PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1189 — 2007
Zastosowania metod ilościowych

Andrzej Dudek, Adam Kurzydłowski
Akademia Ekonomiczna we Wrocławiu

DOBÓR ZMIENNYCH DO ZAGREGOWANYCH MODELI DYSKRYMINACYJNYCH Z WYKORZYSTANIEM ALGORYTMÓW GENETYCZNYCH

1. Wstęp

Drzewa klasyfikacyjne należą do grupy metod eksploracyjnych, które wykorzystuje się w poszukiwaniu i wykrywaniu relacji występujących w określonym zbiorze danych oraz w ustaleniu stopnia dopasowania modelu do posiadanych danych (zob. [Hastie, Tibshirani, Friedman 2001, s. 269]):

$$Y = \sum_{k=1}^{K} a_k I(\mathbf{X} \in S_k), \tag{1}$$

gdzie: Y – zmienna objaśniana,
$\mathbf{x} = [X_1, ..., X_z]$ – wektor zmiennych objaśniających,
a_k – parametry modelu,
I – funkcja wskaźnikowa,
$k = 1, \ldots K$ – numer wyodrębnianej klasy,
S_k – rozłączne obszary (klasy) wielowymiarowej przestrzeni zmiennych objaśniających.

Drzewa klasyfikacyjne o charakterze dyskryminacyjnym są konstruowane wtedy, gdy zmienna objaśniana Y mierzona jest na skali niemetrycznej, tj. nominalnej lub porządkowej (zob. [Weiss, Indurkhya 1995, s. 383]). Celem konstrukcji takiego drzewa klasyfikacyjnego jest znalezienie modelu przypisującego obiekty do znanych kategorii zmiennej objaśnianej.

Każdorazowy podział zbioru następuje na podstawie wartości zmiennej opisującej obiekty. Wybór zmiennej będącej podstawą podziału dokonywany jest na podstawie określonej miary statystycznej będącej jednocześnie miarą jakości podziału. Ważne jest założenie o istnieniu związku między wartościami zmiennych a ich przynależnością do określonej klasy [Gatnar 1998, s. 170]).

Ustalenie i zidentyfikowanie wszystkich możliwych podziałów wartości każdej zmiennej objaśniającej uwzględnionej w badaniu jest punktem wyjścia dla przewidywania przynależności obiektów do klas wyznaczonych przez zmienną objaśnianą. Precyzję, z którą obiekt przydzielany jest do klasy zmiennej objaśnianej, reprezentuje błąd klasyfikacji.

Jedną z metod oceny błędu klasyfikacji jest podział posiadanego zbioru danych na dwie rozłączne części, uczącą (na podstawie której budowany jest model) oraz testową (wykorzystywaną do zbadania jakości modelu).

Wpływ na otrzymywany błąd klasyfikacji ma oczywiście sposób konstrukcji drzewa klasyfikacyjnego. Najodpowiedniejszym sposobem byłoby doprowadzenie do wykonania wszystkich możliwych podziałów zbioru S i jego podzbiorów, a następnie wybranie z nich optymalnych na podstawie ustalonego kryterium. W literaturze można znaleźć kilka praktycznych rozwiązań dających minimalizację błędu klasyfikacji, m.in.: poprzez przycinanie krawędzi (*pruning*), skracanie krawędzi (*shrinking*) czy też łączenie drzew klasyfikacyjnych.

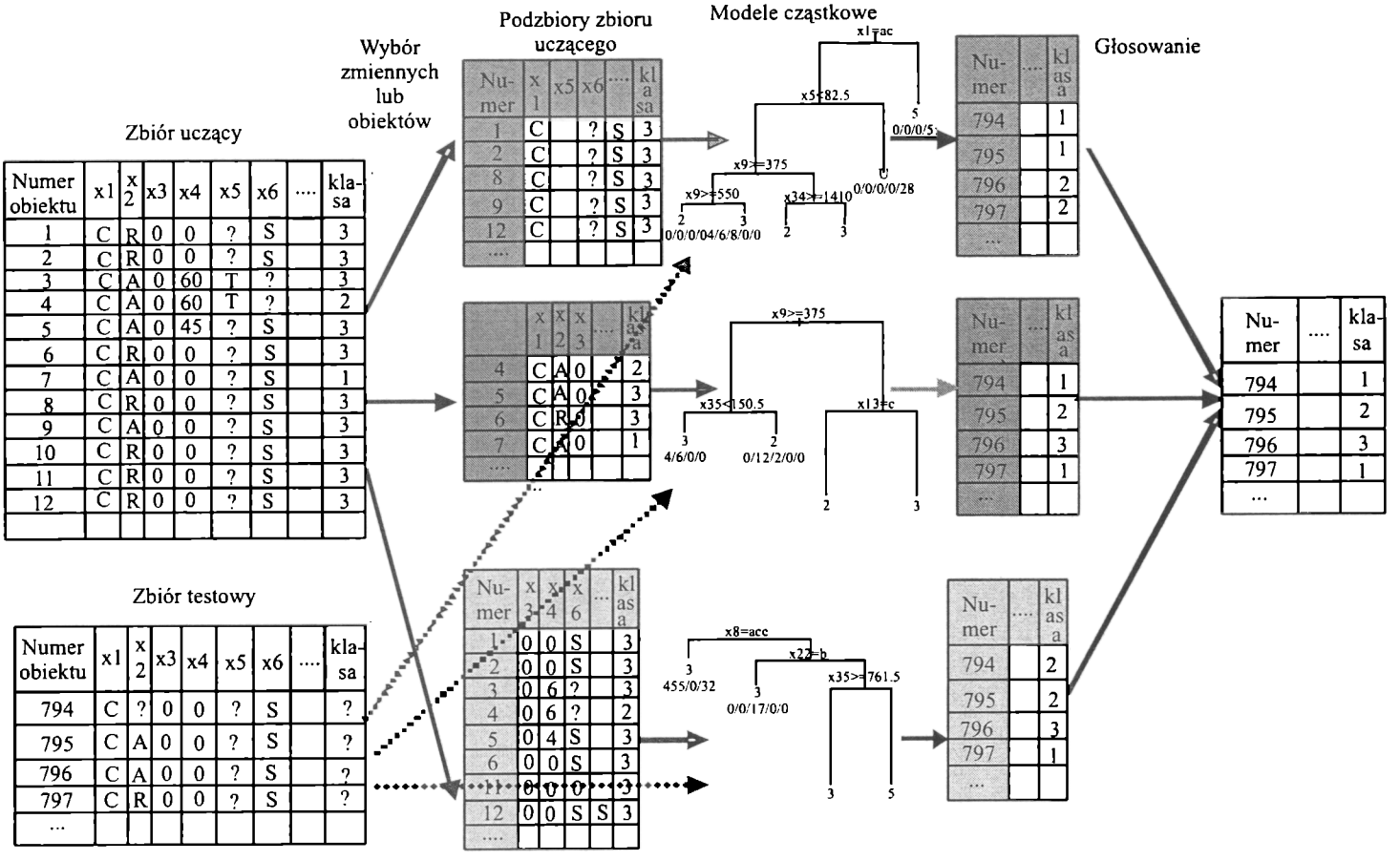
W badaniach z wykorzystaniem drzew klasyfikacyjnych coraz większą rolę odgrywają modele odchodzące od analizy za pomocą pojedynczego drzewa klasyfikacyjnego na rzecz podejścia wielomodelowego, łączącego modele składowe w model zagregowany. Przyczyną tego stanu rzeczy jest to, że modele te dają wyraźnie mniejszy błąd klasyfikacji, o ile zachowana jest niezależność modeli składowych [Breiman 1996].

W artykule zaproponowana zostanie odmiana metody CFSH (*Correlation-based Feature Selection based on Hellwig Heuristic*) wykorzystująca do znalezienia najlepszego podzbioru zmiennych w modelu składowych optymalizację z wykorzystaniem algorytmów genetycznych. Ponadto porównane zostaną błędy klasyfikacji metod CFS, CFSH i nowej metody na zbiorach danych pochodzących z repozytorium Uniwersytetu Kalifornijskiego w Irvine.

2. Agregacja drzew klasyfikacyjnych

Idea agregacji drzew klasyfikacyjnych polega na utworzeniu kilkunastu lub kilkudziesięciu odrębnych modeli drzew klasyfikacyjnych na podstawie posiadanego zbioru danych, a następnie złożeniu ich w jeden model dyskryminacyjny. W przypadku rozpatrywania modeli zagregowanych kluczowym zagadnieniem jest odpowiedni dobór podzbiorów zbioru uczącego, przy czym dobór ten dotyczyć może zarówno obiektów, jak i zmiennych zbioru uczącego.

Osobnym problemem w analizie dyskryminacyjnej z wykorzystaniem drzew klasyfikacyjnych jest wybór miary heterogeniczności, jednak jak wskazują przeprowadzone eksperymenty, nie ma ona takiego wpływu na wielkość błędu klasyfikacji jak wybór właściwej wielkości drzewa. Stąd też w artykule zagadnienie to nie będzie szerzej omawiane.



Rys. 1. Procedura podejścia wielomodelowego

Źródło: opracowanie własne.

Rysunek 1 prezentuje ogólny schemat postępowania w przypadku podejścia wielomodelowego. Ze zbioru uczącego tworzone są podzbiory, zawierające wybrane obiekty lub wybrane zmienne, na podstawie których tworzone są pojedyncze drzewa klasyfikacyjne. Dla każdego z tych drzew dokonywana jest predykcja obiektów ze zbioru testowego. Z powstałych w ten sposób modeli cząstkowych metodą głosowania (z odmianami) tworzony jest ostateczny model zagregowany określający ostateczną przynależność obiektów ze zbioru testowego do klas. Ten model nie ma już reprezentacji w postaci pojedynczego drzewa.

Wśród metod doboru obiektów do modeli zagregowanych najczęściej wykorzystywane są m.in. *bagging*, *boosting*, *adaptive bagging czy arcing*, które zostały przedstawione m.in. w pracach: Breimana [1996; 1998] oraz Freunda i Shapire'a [1996]. W artykule dokładniej omówione zostaną metody doboru zmiennych do modeli zagregowanych. Wśród metod stosowanych w tym celu można wyróżnić:

- *wrapper methods* – metody wykorzystujące sam algorytm klasyfikacyjny drzewa do wyboru zmiennych tworzących model cząstkowy;
- *filter methods* – metody wybierające do modelu te kombinacje zmiennych, dla których wartość funkcji oceniającej jest najwyższa;
- *ranking methods* – metody wybierające do modelu te zmienne, dla których wartość funkcji oceniającej (obliczanej niezależnie dla każdej zmiennej) jest wyższa niż wartość progowa.

Każda z zaprezentowanych grup metod ma swoje wady i zalety. Metody z pierwszej grupy określane są jako nieskomplikowane w użyciu, ale relatywnie wolne. Druga grupa to metody realizujące o wiele szybciej proces doboru zmiennych (np. wskutek zastosowania strategii wspinaczki – *hill climbing*). Trzecią grupę stanowią metody rzadko wykorzystywane w analizach porównawczych wskutek swojej dużej subiektywności.

3. Metody doboru zmiennych do modeli cząstkowych

Pewnym rozwiązaniem dającym dobre rezultaty, jak pokazali w swojej pracy Hall i Smith [2000], okazała się metoda *Correlation-based Feature Selection* (CFS) zaproponowana przez Halla [2000]. Z kolei w pracy Gatnara [2003] przedstawiono metodę *Correlation-based Feature Selection based on Hellwig Heuristic* (CFSH) i empirycznie wykazano, że daje ona mniejsze błędy klasyfikacji niż CFS.

Podobnie jak w oryginalnej metodzie Hellwiga [1969] dotyczącej modeli ekonometrycznych w CFSH spośród wszystkich kombinacji zmiennych do modelu wybierane są te, dla których wartość integralnego nośnika pojemności informacji H_l jest maksymalna.

$$H_l = \sum_{j=1}^{m_l} h_{lj}, \tag{2}$$

$$h_{lj} = \frac{r_j^2}{1 + \sum_{\substack{i=1 \\ i \neq j}}^{m_l} |r_{ij}|}, \tag{3}$$

gdzie: m_l – liczba zmiennych w l-tej kombinacji,

r_j – symetryczny współczynnik niepewności [Theil 1972] liczony dla i-tej zmiennej i dla zmiennej oznaczającej przynależność do klasy (w oryginalnej metodzie Hellwiga współczynnik korelacji między zmienną objaśnianą a i-tą zmienną kandydatką),

r_{ij} – symetryczny współczynnik niepewności liczony dla i-tej i j-tej zmiennej (w oryginalnej metodzie Hellwiga współczynnik korelacji między i-tą a j-tą zmienną kandydatką).

Metoda CFSH daje dla danych empirycznych z UCI *Machnie Learning Repository* najmniejsze błędy klasyfikacji (zob. [Gatnar 2005, s. 83]). Jednak jej praktyczne zastosowanie jest ograniczone tym, iż jest to metoda przeszukująca całą przestrzeń kombinacji zmiennych, więc czas jej wykonania rośnie wykładniczo w stosunku do liczby zmiennych w modelu, co zauważa sam autor metody [Gatnar 2005].

Aby uniknąć tego problemu, zamiast przeszukiwać wszystkie możliwe kombinacje zmiennych, można szukać maksymalnej wartości integralnego nośnika pojemności informacji za pomocą odpowiedniej strategii szukania optimum, np. za pomocą strategii wspinaczkowej (*climbing hill*).

4. Algorytmy genetyczne jako narzędzie optymalizacji

Pewną propozycją autorów jest wykorzystanie algorytmów genetycznych w procesie poszukiwania maksymalnej wartości integralnego nośnika pojemności informacji. Istota algorytmów genetycznych w odróżnieniu od innych algorytmów optymalizacyjnych wyraża się w tym, że [Goldberg 2003, s. 23]):

- nie przetwarzają bezpośrednio parametrów zadania, lecz ich zakodowaną postać;
- prowadzą poszukiwania, wychodząc nie z pojedynczego punktu, lecz z pewnej ich populacji;
- korzystają tylko z funkcji celu, a nie z jej pochodnych lub innych pomocniczych informacji;
- stosują probabilistyczne, a nie deterministyczne reguły wyboru.

Ogólną postać algorytmów genetycznych można przedstawić w sześciu etapach (zob. [Goldberg 2003; Arabas 2002]), a mianowicie:

1. Określenie populacji początkowej.
2. Ocena osobników populacji.
3. Reprodukcja i sukcesja:

a) obliczenie całkowitej funkcji przystosowania,
b) obliczenie cząstkowych funkcji przystosowania,
c) losowanie chromosomów.
4. Operacje genetyczne:
a) krzyżowanie,
b) mutacja,
c) inwersja.
5. Przepisanie osobników z populacji tymczasowej do bazowej.
6. Sprawdzenie kryterium konwergencji i ewentualne powtórzenie etapów 3-5.

Algorytmy genetyczne znajdują wiele zastosowań w rozwiązywaniu problemów optymalizacyjnych. Wśród dziedzin, w których są najczęściej używane, wymienić należy [Goldberg 2003, s. 141]:

- biologię,
- informatykę,
- technikę i badania operacyjne,
- przetwarzanie obrazów i rozpoznawanie wzorców,
- nauki fizyczne,
- nauki społeczne.

5. Algorytmy genetyczne w metodzie CFSH

W zaproponowanej modyfikacji metody CFSH, zamiast obliczać wartości integralnego nośnika pojemności informacji dla każdej kombinacji, znajduje się kombinację maksymalizującą $H = \sum_{j=1}^{m} h_{lj}$ przy użyciu algorytmu GAFIT:

$$h_{lj} = \frac{\omega_j r_j^2}{1 + \sum_{\substack{i=1 \\ i \neq j}}^{m_l} \omega_i \left| r_{ij} \right|}, \tag{4}$$

gdzie: ω_i – waga i-tej zmiennej (wektor ω jest optymalizowany),
m – liczba zmiennych,
r_j – symetryczny współczynnik niepewności liczony dla i-tej zmiennej i dla zmiennej oznaczającej przynależność do klasy,
r_{ij} – symetryczny współczynnik niepewności liczony dla i-tej i j-tej zmiennej.

Zaproponowane rozwiązanie wykorzystujące algorytmy genetyczne sprawdzono dla zbiorów danych udostępnionych przez Uniwersytet Kalifornijski w Irvine. Wyniki klasyfikacji dla zbiorów testowych zaprezentowano w tab. 1. Zmierzono również czas uzyskania rozwiązania, czyli znalezienia modelu zagregowanego. Do obliczeń wykorzystano biblioteki *RPART* i *GAFIT* działające w środowisku ***R***

W tabeli 2 przedstawiono czas realizacji eksperymentu dla poszczególnych zbiorów danych.

Tabela 1. Błędy klasyfikacji dla wybranych zbiorów danych (w %)

Zbiór danych	Pojedyncze drzewo	CFS	CFSH	Nowa metoda
DNA	6,40	5,20	4,51	4,63
Letter	14,00	10,83	5,84	5,91
Satellite	13,80	14,87	10,32	9,98
Soybean	8,00	9,34	6,98	7,02
German credit	29,60	27,33	26,92	21,27
Segmentation	3,70	3,37	2,27	2,75
Sick	1,30	2,51	2,14	2,26
Anneal	1,40	1,22	1,20	1,20
Australian credit	14,90	14,53	14,10	14,23

Źródło: opracowanie własne na podstawie pracy [Gatnar 2005, s. 84].

Tabela 2. Błędy klasyfikacji dla wybranych zbiorów danych

Zbiór danych	Liczba obiektów w zbiorze uczącym	Liczba obiektów w zbiorze testowym	Liczba zmiennych	Czas wykonania (hh:mm:ss)
DNA	2000	1186	180	03:12:45
Letter	15000	5000	16	00:18:27
Satellite	4435	2000	36	00:20:41
Soybean	600	83	35	00:12:43
German credit	900	100	24	00:06:43
Segmentation	2000	310	19	00:07:21
Sick	3400	372	29	00:10:51
Anneal	700	98	38	00:13:53
Australian credit	600	90	15	00:02:20

Źródło: opracowanie własne na podstawie pracy [Gatnar 2005, s. 83].

6. Podsumowanie

Ustalenie zagregowanego drzewa klasyfikacyjnego dającego minimalizację błędu klasyfikacji jest niezwykle trudnym zadaniem dla badacza. Wiadomo bowiem, że na dokładność klasyfikacji w takim przypadku wpływ ma liczba zmiennych uwzględnionych w badaniu. W literaturze znajduje się wiele praktycznych wskazówek oraz konkretnych metod wspomagających badacza w podejmowaniu tego typu decyzji. W praktyce jednak nie istnieje jeden konkretny algorytm, który mógłby zoptymalizować poszukiwanie drzewa odpowiedniej wielkości.

Zaproponowana metoda daje porównywalne błędy klasyfikacji co metoda CFSH. W artykule nie porównano czasów działania metod CFS CFSH i nowej metody. Wynika to z braku gotowego oprogramowania. Symulacja z uwzględnieniem czasu zostanie wykonana po napisaniu kodu odpowiednich funkcji dla tych metod. Można ją będzie rozszerzyć na inne propozycje miary współzależności zmiennych. Ponownie wymagać to będzie zmian w oprogramowaniu.

Literatura

Arabas J., *Wykłady z algorytmów ewolucyjnych*, WNT, Warszawa 2004.

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

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

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., *Symboliczne metody klasyfikacji danych*, PWN, Warszawa 1998.

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

Gatnar E., *O pewnej metodzie redukcji błędu klasyfikacji*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 988, AE, Wrocław 2003, s. 245-253.

Gatnar E., *Dobór zmiennych do zagregowanych modeli dyskryminacyjnych*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1076, AE, Wrocław 2005, s. 79-85.

Goldberg J., *Algorytmy genetyczne i ich zastosowania*, WNT, Warszawa 2003.

Hall M., *Correlation-based Feature Selection for Discrete and Numeric Machine Learning*, [w:] *Proceedings of the 17th International Conference on Machine Learning*, Morgan Kaufmann, San Francisco 2000.

Hall M., Smith L., *Feature Selection for Machine Learning Comparing a Correlation-based Filter Approach to the Wrapper*, http://home.eng.iastate.edu/~julied/classes/ee547/Handouts/Flairs.pdf, 2000.

Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer-Verlag, New York, Berlin, Heidelberg 2001.

Hellwig Z., *O problemie optymalnego wyboru predykant.* „Przegląd Statystyczny" 1969, 3-4.

Theil H., *Statistical Decomposition Analysis*, North-Holland Publishing, Amsterdam 1972.

Weiss S.M., Indurkhya N., *Rule-based Machine Learning Methods for Functional Prediction*, „Journal of Artificial Intelligence Research"1995, no. 3, s. 383-403.

SELECTION OF VARIABLES INTO AGGREGATED DISCRIMINANT MODELS WITH USAGE OF GENETIC ALGORITHMS

Summary

In discriminant analysis studies models using single classification trees are often replaced by models aggregating partial models into one multiple model.

Breiman [1996] showed that significant reduction of classification error can be achieved when independency of partial models is fulfilled. So key role in such approach plays appropriate selection of objects or variables of subsets of training set.

Among methods of selection of objects into aggregated models the most often used are: *boosting, bagging, adaptive bagging, arcing, windowing*. Among methods of selection of variables into aggregated models *Correlation-based Feature Selection* (CFS) developed by Hall [2000] was most effective. Gatnar [2003] proposed *Correlation-based Feature Selection based on Hellwig Heuristic* (CFSH) method and empirically showed that CFSH gave smaller classification errors that CFS.

In this paper some modification of CFSH is proposed and genetic algorithms are used to find best subsets of variables in partial models. Classification errors for CFFS, CFSH and modified method are compared on datasets from *University of California Repository of Machine Learning*.