PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1169 2007
TAKSONOMIA 14
Klasyfikacja i analiza danych – teoria i zastosowania

Andrzej Dudek, Adam Kurzydłowski
Akademia Ekonomiczna we Wrocławiu

PORÓWNANIE METOD DOBORU ZMIENNYCH DO ZAGREGOWANYCH MODELI DYSKRYMINACYJNYCH

1. Wstęp

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.

Potrzeba korzystania z tego typu rozwiązań wynika również ze słabości, jaką odznaczają się drzewa klasyfikacyjne – zarówno te o charakterze dyskryminacyjnym, jak i regresyjnym. Wadą tą jest stabilność, a dokładniej – jej brak. Własność ta przejawia się w tym, że nawet niewielka zmiana wartości zmiennych w zbiorze uczącym S może i często prowadzi do wygenerowania zupełnie innego drzewa. To powoduje z kolei, że użyteczność prognostyczna takiego modelu jest bardzo słaba. Najczęściej wygenerowane na jego podstawie reguły decyzyjne wykorzystywane są do klasyfikacji nowych obiektów, których przynależność klasowa nie jest znana.

W artykule przedstawione zostaną wyniki symulacji porównawczej metod doboru zmiennych do modeli składowych na podstawie repozytorium danych z Uniwersytetu Kalifornijskiego w Irvine, przy czym porównanie błędów klasyfikacji zostanie przeprowadzone dla różnych miar zależności zmiennych. W badaniu wykorzystano metody: *correlation-based feature selection* [Hall 2000], *correlation-based feature selection based on Hellwig heuristics* [Gatnar 2005] oraz modyfikację metody CFSH wykorzystującą do optymalizacji algorytmy genetyczne.

Istota algorytmów genetycznych w odróżnieniu od innych algorytmów optymalizacyjnych wyraża się w tym, że [Goldberg 2003, s. 23]:

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

Ogólną postać algorytmów genetycznych można przedstawić w sześciu krokach (zob. [Goldberg 2003]). Są nimi:

1. Określenie populacji początkowej.
2. Ocena osobników populacji.
3. Reprodukcja i sukcesja.
4. Operacje genetyczne: krzyżowanie, mutacja, inwersja.
5. Przepisanie osobników z populacji tymczasowej do bazowej.
6. Sprawdzenie kryterium konwergencji i ewentualne powtórzenie kroków 3-5.

Algorytmy genetyczne znajdują wiele zastosowań do rozwiązywania problemów optymalizacyjnych. Wśród dziedzin, w których są one najczęściej używane, wymienić należy [Goldberg 2003, s. 141]: biologię, informatykę, technikę i badania operacyjne, przetwarzanie obrazów i rozpoznawanie wzorców, nauki fizyczne, nauki społeczne.

2. Agregacja drzew klasyfikacyjnych

Idea agregacji drzew klasyfikacyjnych (modeli) polega na utworzeniu pewnej liczby (zazwyczaj kilkudziesięciu lub kilkuset) modeli drzew klasyfikacyjnych D_v na podstawie U_v prób uczących wyodrębnionych z posiadanego zbioru danych. Następnie przeprowadzana jest procedura łączenia wszystkich modeli w jeden model zagregowany D^*. W zależności od charakteru drzewa klasyfikacyjnego inaczej ustalane są w poszczególnych podzbiorach modelu zagregowanego wartości zmiennej objaśnianej.

W przypadku klasyfikacji kategoria zmiennej objaśnianej dla każdego obiektu ustalana jest na podstawie liczby głosów oddanych na tę kategorię przez cząstkowe modele dyskryminacyjne. Inaczej mówiąc, dany obiekt przydzielany jest do tej klasy, którą najczęściej wskazywały kolejne modele dyskryminacyjne. Z kolei w przypadku modeli o charakterze regresyjnym przeprowadzana jest operacja uśredniania wartości zmiennej objaśnianej w każdym z wyodrębnionych podzbiorów (zob. [Gatnar, Walesiak 2004, s. 124]).

Kluczowym zagadnieniem w modelach zagregowanych jest odpowiedni dobór 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 (zob. [Gatnar 2001, s. 104]). W związku z tym w artykule zagadnienie to nie będzie szerzej omawiane.

Ogólny schemat postępowania w podejściu wielomodelowym można scharakteryzować następująco. 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.

3. Metody doboru zmiennych do modeli cząstkowych

Ogół metod doboru zmiennych do modelu zagregowanego można podzielić na grupę metod o charakterze losowym i nielosowym. Metodami o charakterze losowym zajmowali się m.in.: Ho [1998]; Breiman [2001], Gatnar [2003]. W artykule poruszona zostanie problematyka metod doboru zmiennych o charakterze nielosowym. Wśród tego typu metod można wyróżnić (zob. [Gatnar 2005, s. 81]):

- *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.

W artykule Dudka i Kurzydłowskiego [2006] przedstawiono porównanie trzech metod dla zbiorów danych udostępnionych przez Uniwersytet Kalifornijski w Irvine. Metodami tymi były: zaproponowana przez Halla [2000] *correlation-based feature selection* (CFS); *correlation-based feature selection based on Hellwig heuristic* (CFSH) przedstawiona w pracy Gatnara [2003] oraz propozycja autorów odnośnie do wykorzystania algorytmów genetycznych w procesie poszukiwania maksymalnej wartości integralnego nośnika pojemności informacji.

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.

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

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

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

r_j – symetryczny współczynnik niepewności [Theil 1972] liczony dla *i*-tej zmiennej oraz 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). Współczynnik ten wyraża się wzorem:

$$r_{ij} = \frac{2\left[E(x_i) + E(x_j) - E(x_i, x_j)\right]}{E(x_i) + E(x_j)}, \tag{3}$$

gdzie $E(x)$ to funkcja entropii, przy czym współczynnik niepewności może przyjmować wartości z przedziału od 0 do 1. Wartości zbliżone do zera oznaczają brak zależności pomiędzy zmiennymi, wartość zbliżona do 1 oznacza zaś, że zmienne są silnie zależne. Miara ta ma więc zbliżone właściwości do modułu współczynnika korelacji Pearsona.

Metoda CFSH daje dla danych empirycznych z UCI *Machine Learning Repository* najmniejsze błędy klasyfikacji, jednak jej praktyczne zastosowanie jest ograniczone faktem, iż jest to metoda przeszukująca całą przestrzeń kombinacji zmiennych, wskutek czego 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*).

W zaproponowanej modyfikacji metody CFSH, zamiast obliczać wartości integralnego nośnika pojemności informacji dla każdej kombinacji, znajduje się kombinację maksymalizującą $H_l = \sum_{j=1}^{m_l} h_{lj}$ z użyciem algorytmu GAFIT[1]:

[1] Algorytm GAFIT, dostępny w pakiecie *GAFIT* środowiska statystycznego *R*, wykorzystuje algorytmy genetyczne do znalezienia odpowiednich współczynników dopasowania krzywych (*curve-fitting*) opisujących model i szukania przez nie optymalnej, z punktu widzenia funkcji celu, kombinacji zmiennych.

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

gdzie: ω_i – waga *i-tej* zmiennej (wektor ω jest optymalizowany),
m – liczba zmiennych.

4. Symulacja porównawcza

Dla zbiorów danych udostępnionych przez Uniwersytet Kalifornijski w Irvine porównano skuteczność metod *correlation-based feature selection* (CFS), *correlation-based feature selection based on Hellwig heuristic* (CFSH) i modyfikacji metody CFSH. Zaproponowana metoda dała błędy klasyfikacji porównywalne do metody CFSH (tab. 1). Tym razem postanowiono rozszerzyć symulację o inne propozycje miar współzależności zmiennych.

Tabela 1. Błędy klasyfikacji dla wybranych zbiorów danych w %

Zbiór danych	Pojedyncze drzewo	CFS	CFSH	Modyfikacja metody CFSH
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].

Zmierzono także 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.

Należy zauważyć, że dla zbioru zawierającego 180 zmiennych czas uzyskania rozwiązania, choć dość długi (ok. 3 godzin), jest jednak zdecydowanie krótszy niż w przypadku sprawdzania wszystkich kombinacji zmiennych (tego drugiego czasu nie udało się autorom zmierzyć, jednak ustalono, że znacznie przekracza on 3 godziny).

Porównano również błędy klasyfikacji, zastępując symetryczny współczynnik niepewności Theila innymi miarami badania zależności zmiennych nominalnych (Phi, V Cramera, Lambdę Goodmana – Kruskala i symetryczny współczynniki niepewności Theila) oraz współczynnikiem korelacji Spearmana i Kendalla. W tabeli 3 przedstawiono wyniki tego porównania.

Tabela 3. Błędy klasyfikacji przy zastosowaniu różnych współczynników zależności zmiennych w %

Zbiór danych	Phi	V Cramera	Lambda Goodmana-Kruskala	Współczynnik Theila	Współczynnik korelacji Spearmana	Współczynnik korelacji Kendalla
DNA	4,77	4,78	4,45	4,63	4,63	5,09
Letter	6,01	6,04	5,81	5,91	5,93	6,12
Satellite	9,70	10,06	9,81	9,98	10,16	11,08
Soybean	6,73	7,35	6,72	7,02	7,85	7,52
German credit	21,58	21,79	20,76	21,27	22,19	23,45
Segmentation	2,76	2,71	2,84	2,75	2,84	2,90
Sick	2,18	2,17	2,33	2,26	2,54	2,42
Anneal	1,23	1,23	1,23	1,20	1,28	1,31
Australian credit	14,12	13,81	13,72	14,23	14,66	15,51

Źródło: opracowanie własne.

Z przeprowadzonego badania wynika, iż wybór miary zależności zmiennych nie ma zbyt dużego wpływu na wielkość błędów klasyfikacji, o ile do procedury wprowadzony zostanie miernik przeznaczony dla danych nominalnych. W przypadku zastosowania klasycznych miar korelacji następuje istotne pogorszenie uzyskiwanych wyników klasyfikacji.

5. Podsumowanie

Ustalenie zagregowanego drzewa klasyfikacyjnego dającego minimalizację błędu klasyfikacji jest niezwykle trudnym zadaniem. Dokładność klasyfikacji w takim przypadku zależy od liczby 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.

Porównanie metod doboru zmiennych pozwala stwierdzić, iż najmniejsze błędy klasyfikacji daje rzeczywiście metoda CFSH. Różnica między metodą CFSH a jej zaproponowaną modyfikacją nie jest jednak zbyt duża, w przypadku zbiorów danych o dużej liczbie zmiennych ta druga metoda jest zaś zdecydowanie szybsza. Symulacja wykazała także, że wybór miary zależności zmiennych nie ma zbyt dużego wpływu na wielkość błędów klasyfikacji, o ile do procedury wybrany zostanie miernik przeznaczony dla danych mierzonych na skali nominalnej, zastosowanie klasycznych miar korelacji istotnie pogarsza natomiast wynik klasyfikacji.

Literatura

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

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

Dudek A., Kurzydłowski A. (2006), *Dobór zmiennych do zagregowanych modeli dyskryminacyjnych z wykorzystaniem algorytmów genetycznych*, materiały z XXXII Konferencji Ekonometryków, Statystyków i Matematyków (w recenzji).

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

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

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

Gatnar E., Walesiak M. (red.) (2004), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, AE, Wrocław.

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

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

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

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

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

THE COMPARISON SELECTION METHODS OF VARIABLES INTO AGGREGATED DISCRIMINANT MODELS

Summary

In discriminant analysis studies, models using single classification trees are often replaced by models aggregating partial models into one multiple model. Between selection methods of objects into aggregated models *boosting, bagging, adaptive bagging, arcing, windowing* are most commonly used. The most effective method of selection of variables into aggregated models is *Correlation-based Feature Selection* (CFS) developed by Hall [2000]. Gatnar [2003] proposed *Correlation-based Feature Selection based on Hellwig Heuristic* (CFSH) method and empirically showed that CFSH gives smaller classification errors that CFS.

In this paper comparison between CFS, CFSH and modification of CFSH method based on optimization through genetic algorithms is presented.