PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1076 2005
TAKSONOMIA 12
Klasyfikacja i analiza danych – teoria i zastosowania

Eugeniusz Gatnar
Akademia Ekonomiczna w Katowicach

DOBÓR ZMIENNYCH DO ZAGREGOWANYCH MODELI DYSKRYMINACYJNYCH

1. Wstęp

W pracach Gatnara [2002; 2003] rozważano modele dyskryminacyjne $D(\mathbf{x})$ mające postać drzew klasyfikacyjnych:

$$y = D(\mathbf{x}) = \sum_{k=1}^{K} a_k I(\mathbf{x}_i \in R_k), \tag{1}$$

gdzie R_k (dla $k = 1, ..., K$) to rozłączne obszary (segmenty) w wielowymiarowej przestrzeni cech, zawierające zbiory obiektów S_k, a_k – parametry modelu, I jest zaś funkcją wskaźnikową.

Modele tego typu mają wiele zalet omawianych szeroko w literaturze (por. [Gatnar 2001]), jednak ich wadą jest brak stabilności (tj. mała zmiana wartości jednej ze zmiennych $x_1,\ldots,x_n$ może spowodować powstanie zupełnie innego modelu). W rezultacie czasami uzyskuje się bardzo dokładne wyniki, a czasami błąd klasyfikacji jest zbyt duży.

Znaczną poprawę dokładności klasyfikacji można uzyskać, stosując podejście wielomodelowe, polegające na łączeniu pojedynczych modeli w jeden model zagregowany.

Breiman [1996] udowodnił, że wyraźna redukcja wielkości błędu klasyfikacji dotycząca modelu zagregowanego jest możliwa wtedy, gdy łączone ze sobą modele składowe różnią się od siebie (są niezależne). Tę ich niezależność uzyskuje się, budując je na podstawie zbiorą uczących $U_1,\ldots,U_M$, zawierających losowo wybrane obiekty lub losowo wybrane zmienne, lub przez stosowanie losowania z wagami dla obiektów.

Łatwo można dowieść, że błąd klasyfikacji modelu zagregowanego jest zawsze mniejszy niż błąd każdego z modeli składowych, przy czym modele te nie muszą być ani dokładne, ani stabilne. Najczęściej ma to miejsce w modelach o prostej postaci. Przykładem takich modeli są właśnie drzewa klasyfikacyjne (1).

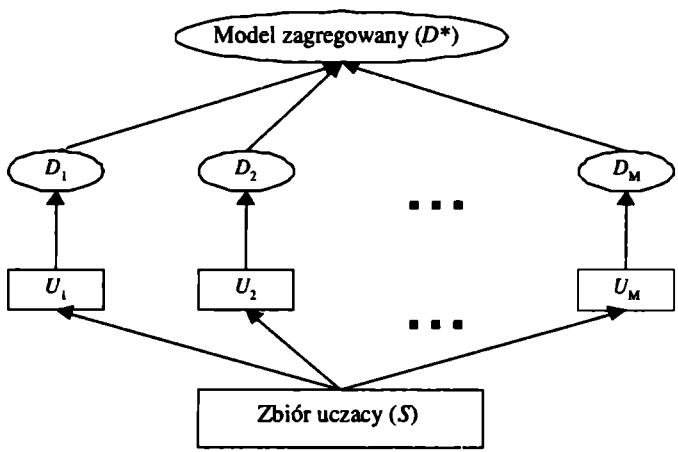
W pracy będzie przedstawiona nowa metoda doboru zmiennych do modeli składowych (w postaci drzew klasyfikacyjnych), wykorzystująca podejście zaproponowane przez Hellwiga [1969].

2. Podejście wielomodelowe

Agregacja modeli polega na wyodrębnieniu M prób uczących $U_1,\ldots,U_M$, na podstawie których są budowane modele w postaci drzew $D_1(\mathbf{x}),\ldots,D_M(\mathbf{x})$. Następnie są one łączone w jeden model $D^*(\mathbf{x})$ (por. [Gatnar 2002, s. 219]). Klasa obiektu jest ustalana w wyniku „głosowania":

$$D^*(\mathbf{x}) = \arg\max_l \left\{ \sum_{m=1}^{M} I(\hat{D}_m(\mathbf{x}) = l) \right\} . \qquad (2)$$

W literaturze mówi się o kilku metodach agregacji modeli. Podejście stochastyczne opiera się na losowaniu ze zwracaniem obiektów ze zbioru uczącego $S = \{(\mathbf{x}_1, y_1),\ldots,(\mathbf{x}_n, y_n)\}$ do kolejnych prób uczących, z których każda liczy także po n obiektów (rys. 1). W podejściu deterministycznym zaś jest wykorzystywany system wag przydzielanych obserwacjom dobieranym do tych prób.



Rys. 1. Podejście wielomodelowe
Źródło: opracowanie własne.

Każdy z modeli składowych modelu (2) można zbudować, opierając się na próbie uczącej, w której znajdują się wszystkie obserwacje, lecz zmienne są wybrane losowo.

3. Losowy dobór zmiennych do modelu

W pracy Gatnara [2003] omówiono metodę stochastyczną, polegającą na losowym doborze zmiennych $x_1,\ldots,x_n$ do prób uczących $U_1,\ldots,U_M$. Metoda ta jest prostsza od wymienionych, że pozwala tworzyć modele składowe „równolegle" i nie wymaga ona specjalnego programu komputerowego, a jedynie ogólnodostępne pakiety do tworzenia drzew klasyfikacyjnych, np. STATISTICA, AnswerTree, CART itd.

Metoda ta wydaje się najbardziej odpowiednia wtedy, gdy:

- wymiar przestrzeni zmiennych jest duży,
- występuje redundacja informacji,
- liczba obserwacji w zbiorze uczącym jest mała w porównaniu z wymiarem przestrzeni,
- modele składowe są wrażliwe na „przekleństwo wielowymiarowości" (*curse of dimensionality*).

Gatnar [2004] pokazał, że liczba zmiennych dobieranych do modeli składowych (wymiar przestrzeni cech) wpływa na wielkość błędu klasyfikacji modelu $D^*(x)$. Dlatego liczbę tę należy ograniczać, losując np. połowę z nich, tj. $L/2$. Dodatkową redukcję wymiaru przestrzeni klasyfikacji można uzyskać przez zastosowanie formalnego kryterium doboru zmiennych do modeli składowych (prób uczących).

4. Metody doboru zmiennych do modeli zagregowanych

Metody doboru zmiennych x_1, ..., x_n do modelu dyskryminacyjnego można podzielić na trzy rodzaje:

1. Metody wykorzystujące sam algorytm klasyfikacyjny do oceny przydatności każdego z wybranych podzbiorów zmiennych (*wrapper methods*).
2. Metody usuwające zmienne niepożądane przed klasyfikacją lub dobierające „najlepsze" zmienne do modelu (*filter methods*).
3. Metody oceniające pojedyncze zmienne i porównujące poziom oceny z wybranym arbitralnie progiem (*ranking methods*).

Metody należące do pierwszej grupy są bardzo wygodne, lecz wolne. Z kolei te z grupy drugiej są znacznie szybsze, pod warunkiem że rozwiązano problem przeszukiwania dużej przestrzeni rozwiązań, np. stosując strategię wspinaczki (*hill climbing)* lub strategię *best-first*. Trzecia grupa to metody bardzo subiektywne i, w związku z tym trudne do wykorzystywania w analizach porównawczych.

Wśród metod korelacyjnych znane są rozwiązania, które można zaliczyć do jednej z trzech grup:

- proste – wykorzystują wyłącznie korelacje zmiennych x_1, ..., x_n ze zmienną y,
- zaawansowane (kontekstowe) – wykorzystują także korelacje między zmiennymi x_1, ..., x_n (np. metoda CFS zaproponowana w pracy Halla [2000]),

- dodawania wartości (*merit-based*) – np. metoda zaproponowana przez Honga [1997] przydziela zmiennej x_i wartość, którą jest stopień podobieństwa, w jakim inne zmienne potrafią sklasyfikować te same obserwacje co zmienna x_i.

5. Proponowana metoda (CFSH)

W celu doboru zmiennych do prób uczących $U_1,\ldots,U_M$ można wykorzystać metodę korelacyjną, zaproponowaną na gruncie ekonometrii przez prof. Zdzisława Hellwiga [1969]. Metoda ta należy do grupy metod kontekstowych, ponieważ bierze pod uwagę związki pomiędzy zmiennymi objaśniającymi (predyktorami).

Metoda CFSH (*correlation-based feature selection based on Hellwig heuristic*) składa się z dwóch etapów.

1. Powtarzaj od $m = 1$ do M:

a) wybierz losowo $L/2$ zmiennych ze zbioru uczącego do próby U_m,

b) wybierz najlepszy podzbiór zmiennych (F_m) ze zbioru U_m, zgodnie z metodą Hellwiga,

c) utwórz model D_m w postaci drzewa klasyfikacyjnego na podstawie zbioru zmiennych F_m.

2. Połącz modele składowe w model zagregowany $D^*(x)$, stosując zasadę majoryzacji (2).

Metoda Hellwiga polega na wyborze ze zbioru wszystkich możliwych kombinacji zmiennych ($F_1, F_2, \ldots, F_K$) tej, która maksymalizuje tzw. integralną pojemność informacji:

$$H^* = \max_l \{H_l\}, \tag{3}$$

gdzie:

$$H_l = \sum_{j=1}^{m_l} h_{lj}. \tag{4}$$

We wzorze (4) h_{lj} oznacza indywidualną pojemność informacji, tj. ilość informacji, jaką niesie zmienna x_j w kombinacji F_l:

$$h_{lj} = \frac{r_j^2}{1 + \sum_{\substack{i=1 \\ i \neq j}}^{m_l} r_{ij}}. \tag{5}$$

a $l = 1,\ldots,m_l$.

Zmienne mierzone na skalach mocnych (metryczne) poddano dyskretyzacji metodą zaproponowaną w pracy [Fayyad, Irani 1993]. Ze względu na to, że zmienna y reprezentuje klasę (zmienna nominalna), zamiast współczynnika korelacji Pearsona

w formule (5) do pomiaru związków między zmiennymi wykorzystano współczynnik niepewności [Press i in. 1989]:

$$r_{ij} = \frac{2[E(x_i) + E(x_j) - E(x_i, x_j)]}{E(x_i) + E(x_j)}, \tag{6}$$

gdzie $E(x)$ to funkcja entropii.

Podobne rozwiązanie zaproponował Hall [2000] w swoim algorytmie CFS, który do każdej z prób uczących $U_1,\ldots,U_M$ dobierał zbiór zmiennych F_l, maksymalizujący wartość wyrażenia:

$$CFS(F_l) = \frac{m_l \cdot \overline{r}_j}{\sqrt{m_l + m_l(m_l - 1) \cdot \overline{r}_{ij}}}, \tag{7}$$

gdzie: $\overline{r}_j$ to przeciętna wartość współczynnika (6) pomiędzy zmienną y a pozostałymi zmiennymi objaśniającymi w kombinacji F_l, a $\overline{r}_{ij}$ to przeciętna wartość współczynnika (6) pomiędzy parami zmiennych objaśniających w tej kombinacji.

6. Przykład

Dysponując zbiorami danych udostępnionymi przez Uniwersytet Kalifornijski w Irvine [Blake, Keogh, Merz 1998], wykonano obliczenia za pomocą programu *Rpart* [Therneau, Atkinson 1997], który pracuje w środowisku **R** i służy do budowy drzew klasyfikacyjnych.

W trakcie eksperymentu zbudowano modele zagregowane, liczące po 100 drzew klasyfikacyjnych pełnej wielkości, tj. bez stosowania procedury przycinania krawędzi (*pruning*).

Tabela 1. Zbiory danych wykorzystane w eksperymentach

Zbiór danych	Liczba obiektów w zbiorze uczącym	Liczba obiektów w zbiorze testowym	Liczba zmiennych	Liczba klas
DNA	2000	1186	180	3
Letter	15000	5000	16	26
Satellite	4435	2000	36	6
Soybean	600	83	35	19
German credit	900	100	24	2
Segmentation	2000	310	19	8
Sick	3400	372	29	2
Anneal	800	98	38	5
Australian credit	600	90	15	2

Źródło: opracowanie własne.

Wyniki klasyfikacji dla zbiorów testowych zaprezentowanych w tab. 1 zamieszczono w tab. 2. Porównania dotyczą jedynie pojedynczego modelu w postaci

drzewa klasyfikacyjnego oraz metody CFS, ponieważ w pracy Halla [2000] wykazano, że metoda ta daje dokładniejsze modele niż użycie innych metod doboru zmiennych.

Tabela 2. Błędy klasyfikacji dla wybranych zbiorów danych (w %)

Zbiór danych	Pojedyncze drzewo	CFS	CFSH
DNA	6,40	5,20	4,51
Letter	14,00	10,83	5,84
Satellite	13,80	14,87	10,32
Soybean	8,00	9,34	6,98
German credit	29,60	27,33	26,92
Segmentation	3,70	3,37	2,27
Sick	1,30	2,51	2,14
Anneal	1,40	1,22	1,20
Australian credit	14,90	14,53	14,10

Źródło: opracowanie własne.

Na podstawie wyników przedstawionych w tab. 2 można wyciągnąć wniosek, że metoda CFSH daje dokładniejsze modele niż metoda CFS.

Pewnym problemem jest jednak czas obliczeń, związany z potrzebą przeszukiwania w metodzie CFSH dosyć dużej przestrzeni rozwiązań. Należy rozważać bowiem wszystkie możliwe kombinacje $L/2$ zmiennych.

7. Podsumowanie

Rozważania nad losowym doborem zmiennych do modeli dyskryminacyjnych ukazują wpływ liczby tych zmiennych (wymiaru przestrzeni zmiennych) na dokładność klasyfikacji modelu zagregowanego. Wprowadzenie do modelu większej liczby zmiennych powoduje zwiększenie błędu klasyfikacji.

Aby zmniejszyć tę liczbę, zastosowano korelacyjną metodę doboru zmiennych do modelu. Zaproponowana metoda CFSH generuje modele dyskryminacyjne dokładniejsze od tych, które uzyskał Hall [2000]. Ponieważ wykazał on, że metoda CFS jest najlepsza (w sensie minimalizacji błędu klasyfikacji) metodę CFSH porównano jedynie z metodą CFS.

Literatura

Blake C., Keogh E., Merz C.J., *UCI Repository of Machine Learning Databases*, Department of Information and Computer Science, University of California, Irvine, CA 1998.

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

Breiman L., *Bagging Predictors*, „Machine Learning" 1996, 24, s. 123-140.
Breiman L., *Random Forests*, „Machine Learning" 2001, 45, s. 5-32.
Dietterich T., Bakiri G., *Solving Multiclass Learning Problem via Error-Correcting Output Codes*, „Journal of Artificial Intelligence Research" 1995, 2, s. 263-286.
Fayyad U.M., Irani K.B., *Multi-Interval Discretisation Of Continuous-Valued Attributes*, [w:] *Proceedings of the XIII International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, San Francisco 1993, s. 1022-1027.
Freund Y., Schapire R.E., *A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting*, „Journal of Computer and System Sciences" 1997, 55, s. 119-139.
Gatnar E., *Nieparametryczna metoda dyskryminacji i regresji*, PWN, Warszawa 2001.
Gatnar E., *Agregacja modeli dyskryminacyjnych*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 942, AE, Wrocław 2002, s. 217-226.
Gatnar E., *O pewnej metodzie redukcji błędu klasyfikacji*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 988, AE, Wrocław 2003, s. 245-253.
Gatnar E., *Problem wymiaru przestrzeni cech w klasyfikacji*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1022, AE, Wrocław 2004, s. 477-484.
Hall M., *Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning*, [w:] *Proceedings of the 17th International Conference on Machine Learning*, Morgan Kaufmann, San Francisco 2000.
Hellwig Z., *O problemie optymalnego wyboru predykant*, „Przegląd Statystyczny" 1969, 3-4.
Ho T.K., *The Random Subspace Method for Constructing Decision Forests*, IEEE Trans. on Pattern Analysis and Machine Learning, 1998, 20, s. 832-844.
Hong S.J., *Use of Contextual Information for Feature Ranking and Discretization*, *IEEE Transactions on Knowledge and Data Engineering*, 1997, 9, s. 718-730.
Kira A., Rendell L., *A Practical Approach to Feature Selection*, [w:] *Proceedings of the 9th International Conference on Machine Learning*, red. D. Sleeman, P. Edwards, Morgan Kaufmann, San Francisco 1992, s. 249-256.
Kohavi R., John G.H., *Wrappers for Feature Subset Selection*, „Artificial Intelligence" 1997, 97, s. 273-324.
Press W.H., Flannery B.P., Teukolsky S.A., Vetterling W.T., *Numerical Recipes in Pascal*, Cambridge University Press, Cambridge 1989.
Therneau T.M., Atkinson E.J., *An Introduction to Recursive Partitioning Using the RPART Routines*, Mayo Foundation, Rochester 1997.

FEATURE SELECTION TO AGGREGATED CLASSIFICATION MODELS

Summary

Significant improvement of classification accuracy can be obtained by aggregation of multiple models. Known methods in this field are mostly based on sampling cases from the training set, or changing weights for cases.

Further reduction of classification error can be achieved by random selection of variables to the training subsamples or directly to the model. In this paper we propose a new correlation-based feature selection method for classifier ensembles (CFSH) that is contextual (uses feature intercorrelations) and based on the Hellwig heuristic. It gives more accurate aggregated models than those built with other correlation-based feature selection methods.