C_j) = N_j / N$, a prawdopodobieństwo warunkowe jako:

$$p(\mathbf{d} \mid C_j) = \prod_{m=1}^{M} \frac{Z_{j,d_m}^m}{N_j}. \tag{6}$$

Wtedy wartość wskaźnika dyskryminacji dla klasy C_j można obliczyć jako:

$$W_j(\mathbf{x}_i) = \frac{N_j}{N} \prod_{m=1}^{M} \frac{Z_{j,d_m}^m}{N_j}. \tag{7}$$

Do łączenia modeli bazowych można także zastosować metodę zaproponowaną przez Huanga i Suena [1995], którą autorzy nazwali *Behavior-Knowledge Space* (BKS). Nazwa ta pochodzi stąd, że decyzje poszczególnych modeli bazowych $D_1, ..., D_M$ co do przynależności obserwacji $\mathbf{x}_i$ są zapamiętywane w tablicy imitującej M-wymiarową przestrzeń. Można więc przyjąć, że tablica BKS zawiera wiedzę o zachowaniu modeli dyskryminacyjnych. W tab. 1 znajduje się przykład tablicy BKS dla $M = 2$.

Tabela 1. Tablica BKS dla $M = 2$

D_1 / D_2	1	...	J
1	$N_{1,1}(1), T_{1,1}, R_{1,1}$	...	$N_{1,J}(1), N_{1,J}(J), T_{1,J}, R_{1,J}$
...	...	...	...
J	$N_{J,1}(1), N_{J,1}(J), T_{1,1}, R_{1,1}$	...	$N_{J,J}(J), T_{J,J}, R_{J,J}$

Źródło: opracowanie własne.

Każdy model bazowy D_m przydziela obserwację $\mathbf{x}_i$ do jednej z klas $C_1, ..., C_J$, zatem w każdym wymiarze tablicy znajduje się J komórek[1]. W sumie tablica BKS składa się z $M \times J \times J$ komórek, które są hiperkostkami o współrzędnych $BKS(d_1, ..., d_M)$. W każdej komórce zaś zawarte są następujące informacje:

- liczba obserwacji należących do poszczególnych klas:

$$N_{d_1,...,d_M}(j), \tag{8}$$

- liczba wszystkich obserwacji w komórce:

$$T_{d_1,...,d_M} = \sum_{j+1}^{J} N_{d_1,...,d_M}(j), \tag{9}$$

- klasa dominująca (najliczniejsza): $R_{d_1,...,d_M}$,

którą określa się za pomocą głosowania:

$$R_{d_1,...,d_M} = \arg\max_j \{N_{d_1,...,d_M}(j)\}, \tag{10}$$

remisy zaś są rozstrzygane arbitralnie (np. losowo).

[1] W oryginalnej propozycji Huang i Suen [1995] znajduje się jeszcze dodatkowa klasa $J + 1$. Są do niej przydzielane te obserwacje, których modele składowe nie potrafią jednoznacznie sklasyfikować.

Decyzja modelu zagregowanego $D^*(\mathbf{x}_i)$ o przynależności obserwacji $\mathbf{x}_i$ do jednej z klas $C_1, ..., C_J$ jest podejmowana na podstawie reguły:

$$\hat{D}^*(\mathbf{x}_i) = R_{d_1,...,d_M}, \tag{11}$$

dla $T_{d_1,...,d_M} > 0$ oraz:

$$\frac{N_{d_1,...,d_M}(R_{d_1,...,d_M})}{T_{d_1,...,d_M}} \geq \lambda, \tag{12}$$

gdzie parametr $\lambda \in [0, 1]$ steruje jakością predykcji.

Wernecke [1992] zaproponował podobną metodę, lecz zamiast liczebności w komórkach tablicy BKS znajdują się przedziały ufności dla liczebności obserwacji należących do poszczególnych klas ($N_{d_1,...,d_M}(j)$) dla poziomu ufności 0,95. Jeżeli przedziały te są nierozłączne, to model zagregowany wykorzystuje wynik predykcji modelu, dla którego prawdopodobieństwo:

$$p(D_m(\mathbf{x}_i) \neq y \wedge D_m(\mathbf{x}_i) = d_m) \tag{13}$$

jest najmniejsze.

Przedział ufności, na poziomie ufności $\alpha = 0{,}95$, dla liczebności $\hat{k}_j$, w pewnej komórce tablicy BKS ma postać:

$$\left(k_j - 1{,}96\sqrt{\frac{k_j(k-k_j)}{k}} + 0{,}5; k_j + 1{,}96\sqrt{\frac{k_j(k-k_j)}{k}} - 0{,}5 \right), \tag{14}$$

gdzie $k_j = N_{d_1,...,d_M}(j)$ to liczba obserwacji należących do klasy C_j w komórce tablicy BKS $(d_1, ..., d_M)$, zaś $k = \sum_{j=1}^{J} k_j$.

3.2. Funkcje dla wektora prawdopodobieństw *a posteriori*

Wynik predykcji modelu bazowego D_m dla obserwacji $\mathbf{x}_i$ może być także wektorem prawdopodobieństw *a posteriori* dla zbioru klas. Jeżeli taki wektor jest generowany przez wszystkie modele bazowe, to powstaje macierz o wymiarach $M \times J$, którą Kuncheva i in. [2001] nazywają profilem decyzyjnym. Macierz ta ułatwia zapis wyników predykcji modeli, które podlegają łączeniu, i ma postać:

$$p(\mathbf{x}_i) = [p_{m,j}(\mathbf{x}_i)] = \begin{bmatrix} p_{1,1}(\mathbf{x}_i) & \dots & p_{1,J}(\mathbf{x}_i) \\ \dots & \dots & \dots \\ p_{M,1}(\mathbf{x}_i) & \dots & p_{M,J}(\mathbf{x}_i) \end{bmatrix}. \tag{15}$$

Każdy jej wiersz to wynik predykcji modelu D_m, tj. $\hat{D}_m(\mathbf{x}_i) = [p_{m,1}(\mathbf{x}_i), ..., p_{m,J}(\mathbf{x}_i)]$, zaś każda kolumna to wektor prawdopodobieństw *a posteriori* dla klasy C_j, generowanych przez M modeli bazowych, tj. $[p_{1,j}(\mathbf{x}_i), ..., p_{M,j}(\mathbf{x}_i)]^T$.

Każda z metod łączenia wyników predykcji w postaci profilu decyzyjnego (15) wyznacza dla klasy C_j wartość wskaźnika dyskryminacji W_j. Aby uzyskać wynik predykcji modelu zagregowanego w postaci nazwy klasy, należy zastosować regułę (3).

Prosta metoda agregacji może wykorzystywać wartość maksymalną:

$$W_j(\mathbf{x}_i) = \max_m \{p_{m,j}(\mathbf{x}_i)\}, \tag{16}$$

co oznacza wskazanie klasy o największej wartości prawdopodobieństwa *a posteriori* uzyskanej w wyniku predykcji modeli bazowych. Istnieje także metoda wykorzystująca wartość najmniejszą:

$$W_j(\mathbf{x}_i) = \min_m \{p_{m,j}(\mathbf{x}_i)\}, \tag{17}$$

oraz medianę:

$$W_j(\mathbf{x}_i) = \operatorname{med}_m \{p_{m,j}(\mathbf{x}_i)\}. \tag{18}$$

Należy jednak zaznaczyć, że metody (16) oraz (17) są bardzo czułe na asymetrię wartości prawdopodobieństw $p_{1,j}(\mathbf{x}_i), ..., p_{M,j}(\mathbf{x}_i)$.

Następna grupa metod łączenia wyników predykcji wykorzystuje operatory arytmetyczne. Pierwszy z nich polega na wyznaczeniu średniej:

$$W_j(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} p_{m,j}(\mathbf{x}), \tag{19}$$

która jest wrażliwa na wartości ekstremalne, tj. bardzo wysokie i bardzo niskie wartości prawdopodobieństw $p_{m,j}(\mathbf{x})$. Wtedy można usunąć wyniki skrajnie optymistycznych i pesymistycznych modeli, np. 10% wartości najwyższych i najniższych.

Interesującą formułą agregacji jest iloczyn prawdopodobieństw:

$$W_j(\mathbf{x}) = \frac{1}{M} \prod_{m=1}^{M} p_{m,j}(\mathbf{x}), \tag{20}$$

który jednak jest bardzo wrażliwy na wyniki predykcji najbardziej pesymistycznego modelu bazowego. Inaczej mówiąc, jeżeli jakiś model D_m generuje małą wartość prawdopodobieństwa dla klasy C_j, to ma ona małe szanse stać się wynikiem

predykcji modelu zagregowanego. Jeżeli jednak prawdopodobieństwa $p_{m,j}(\mathbf{x}_i)$ są dobrze oszacowane, to (20) jest najlepszym estymatorem prawdopodobieństwa *a posteriori* $p(C_j \mid \mathbf{x}_i)$.

Inny sposób łączenia modeli zaproponowali Kuncheva i in. [2001]. Został on oparty na tzw. szablonach decyzyjnych (*decision templates*), tworzonych dla poszczególnych klas, wyznaczone jako średnie wartości z profilów decyzyjnych (15) dla kolejnych obserwacji ze zbioru uczącego, które należą do klasy C_j:

$$S_j = \frac{1}{N_j} \sum_{\mathbf{x}_i \in Z_j} P(\mathbf{x}_i), \tag{21}$$

przy czym Z_j jest zbiorem obserwacji ze zbioru uczącego, które należą do klasy C_j, N_j zaś jest liczebnością tego zbioru. Szablon dla tej klasy ma postać:

$$S_j = [s_{m,k}(j)] = \begin{bmatrix} s_{1,1}(j) & \dots & s_{1,J}(j) \\ \dots & \dots & \dots \\ s_{M,1}(j) & \dots & s_{M,J}(j) \end{bmatrix}, \tag{22}$$

gdzie $s_{m,k}(j)$ można zinterpretować jako przeciętne prawdopodobieństwo, obliczone na podstawie obserwacji należących do klasy C_j, tego, że model D_m przydziela obserwacje do klasy C_k. Maksymalną wartość prawdopodobieństwo to osiąga wtedy, gdy $k = j$.

Dla obserwacji $\mathbf{x}_i$ ze zbioru rozpoznawanego najpierw zostanie zbudowany jej profil decyzyjny (15), a następnie obliczona wartość pewnej miary odległości δ pomiędzy $P(\mathbf{x}_i)$ a szablonami decyzyjnymi poszczególnych klas. Zwykle wykorzystuje się w tym celu kwadrat odległości euklidesowej:

$$\delta\left(P(\mathbf{x}_i), S_j\right) = \frac{1}{M \cdot J} \sum_{m=1}^{M} \sum_{k=1}^{J} [s_{m,k}(j) - p_{m,k}(\mathbf{x}_i)]^2 . \tag{23}$$

Grabisch [1995] zaproponował zastosowanie w tym celu symetrycznej różnicy, zaczerpniętej z teorii zbiorów rozmytych:

$$\delta\left(P(\mathbf{x}_i), S_j\right) = \frac{1}{M \cdot J} \sum_{m=1}^{M} \sum_{k=1}^{J} \max\left\{\min\{s_{m,k}(j), 1 - p_{m,k}(\mathbf{x}_i)\},\ \min\{1 - s_{m,k}(j), p_{m,k}(\mathbf{x}_i)\}\right\}. \tag{24}$$

Wtedy wskaźnik dyskryminacji dla klasy C_j można wyrazić jako podobieństwo:

$$W_j(\mathbf{x}_i) = 1 - \delta(P(\mathbf{x}_i), S_j). \tag{25}$$

4. Zbiór danych

Wykorzystany w eksperymentach zbiór EUROPA dotyczy państw Europy i jest kompilacją danych pochodzących z Banku Światowego oraz z bazy danych AMECO, tworzonej i utrzymywanej przez Komisję Europejską.

Bank Światowy klasyfikuje kraje na cztery grupy ze względu na poziom dochodu narodowego brutto na głowę mieszkańca (GNI *per capita*[2]):

- H – wysoki poziom (10 726 dol. i więcej),
- UM – średnio wysoki (3466-10 725 dol.),
- LM – średnio niski (876-3465 dol.),
- L – niski poziom (875 dol. i mniej).

Do jednej z tych czterech klas należy każda obserwacja w zbiorze uczącym.

Z kolei baza danych AMECO zawiera roczne dane makroekonomiczne dla 25 krajów Unii Europejskiej, krajów należących do strefy euro, krajów kandydujących oraz pozostałych krajów OECD (USA, Japonia, Kanada, Szwajcaria, Norwegia, Islandia, Meksyk, Korea, Australia i Nowa Zelandia).

Wspomniane kraje są charakteryzowane przez ponad 700 zmiennych, z których wybrano do analizy jedynie 6: wydatki budżetowe, wydatki na konsumpcję, wydatki rządowe, wielkość eksportu i importu oraz stopę bezrobocia.

W sumie w zbiorze EUROPA znajduje się 750 obserwacji:

- 15 krajów „starej" Unii obserwowanych w latach 1970-2005,
- 14 krajów „nowej" Unii (10 nowych oraz 4 kraje kandydujące: Bułgaria, Rumunia, Turcja, Chorwacja) obserwowanych w latach 1991-2005.

5. Rezultaty eksperymentów obliczeniowych

W zależności od tego, czy postać funkcji agregacji Ψ jest przyjęta z góry przez badacza i niezależna od danych, czy też podlega modyfikacjom w procesie ustalania parametrów modeli, można wszystkie omówione wcześniej metody podzielić na dwie grupy:

- niewymagające uczenia, które bezpośrednio łączą wyniki predykcji modeli bazowych, np. głosowanie (2) czy uśrednianie (19),
- wymagające uczenia po to, by określić niektóre parametry funkcji agregacji Ψ, np. w przypadku naiwnej metody Bayesa (7) czy metody szablonów decyzyjnych (22).

Pierwszą grupę metod stosujemy wtedy, gdy modele bazowe $D_1, ..., D_M$ mają zbliżony poziom błędu predykcji, a wyniki predykcji nie są skorelowane lub są skorelowane ujemnie. Modele łączone za pomocą tego rodzaju funkcji Ψ dają poprawę dokładności predykcji wtedy, gdy powstały w różnych podprzestrzeniach pierwotnej przestrzeni zmiennych. Kolejnym warunkiem jest to, by zbiór uczący był odpowiednio obszerny, a postać funkcji agregacji dobrze wybrana.

[2] www.worldbank.org/Home/Data/CountryClassification.

Metody z drugiej grupy pozwalają dopasować model zagregowany do danych z wykorzystaniem dodatkowego zbioru uczącego. Oznacza to tak naprawdę określenie liczebności w tablicy BKS (tab. 1), a także postać profili decyzyjnych (15) czy też tablic zgodności klasyfikacji $\mathbf{Z}^m$ w metodzie Bayesa (7).

Opierając się na zbiorze uczącym EUROPA, zbudowano 100 modeli zagregowanych, z których każdy składał się z 10 modeli indywidualnych ($M = 10$). Modele bazowe łączono za pomocą wybranych 10 funkcji agregacji opisanych powyżej (pominięto metody dla wyników predykcji w postaci uporządkowanej listy klas).

Modele bazowe w tym eksperymencie miały postać drzew klasyfikacyjnych i zostały zbudowane za pomocą procedury `rpart` [Therneau, Atkinson 1997] na podstawie zbioru uczącego (750 obserwacji). Modele $D_1, ..., D_M$ połączono za pomocą wybranych 10 funkcji agregacji i dla tak przygotowanego modelu D^* obliczono błąd klasyfikacji, stosując trzy metody[3]:

- cały zbiór uczący jako testowy (tzw. błąd zastąpienia),
- zbiór testowy wybrany losowo i zawierający 200 obserwacji,
- 10-częściowe sprawdzanie krzyżowe (cross-validation).

Wielkości błędu klasyfikacji znajdują się w tab. 2.

Tabela 2. Błąd klasyfikacji dla wybranych metod agregacji modeli bazowych

Błąd	mean	med	min	max	prod	vote	bayes	BKS	Wer	DT
Zastąpienia	0,149	0,163	0,165	0,171	**0,143**	0,145	0,193	0,203	0,214	0,209
Testowy	0,208	0,219	0,221	0,223	**0,195**	0,198	0,224	0,235	0,267	0,248
10-CV	0,201	0,215	0,218	0,220	0,196	**0,195**	0,219	0,232	0,259	0,241

Źródło: opracowanie własne za pomocą programu **R**.

W tab. 2 pokazano wyraźny podział metod łączenia modeli na dwie grupy. Pierwsze 6 metod ma wyraźnie mniejszy błąd klasyfikacji. Są to metody niewymagające uczenia, z których najbardziej dokładna jest metoda iloczynu (`prod`). Drugą grupę tworzą metody wymagające uczenia (`bayes`, `BKS`, `DT` i `Wer`), dla których ten błąd jest już nieco wyższy.

6. Wnioski

W prowadzonych eksperymentach najmniejszy błąd klasyfikacji uzyskano, gdy modele bazowe były łączone za pomocą metody iloczynu (`prod`). Dzieje się tak zwłaszcza wtedy, gdy wartości prawdopodobieństw *a posteriori* dla klas są dobrze oszacowane, a wyniki predykcji dla tych modeli nie są skorelowane.

[3] Szerzej na temat sposobów oceny błędu klasyfikacji pisze Gatnar [2001, s. 96-100].

Spośród metod niewymagających uczenia metoda maksymalnej wartości prawdopodobieństw *a posteriori* (max) daje największą wartość błędu klasyfikacji, ponieważ jest nadmiernie dopasowana do zbioru uczącego.

Z kolei metoda głosowania (vote) daje mały błąd klasyfikacji, gdy wynik predykcji modeli bazowych ma postać klasy. Poza tym metoda ta jest prostsza i szybsza pod względem obliczeń niż na przykład metoda szablonów decyzyjnych (DT).

Literatura

Domingos P., Pazzani M. (1997), *On the Optimality of the Simple Bayesian Classifier under Zero-One Loss*, „Machine Learning" 29, s. 103-130.

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

Grabisch M. (1995), *On Equivalence Classes of Fuzzy Connectives – the Case of Fuzzy Integrals*, „IEEE Transactions on Fuzzy Systems" 3(1), s. 96-109.

Huang Y.S., Suen C.Y. (1995), *A Method of Combining Multiple Experts for the Recognition of Unconstrained Handwritten Numerals*, „IEEE Transactions on Pattern Analysis and Machine Intelligence" 17, s. 90-93.

Kuncheva L., Bezdek J.C., Duin R. (2001), *Decision Templates for Multiple Classifier Fusion: An Experimental Comparison*, „Pattern Recognition" 34, s. 299-314.

Therneau T.M., Atkinson E.J. (1997), *An Introduction to Recursive Partitioning using the RPART Routines*, Mayo Foundation, Rochester.

Wernecke K.-D. (1992), *A Coupling Procedure for Discrimination of Mixed Data*, „Biometrics" 48, s. 497-506.

THE IMPACT OF COMBINATION FUNCTION ON THE CLASSIFICATION ERROR IN CLASSIFIER FUSION

Summary

Classifier combining is an important method in a nonparametric discriminant analysis. In this approach M base (local) models are combined into the aggregated (global) model with the use of the combination function Ψ :

$$\hat{D}^*(\mathbf{x}_i) = \Psi\left(\hat{D}_1(\mathbf{x}_i), \ldots, \hat{D}_M(\mathbf{x}_i)\right),$$

where $\hat{D}_m(\mathbf{x})$ is the prediction of the m-th base model.

Several different types of the function Ψ can be found in the literature. The most commonly used are: majority vote, average, maximum and product. Moreover, some more sophisticated combination functions have been proposed, such as fuzzy integrals or decision templates.

In this paper ten combination functions Ψ have been compared and their impact on the classification error of the model D^* has been investigated.