PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1126 — 2006
TAKSONOMIA 13
Klasyfikacja i analiza danych – teoria i zastosowania

Eugeniusz Gatnar
Akademia Ekonomiczna w Katowicach

WYKORZYSTANIE MIARY HAMANNA DO OCENY PODOBIEŃSTWA MODELI W PODEJŚCIU WIELOMODELOWYM

1. Wstęp

W nieparametrycznej analizie dyskryminacyjnej [Gatnar 2001] coraz większe znaczenie ma podejście wielomodelowe (agregacja modeli), polegające na łączeniu K modeli składowych (lokalnych) $C_1(\mathbf{x}),\ldots,C_K(\mathbf{x})$ w jeden model globalny $C^*(\mathbf{x})$, np. na zasadzie majoryzacji (głosowania większościowego):

$$C^*(\mathbf{x}) = \arg\max_y \left\{ \sum_{k=1}^{K} I(C_k(\mathbf{x}) = y) \right\}. \tag{1}$$

Tumer i Ghosh [1996] udowodnili, że błąd klasyfikacji dla modelu zagregowanego $C^*(\mathbf{x})$ zależy od stopnia podobieństwa („korelacji") modeli składowych. Inaczej mówiąc, im bardziej modele składowe są do siebie niepodobne, tj. zupełnie inaczej klasyfikują te same obiekty, tym bardziej dokładny jest model $C^*(\mathbf{x})$.

W literaturze zaproponowano kilka miar pozwalających ocenić podobieństwo (zgodność, korelację) modeli składowych w podejściu wielomodelowym. Przedstawiają je np. Kuncheva i Whitaker [2003].

W niniejszym artykule zostaną omówione podstawowe własności znanych do tej pory miar, zostanie także zaproponowana nowa miara podobieństwa modeli dyskryminacyjnych, oparta na współczynniku Hamanna [1961], dająca lepszą ocenę podobieństwa modeli $C_1(\mathbf{x}),\ldots,C_K(\mathbf{x})$.

2. Metody łączenia modeli

Podejście wielomodelowe polega na tym, że mając zbiór uczący $T = \{(\mathbf{x}_1, y_1),\ldots,(\mathbf{x}_N, y_N)\}$, zawierający obiekty o znanej przynależności do klas,

możemy na jego podstawie utworzyć zbiór prób uczących $T_1,\ldots,T_K$. Próby te służą do budowy modeli składowych $C_1(\mathbf{x}),\ldots,C_K(\mathbf{x})$, które są następnie łączone i dają w rezultacie model zagregowany $C^*(\mathbf{x})$.

W ostatnim czasie w literaturze zaproponowano kilka metod łączenia modeli składowych. Różnią się one albo sposobem, w jaki tworzone są modele składowe, tj. sposobem konstrukcji zbiorów uczących, albo sposobem łączenia wyników klasyfikacji dla obiektów ze zbioru testowego.

Ogólnie rzecz biorąc, wszystkie metody budowy modeli składowych możemy podzielić na trzy grupy:

- manipulujące obiektami ze zbioru uczącego, np. *bagging* [Breiman 1996], *boosting* [Freund, Shapire 1997] oraz *arcing* [Breiman 1998],
- manipulujące zmiennymi charakteryzującymi obiekty w zbiorze uczącym, np. *random subspaces* [Ho 1998] oraz *random forests* [Breiman 2001],
- manipulujące wartościami zmiennej objaśnianej (y, np. *error-correcting output coding* [Dietterich, Bakiri 1995]).

Z kolei sposoby łączenia (agregacji) wyników klasyfikacji przez modele składowe można zaliczyć do jednej z trzech grup:

- metod uśredniających (*average methods*), np. średniej arytmetycznej lub średniej ważonej,
- metod nieliniowych, np. głosowania większościowego (*majority vote*), wyboru wartości największej, metody zliczania Bordy,
- uogólniania za pomocą stosu (*stacked generalisation*), zaproponowanego przez Wolperta [1992].

3. Zróżnicowanie modeli

W wielu praktycznych zastosowaniach obserwowano, że poprawa dokładności klasyfikacji zależy od korelacji między modelami składowymi. Formalny dowód tej zależności został przedstawiony przez Hansena i Salamona [1990]. Zaproponowali oni formułę określającą wielkość błędu modelu zagregowanego $C^*(\mathbf{x})$, który powstaje poprzez połączenie K modeli składowych na zasadzie majoryzacji:

$$e(C^*) = 1 - \sum_{i=0}^{\lfloor K/2 \rfloor} \binom{K}{i} p^{K-i}(1-p)^i , \tag{2}$$

gdzie $p = 1 - e(C_i)$ oznacza dokładność i-tego modelu składowego, przy czym dla uproszczenia przyjmuje się, że $e(C_1) = e(C_2) = \ldots = e(C_K)$. Przykładowe wartości błędu (2) pokazano w tab. 1.

Tabela 1. Błędy dla modelu zagregowanego wyznaczone za pomocą formuły (2)

	K=3	K=5	K=7	K=9
p=0,6	0,352	0,3174	0,2898	0,2666
p=0,7	0,216	0,1631	0,126	0,0988
p=0,8	0,104	0,0579	0,0333	0,0196
p=0,9	0,028	0,0086	0,0027	0,0009

Tumer i Ghosh [1996] dokonali dekompozycji błędu predykcji modelu zagregowanego na dwa składniki (przy założeniu agregacji przez uśrednianie):

$$e(C^*) = e^B(C^*) + \frac{1 + r(K-1)}{K} e(C_i), \tag{3}$$

gdzie: $e^B(C^*)$ – błąd modelu optymalnego w sensie bayesowskim,

r – współczynnik korelacji między wynikami klasyfikacji modeli składowych,

$e(C_i)$ – wielkość błędu pojedynczego modelu.

Z analizy formuły (3) widać, że błąd modelu zagregowanego spada wraz ze spadkiem wielkości zależności r między modelami składowymi modelu zagregowanego. Oznacza to, że należy łączyć takie modele, które są jak najbardziej różne (*diverse*).

Jeśli chodzi o zróżnicowanie modeli składowych, to Sharkey i Sharkey [1997] rozróżniają jego cztery poziomy:

1) każdy obiekt jest błędnie klasyfikowany nie więcej niż przez jeden model,
2) każdy obiekt jest błędnie klasyfikowany nie więcej niż przez połowę modeli (majoryzacja daje zawsze model dokładny),
3) każdy obiekt jest poprawnie klasyfikowany przez przynajmniej jeden z modeli,
4) niektóre obiekty nie są poprawnie klasyfikowane przez żaden z modeli.

Poziom zróżnicowania modeli składowych wpływa na metodę ich agregacji. Na przykład zasada majoryzacyjna (1) jest odpowiednia dla modeli, które wykazują poziom zróżnicowania 1 lub 2. W przeciwnym wypadku należy wykorzystać inne, bardziej wyszukane metody agregacji, np. uogólnianie za pomocą stosu (*stacked generalization*).

4. Miary zróżnicowania modeli

Najprostszy sposób pomiaru zgodności klasyfikacji dwóch modeli $C_i(\mathbf{x})$ i $C_j(\mathbf{x})$ to wykorzystanie współczynnika:

$$Z(i,j) = \frac{1}{N} \sum_{n=1}^{N} I(C_i(\mathbf{x}_n) = C_j(\mathbf{x}_n)). \tag{4}$$

Formuła (4) nie bierze jednak pod uwagę tego, czy oba modele poprawnie rozpoznały klasę dla obiektu $\mathbf{x}_n$.

Lepsze rozwiązanie polega zatem na zastosowaniu wartości zero-jedynkowych (*oracle labels*):

$$O_k(\mathbf{x}_i) = \begin{cases} 1 & C_k(\mathbf{x}_i) = y_i \\ 0 & C_k(\mathbf{x}_i) \neq y_i \end{cases}. \tag{5}$$

Wartość O_i=1 oznacza, że model $C_i(\mathbf{x})$ poprawnie rozpoznał klasę y_i obiektu $\mathbf{x}_i$, a O_i=0 oznacza, że się pomylił. Wtedy zgodność klasyfikacji pary modeli $C_i(\mathbf{x})$ i $C_j(\mathbf{x})$ można analizować za pomocą tablicy kontyngencji o wymiarach 2×2 (tab. 2).

Tabela 2. Tablica kontyngencji dla wyników klasyfikacji

	O_j=1	O_j=0
O_i=1	a	b
O_i=0	c	d

Najbardziej znaną miarą zgodności jest zero-jedynkowa wersja współczynnika korelacji Pearsona:

$$r(i,j) = \frac{ad - bc}{(a+b)(c+d)(a+c)(b+d)}. \tag{6}$$

Giacinto i Roli [2001] zaproponowali miarę opartą na błędzie i nazwali ją łącznym zróżnicowaniem (*compound diversity*):

$$CD(i,j) = \frac{d}{a+b+c+d}. \tag{7}$$

Ta miara jest także nazywana miarą podwójnego błędu (*double-fault measure*) ponieważ jest frakcją obiektów błędnie sklasyfikowanych przez oba modele.

Partridge i Yates [1996] oraz Margineantu i Dietrich [1997] zaproponowali miarę nazwaną zróżnicowaniem wewnątrzgrupowym (*within-set generalization diversity*). Ta miara jest po prostu statystyką kappa zaproponowaną przez Fleissa [1981] do pomiaru rzetelności (*measure of interrater reliability*). Mierzy ona poziom skorygowany zgodności dwóch modeli klasyfikacyjnych:

$$K(i,j) = \frac{2(ac - bd)}{(a+b)(c+d) + (a+c)(b+d)}. \tag{8}$$

Skalak [1996] wykorzystał miarę niezgodności (*disagreement measure*), aby ocenić zróżnicowanie dwóch modeli:

$$DM(i,j) = \frac{b+c}{a+b+c+d}. \tag{9}$$

Jest to frakcja liczby obiektów, które zostały błędnie sklasyfikowane przez przynajmniej jeden z modeli.

W literaturze znane są także miary zróżnicowania dla większej liczby modeli, tzw. miary globalne (*non-pairwise*).

Na przykład Hansen i Salamon [1990] zaproponowali tzw. miarę trudności (*measure of difficulty*). Jest to wariancja zmiennej reprezentującej frakcję modeli, które poprawnie klasyfikują wybrany losowo obiekt $\mathbf{x}_n$.

Partridge i Krzanowski [1997] przygotowali uogólnioną miarę zróżnicowania (*generalized diversity measure*), zaś Cunnigham i Carney [2000] wykorzystali do pomiaru różnic między modelami funkcję entropii.

5. Miara Hammana

Kuncheva i inni [2000] zaproponowali statystykę Q Yule'a do oceny stopnia zróżnicowania modeli klasyfikacyjnych:

$$Q(i,j) = \frac{ad - bc}{ad + bc}. \tag{10}$$

Jest to miara symetryczna i przyjmuje wartości pomiędzy –1 i 1, wartość 0 zaś oznacza statystyczną niezależność dwóch modeli klasyfikacyjnych.

Statystyka Yule'a jest dobrą miarą oceny podobieństwa wyników klasyfikacji modeli $C_i(\mathbf{x})$ i $C_j(\mathbf{x})$, lecz w niektórych przypadkach jej wartość może być niezdefiniowana, np. gdy $a = 0$ i $b = 0$. Co więcej, statystyka Q nie pozwala odróżnić dwóch skrajnie różnych rozkładów wyników klasyfikacji, pokazanych w tabelach 3 i 4.

Tabela 3. Rozkład wyników klasyfikacji dla K=100

	O_j=1	O_j=0
O_i=1	49	1
O_i=0	50	0

Tabela 4. Rozkład wyników klasyfikacji dla K=100

	O_j=1	O_j=0
O_i=1	1	49
O_i=0	50	0

W przypadku tab. 3 i 4 wartość statystyki Q jest ta sama i wynosi –1, wskazując najsilniejszą ujemną korelację pomiędzy modelami. Jednak w tab. 3 znajdują się wyniki klasyfikacji dla dwóch modeli, które w ogóle nie są związane ze sobą.

Aby przezwyciężyć wspomniane ograniczenia statystyki Q Yule'a, proponujemy wykorzystanie do oceny zróżnicowania modeli składowych współczynnika Hamanna.

Hamann [1961] zaproponował binarny współczynnik podobieństwa, który jest różnicą między klasyfikacjami zgodnymi i niezgodnymi w stosunku do liczby wszystkich obserwacji:

$$H(i,j)=\frac{(a+d)-(b+c)}{a+b+c+d}. \tag{11}$$

Jej wartość znajduje się w przedziale [–1,1], przy czym wartość 0 wskazuje na równą liczbę klasyfikacji zgodnych i niezgodnych, –1 oznacza idealną niezgodność, a 1 – idealną zgodność.

Miara Hamanna pozwala odróżnić dwa rozkłady w tab. 3 i 4; dla tab. 4 przyjmuje wartość $H(i, j) = -0{,}98$, co wskazuje na silną zależność ujemną, a dla tab. 3 przyjmuje wartość $H(i, j) = -0{,}02$, oznaczającą brak zależności.

Dla modelu zagregowanego współczynnik Hamanna jest liczony jako średnia dla wszystkich par modeli składowych:

$$H(C^*)=\frac{2}{K(K-1)}\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}H(i,j). \tag{12}$$

6. Wyniki eksperymentów

Aby porównać dokładność identyfikacji zróżnicowania modeli składowych przez obie miary, tj. Yule'a i Hamanna, przygotowano eksperyment podobny do tego, który opisali Kuncheva i inni [2000]. Wykorzystano zbiór testowy zawierający 10 obserwacji (N=10) oraz trzy modele składowe: $C_1(\mathbf{x}), C_2(\mathbf{x}), C_3(\mathbf{x})$, z których każdy ma taki sam błąd klasyfikacji $e(C_i) = 0{,}4$ (co oznacza, że 6 z 10 obserwacji jest klasyfikowanych poprawnie). W sumie daje to 28 różnych modeli zagregowanych.

Do oceny zróżnicowania modeli składowych wykorzystano miarę Hamanna (tab. 5) oraz zasadę majoryzacji.

Tabela 5. Wartości miar H i Q oraz błędu klasyfikacji dla 28 modeli

Model	H_{12}	H_{13}	H_{23}	H	Q	$e(D^*)$
1	2	3	4	5	6	7
1	0,2	–0,6	–0,6	–0,33	–0,56	0,2
2	–0,2	–0,6	–0,2	–0,33	–0,67	0,2
3	–0,2	–0,2	–0,2	–0,20	–0,50	0,1
4	0,6	–0,6	–0,6	–0,20	–0,37	0,3
5	0,2	–0,6	–0,2	–0,20	–0,39	0,3
6	–0,2	–0,2	–0,2	–0,20	–0,50	0,3
7	0,2	–0,2	–0,2	–0,07	–0,22	0,2
8	1,0	–0,6	–0,6	–0,07	–0,33	0,4
9	0,6	–0,6	–0,2	–0,07	–0,21	0,4
10	0,2	–0,6	0,2	–0,07	–0,11	0,4
11	0,2	–0,2	–0,2	–0,07	–0,22	0,4
12	0,6	–0,2	–0,2	0,07	–0,04	0,3
13	0,2	–0,2	0,2	0,07	0,05	0,3
14	0,2	0,2	0,2	0,20	0,33	0,2
1	2	3	4	5	6	7

cd. tab. 5

15	0,6	–0,2	–0,2	0,07	–0,04	0,5
16	1,0	–0,2	–0,2	0,20	0,00	0,4
17	0,2	–0,2	0,2	0,07	0,05	0,5
18	0,6	–0,2	0,2	0,20	0,24	0,4
19	0,2	0,2	0,2	0,20	0,33	0,4
20	0,6	0,2	0,2	0,33	0,51	0,3
21	0,2	0,2	0,2	0,20	0,33	0,6
22	0,6	0,2	0,2	0,33	0,51	0,5
23	1,0	0,2	0,2	0,47	0,55	0,4
24	0,6	0,2	0,6	0,47	0,70	0,4
25	0,6	0,6	0,6	0,60	0,88	0,3
26	0,6	0,6	0,6	0,60	0,88	0,5
27	1,0	0,6	0,6	0,73	0,92	0,4
28	1,0	1,0	1,0	1,00	1,00	0,4

Współczynnik korelacji Pearsona między miarami Hamanna i Yule'a wynosi 0,972, co oznacza silną zależność. Jeśli zaś chodzi o korelację między obiema miarami a błędem klasyfikacji, to wyniki są podobne, chociaż trochę lepszy był współczynnik Hamanna: $H = 0{,}462$, a $Q = 0{,}431$.

Podobne wyniki osiągnięto dla zbioru testowego zawierającego 10 obserwacji ($N = 10$). Wtedy otrzymano 36 151 różnych modeli zagregowanych. Oceny zróżnicowania wyniosły odpowiednio: $H = 0{,}662$, zaś $Q = 0{,}531$.

7. Podsumowanie

Statystyka Hamanna jest miarą bardziej dokładną niż współczynnik Yule'a, pozwalającą zidentyfikować różnice między modelami składowymi w podejściu wielomodelowym. Możliwość prawidłowej oceny stopnia zróżnicowania modeli składowych jest ważna, ponieważ łączone powinny być jedynie te, które są jak najbardziej do siebie niepodobne, tj. zupełnie inaczej klasyfikują te same obserwacje. Zgodnie z twierdzeniem Tumera i Ghosha [1996], taki model zagregowany daje najbardziej dokładne wyniki predykcji.

Literatura

Breiman L. (1996), *Bagging Predictors*, „Machine Learning" nr 24, s. 123-140.

Breiman L. (1998), *Arcing Classifiers*, „Annals of Statistics" nr 26, s. 801-849.

Breiman L. (1999), *Using Adaptive Bagging to Debias Regressions*, Technical Report 547, Department of Statistics, University of California, Berkeley.

Breiman L. (2001), *Random Forests*, „Machine Learning" nr 45, s. 5-32.

Cunnigham P., Carney J. (2000), *Diversity versus Quality in Classification Ensembles Based on Feature Selection*, Proceedings of European Conference on Machine Learning, LNCS, vol. 1810, Springer, Berlin, s. 109-116.

Dietterich T., Bakiri G. (1995), *Solving Multiclass Learning Problem via Error-Correcting Output Codes*, „Journal of Artificial Intelligence Research" nr 2, s. 263-286.

Fleiss J.L. (1981), *Statistical Methods for Rates and Proportions*, John Wiley and Sons, New York.

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" nr 55, s. 119-139.

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

Giacinto G., Roli F. (2001), *Design of Effective Neural Network Ensembles for Image Classification Processes*, „Image Vision and Computing Journal" nr 19, s. 699-707.

Hamann U. (1961), *Merkmalsbestand Und Verwandtschafsbeziehungen Der Farinosae. Ein Beitragzum System Der Monokotyledonen*, „Willdenowia" nr 2, s. 639-768.

Hansen L.K., Salamon P. (1990), *Neural Network Ensembles*, „IEEE Transactions on Pattern Analysis and Machine Intelligence" nr 12, s. 993-1001.

Ho T.K. (1998), *The Random Subspace Method for Constructing Decision Forests*, „IEEE Transactions on Pattern Analysis and Machine Intelligence" nr 20, s. 832-844.

Kuncheva L., Whitaker C., Shipp D., Duin R. (2000), *Is Independence Good for Combining Classifiers*, Proceedings of the 15th International Conference on Pattern Recognition, Barcelona, Spain, s. 168-171.

Kuncheva L., Whitaker C. (2003), *Measures of Diversity in Classifier Ensembles and their Relationship With the Ensemble Accuracy*, „Machine Learning" nr 51, s. 181-207.

Margineantu M.M., Dietterich T.G. (1997), *Pruning Adaptive Boosting*, Proceedings of the 14th International Conference on Machine Learning, Morgan Kaufmann, San Mateo, s. 211-218.

Partridge D., Krzanowski W.J. (1997), *Software Diversity: Practical Statistics for Its Measurement and Exploitation*, „Information and Software Technology" nr 39, s. 707-717.

Partridge D., Yates W.B. (1996), *Engineering Multiversion Neural-Net Systems*, „Neural Computation" nr 8, s. 869-893.

Sharkey A., Sharkey N. (1997), *Diversity, Selection, and Ensembles of Artificial Neural Nets, Neural Networks and Their Applications*, NEURAP-97, s. 205--212.

Skalak D.B. (1996), *The Sources of Increased Accuracy for Two Proposed Boosting Algorithms*, Proceedings of the American Association for Artificial Intelligence AAAI-96, Morgan Kaufmann, San Mateo.
Therneau T.M., Atkinson E.J. (1997), *An Introduction to Recursive Partitioning Using the RPART Routines*, Mayo Foundation, Rochester.
Tumer K., Ghosh J. (1996), *Analysis of Decision Boundaries in Linearly Combined Neural Classifiers*, „Pattern Recognition" nr 29, s. 341-348.
Wolpert D. (1992), *Stacked Generalization*, „Neural Networks" nr 5, s. 241-259.

THE USE OF HAMANN'S COEFFICIENT AS DIVERSITY MEASURE IN MULTIPLE-MODEL APPROACH

Summary

Combining multiple classifiers into an ensemble has proved to be very successful in the past decade. The key issue is the diversity of the component classifiers, because unrelated members form an accurate ensemble.

In this paper we propose a new pairwise measure of diversity for classifier ensembles based on Hamann's similarity coefficient.