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

Dorota Rozmus
Akademia Ekonomiczna w Katowicach

BADANIE RELACJI MIĘDZY ZRÓŻNICOWANIEM POJEDYNCZYCH MODELI A BŁĘDEM MODELU ZAGREGOWANEGO

1. Wstęp

Zastosowanie modeli zagregowanych w klasyfikacji i regresji cieszy się bardzo dużym zainteresowaniem w ostatnich latach. Budowa takiego modelu polega na tym, że wartości teoretyczne wielu pojedynczych modeli łączone są ze sobą za pomocą różnych operatorów agregacji. W przypadku klasyfikacji przeważnie stosuje się głosowanie majoryzacyjne, a w regresji – uśrednianie.

Oprócz wysokiej jakości pojedynczych modeli, istnieje także drugi czynnik wpływający na jakość modelu zagregowanego. Czynnik ten określany jest jako zróżnicowanie modeli cząstkowych [Ruta, Gabrys 2001]. Jest wiele sposobów wprowadzania zróżnicowania do zbioru pojedynczych modeli, np. poprzez ich budowę na podstawie różnych podzbiorów zmiennych objaśniających czy też obserwacji wybranych z pierwotnego zbioru uczącego. Gdy wartością zmiennej objaśnianej jest klasa (dyskryminacja), pojęcie zróżnicowania modeli nie jest dokładnie zdefiniowane [Ruta, Gabrys 2002]. W regresji natomiast znaczenie zróżnicowania może zostać pokazane przez dekompozycje błędu modelu zagregowanego przedstawione przez Ueda i Nakano [1996] oraz Krogha i Vedelsby'ego [1995]

Badania empiryczne zaprezentowane w tym artykule miały na celu sprawdzenie, czy dekompozycje te pozwolą na sformułowanie relacji zachodzących między stopniem zróżnicowania pojedynczych modeli a błędem modelu zagregowanego. Ponadto celem tego opracowania było także sprawdzenie, która z tych dwóch dekompozycji jest bardziej użyteczna w badaniu wspomnianej relacji. Ujawnienie się jasnej relacji między stopniem zróżnicowania a błędem modelu zagregowanego jest ważne, bo coraz częściej pojawiają się algorytmy, które taki czynnik, jak zróż-

nicowanie modeli, w sposób jawny biorą pod uwagę w procesie konstrukcji modelu zagregowanego (np. [Opitz 1999; Carney, Cunningham 1999]).

2. Dekompozycje błędu modelu zagregowanego w regresji

Ueda i Nakano [1996] pokazali, że średni błąd kwadratowy podlega dekompozycji na trzy składowe:

$$E\{(\hat{y}_m - y_n)^2\} = \overline{\text{bias}}^2 + \frac{1}{K}\overline{\text{var}} + \left(1 - \frac{1}{K}\right)\overline{\text{cov}}, \tag{1}$$

tj. uśrednione obciążenie pojedynczych modeli:

$$\overline{\text{bias}} = \frac{1}{K}\sum_{k=1}^{K}\left(E\{\hat{y}_k\} - y_n\right), \tag{2}$$

uśrednioną wariancję:

$$\overline{\text{var}} = \frac{1}{K}\sum_{k=1}^{K} E\left\{(\hat{y}_k - E\{\hat{y}_k\})^2\right\}, \tag{3}$$

oraz uśrednioną kowariancję:

$$\overline{\text{cov}} = \frac{1}{K(K-1)}\sum_{k}\sum_{k\neq j} E\left\{(\hat{y}_k - E\{\hat{y}_k\})(\hat{y}_j - E\{\hat{y}_j\})\right\}. \tag{4}$$

Na podstawie tej dekompozycji widać, że błąd modelu zagregowanego zależy także od stopnia skorelowania między wartościami teoretycznymi pojedynczych modeli, który mierzony jest przez trzecią składową tej dekompozycji, czyli kowariancję.

Dekompozycja Ueda i Nakano pokazuje, jak mierzyć zróżnicowanie modeli, nie pokazuje natomiast, jaka jest relacja między stopniem zróżnicowania pojedynczych modeli a ich jakością. Okazuje się jednak, że te dwa czynniki są ze sobą powiązane, co pokazuje druga dekompozycja autorstwa Krogha i Vedelsby'ego [1995].

Wykazali oni, że w przypadku pojedynczej obserwacji, błąd predykcji na podstawie modelu zagregowanego jest różnicą dwóch wielkości:

$$\varepsilon_{agr} = (\hat{y}_m - y_n)^2 = \sum_{k=1}^{K} w_k(\hat{y}_k - y_n)^2 - \sum_{k=1}^{K} w_k(\hat{y}_k - \hat{y}_m)^2, \tag{5}$$

gdzie: y_n to rzeczywista wartość zmiennej objaśnianej,

$\hat{y}_k$ – wartość teoretyczna k-tego pojedynczego modelu,

$\hat{y}_m$ – wartość teoretyczna modelu zagregowanego, obliczana jako:

$$\hat{y}_m = \sum_{k=1}^{K} w_k \hat{y}_k, \tag{6}$$

oraz $\sum_{k=1}^{K} w_k = 1$.

Pierwszy element mierzy uśredniony błąd pojedynczych modeli i określa, jaka jest ich jakość. Drugi natomiast, zwany czynnikiem zróżnicowania (*ambiguity term*), określa stopień rozproszenia wartości teoretycznych poszczególnych modeli wokół wartości teoretycznej modelu zagregowanego, czyli mierzy ich zróżnicowanie. W związku z tym, że jest on zawsze dodatni, odjęcie go od pierwszego oznaczać będzie, że błąd modelu zagregowanego będzie równy uśrednionemu błędowi pojedynczych modeli lub będzie od niego mniejszy. Im większa wartość czynnika zróżnicowania, tym większa powinna być redukcja błędu przez model zagregowany. Jednakże, jeżeli zróżnicowanie pojedynczych modeli rośnie, to rośnie także wartość pierwszego elementu tej dekompozycji. Oznacza to zatem, że maksymalizacja samego tylko zróżnicowania nie wystarcza, by zapewnić wysoką jakość modelu zagregowanego. Potrzebne jest zapewnienie właściwej relacji między stopniem zróżnicowania a jakością pojedynczych modeli, by osiągnąć niski błąd modelu zagregowanego.

3. Badanie empiryczne

Celem empirycznej części artykułu jest zbadanie, czy zachodzą ścisłe relacje między stopniem zróżnicowania pojedynczych modeli, mierzonym za pomocą przedstawionych dekompozycji, a błędem modelu zagregowanego. Zastosowane metody agregacji to *bagging* [Breiman 1996], polegający na losowym doborze obserwacji do kolejnych prób uczących; metoda losowych podprzestrzeni [Ho 1998], polegająca na losowym doborze cech do kolejnych zbiorów uczących oraz metoda *random forest* [Breiman 2001], w której losowemu doborowi podlegają zarówno cechy, jak i obiekty.

W odniesieniu do dekompozycji (5) najpierw wyliczono jej elementy składowe, a potem zbadano relację między kowariancją a wartością błędu modelu zagregowanego. Dla dekompozycji Krogha i Vedelsby'ego w pierwszym kroku wyliczony zaś został udział czynnika zróżnicowania w sumie obydwu składników tej dekompozycji, a następnie zbadano jego relacje w stosunku do błędu modelu zagregowanego.

Do eksperymentów zastosowano ogólnodostępne zbiory z repozytorium baz danych Uniwersytetu Kalifornijskiego [Blake, Keogh, Merz 1988], dzieląc je za każdym razem na część uczącą (80%) i testową (20%). Ich charakterystyka znajduje się w tab. 1.

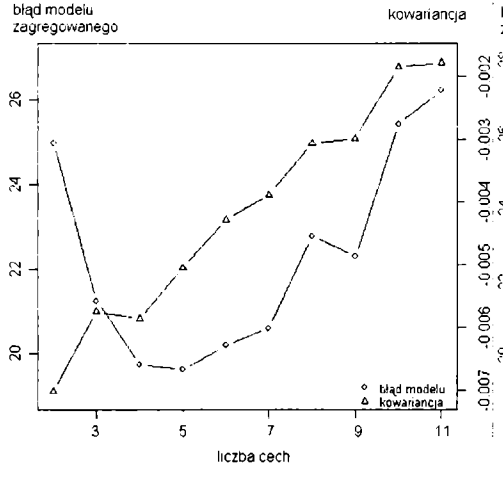
Tabela 1. Zastosowane zbiory danych

Nazwa	Liczba obserwacji	Liczba zmiennych objaśniających
Boston	506	13
Ozon	366	12
Friedman 1	500	10

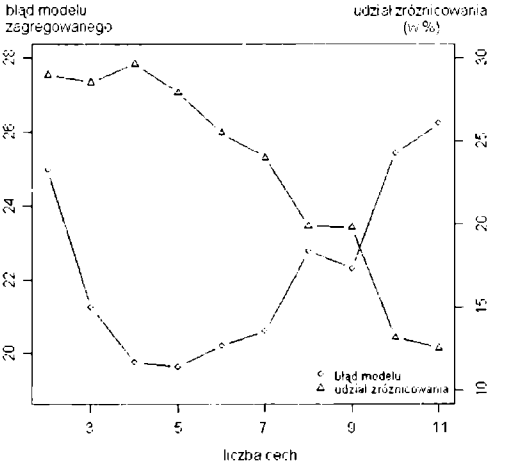
Źródło: opracowanie własne na podstawie: http://www.ics.uci.edu/~mlearn/MLSummary.html.

Pierwszy eksperyment miał na celu zastosowanie przedstawionych dekompozycji do zbadania relacji między błędem modelu zagregowanego a stopniem zróżnicowania pojedynczych modeli w metodzie losowych podprzestrzeni w zależności od liczby cech dobieranych do prób uczących. W tym celu dla zbioru *Boston* losowano od 2 do 12 cech, dla zbioru *Ozon* – od 2 do 11, a dla zbioru *Friedman 1* – od 2 do 9, dodając za każdym razem kolejną cechę. Model zagregowany obejmował 50 pojedynczych modeli. Wartości błędu modelu zagregowanego oraz składowych dekompozycji wyliczone zostały na podstawie zbiorów testowych. Wszystkie wyniki zilustrowano wykresami. Na rysunkach 1-6 na jednym wykresie przedstawiono wartości błędu modelu zagregowanego oraz kowariancji (udziału zróżnicowania). Ze względu na duże różnice w rzędzie wielkości ilustrowanych wartości lewa oś stanowi skalę dla wartości błędu modelu, a prawa – dla kowariancji (udziału zróżnicowania). Ograniczone rozmiary opracowania powodują, że pokazane będą tylko wybrane wyniki, ale przedstawione wnioski dotyczą całości badań.

Wyniki empiryczne pokazują brak wyraźnej relacji między zróżnicowaniem mierzonym za pomocą kowariancji a błędem modelu zagregowanego. Ilustracją tego faktu jest rys. 1, na którym malejąca wartość kowariancji łączy się początkowo z malejącą, a następnie rosnącą wartością błędu modelu zagregowanego. Wyniki uzyskane na podstawie dekompozycji Krogha i Vedelsby'ego zaś pokazują (rys. 2), że wyższy udział czynnika zróżnicowania w sumie składowych tej dekompozycji łączy się z niższą wartością błędu modelu zagregowanego. Wynika stąd zatem, że zapewnienie odpowiedniej relacji między stopniem zróżnicowania modeli cząstkowych a ich jakością pozwala na obniżenie błędu modelu zagregowanego.

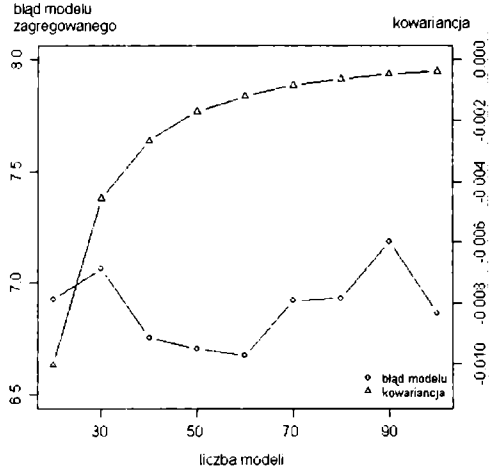


Rys. 1. Wartości kowariancji i błędu modelu zagregowanego w metodzie losowych podprzestrzeni dla zbioru *Ozon*
Źródło: opracowanie własne.

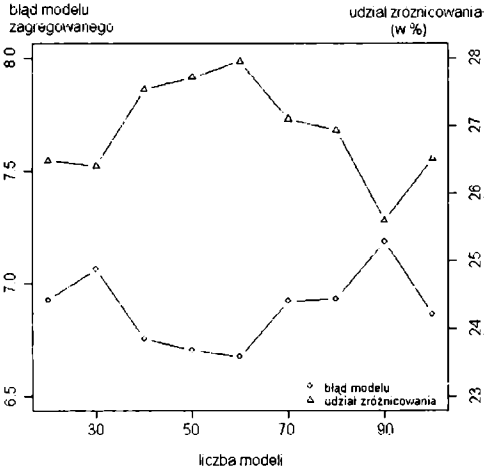


Rys. 2. Wartości udziału czynnika zróżnicowania w dekompozycji Krogha i Vedelsby'ego i błędu modelu zagregowanego w metodzie losowych podprzestrzeni dla zbioru *Ozon*
Źródło: opracowanie własne.

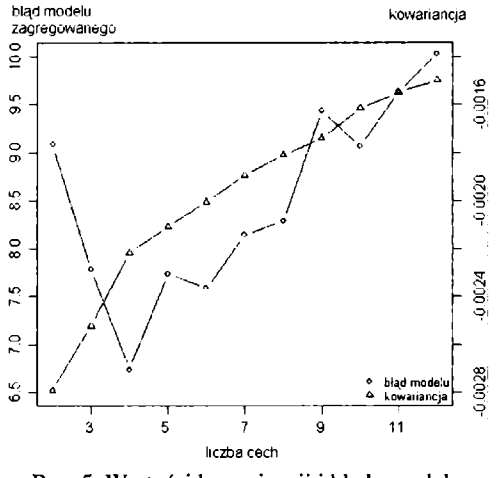
Celem drugiego eksperymentu było zastosowanie omawianych dekompozycji do zbadania, jak w metodzie *bagging* kształtuje się relacja między błędem modelu zagregowanego a stopniem zróżnicowania modeli w zależności od liczby modeli składowych. W związku z tym zbudowano model zagregowany na podstawie od 20 do 100 modeli, dodając za każdym razem kolejne 10 modeli.



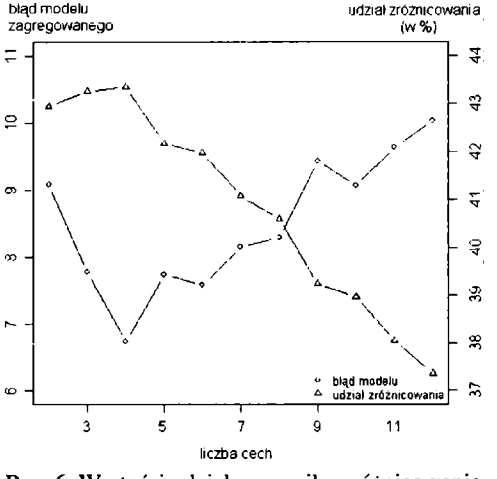
Rys. 3. Wartości kowariancji i błędu modelu agregowanego w metodzie *bagging* dla zbioru *Friedman 1*
Źródło: opracowanie własne.



Rys. 4. Wartości udziału czynnika zróżnicowania w dekompozycji Krogha i Vedelsby'ego i błędu modelu zagregowanego w metodzie *bagging* dla zbioru *Friedman1*
Źródło: opracowanie własne.



Rys. 5. Wartości kowariancji i błędu modelu zagregowanego w metodzie *random forest* dla zbioru *Boston*
Źródło: opracowanie własne.



Rys. 6. Wartości udziału czynnika zróżnicowania w dekompozycji Krogha i Vedelsby'ego i błędu modelu zagregowanego w metodzie *random forest* dla zbioru *Boston*
Źródło: opracowanie własne.

Wyniki z zastosowaniem dekompozycji Ueda i Nakano, pokazane na ryś. 3, potwierdzają brak wyraźnego związku między zróżnicowaniem mierzonym za pomocą kowariancji a błędem modelu zagregowanego. Dopiero druga z zastosowanych dekompozycji (autorstwa Krogha i Vedelsby'ego) pozwala sformułować relację między stopniem zróżnicowania modeli a błędem modelu zagregowanego: większemu udziałowi czynnika zróżnicowania w sumie składowych tej dekompozycji towarzyszy niższa wartość błędu modelu zagregowanego.

Dla metody agregacji *random forest*, przy zastosowaniu obydwu dekompozycji, zbadano relację między wartością błędu modelu zagregowanego a zróżnicowaniem poszczególnych modeli w zależności od liczby cech losowo dobieranych na każdym etapie budowy drzewa regresyjnego. Losowano analogiczną liczbę cech jak w przypadku metody losowych podprzestrzeni, a model zagregowany obejmował 100 składowych.

Wyniki przedstawione na rys. 5 potwierdzają wcześniejsze wnioski: pomiar zróżnicowania pojedynczych modeli za pomocą kowariancji nie pozwala wnioskować o wartości błędu modelu zagregowanego. Na podstawie wyników uzyskanych z zastosowaniem dekompozycji Krogha i Vedelsby'ego (rys. 6) można zaś stwierdzić, że niższej wartości błędu modelu zagregowanego towarzyszy wyższy udział czynnika zróżnicowania w sumie składowych tej dekompozycji.

4. Wnioski

Przechodząc do sformułowania wniosków ostatecznych, można stwierdzić, że błąd modelu zagregowanego zależy zasadniczo od dwóch czynników: jakości pojedynczych modeli (mierzonej przeciętnym ich błędem) i stopnia ich zróżnicowania. Wyniki empiryczne uzyskane na podstawie dekompozycji Ueda i Nakano wskazują na brak jasnej relacji między stopniem zróżnicowania modeli mierzonego za pomocą kowariancji a błędem modelu zagregowanego. Dekompozycja Krogha i Vedelsby'ego pokazuje zaś, że zwiększanie udziału czynnika zróżnicowania w sumie składników tej dekompozycji pozwala na obniżenie błędu modelu zagregowanego. Dodatkowy wniosek, jaki można wyciągnąć na podstawie przeprowadzonych badań, jest taki, że zapewnienie właściwej relacji między stopniem zróżnicowania modeli cząstkowych a ich przeciętnym błędem zapewni redukcję błędu modelu zagregowanego.

Uzyskane wyniki empiryczne pozwalają stwierdzić, że z dwóch dekompozycji przedstawionych w tym artykule dekompozycja Krogha i Vedelsby'ego może zostać uznana za bardziej użyteczną w analizie relacji między stopniem zróżnicowania modeli a błędem modelu zagregowanego. Dekompozycja ta bowiem pozwala na sformułowanie zależności zachodzącej między jakością pojedynczych modeli, stopniem ich zróżnicowania a błędem modelu zagregowanego.

Literatura

Blake C., Keogh E., Merz C.J. (1988), *UCI Repository of Machine Learning Databases*, Department of Information and Computer Science, University of California, Irvine.

Breiman L. (1996), *Bagging Predictors*, „Machine Learning" 26 (2).

Breiman L. (2001), *Random Forests*, „Machine Learning" 45.

Carney J., Cunningham P. (1999), *Tuning Diversityin Bagged Neural Network Ensembles*, "Technical Report TCD-CS-1999-44", Trinity College Dublin.

Ho T. K. (1998) *The Random Subspace Method for Constructing Decision Forests*, "IEEE Transactionson Pattern Analysis and Machine Intelligence" 20, 8.

Krogh A., Vedelsby J. (1995), *Neural Network Ensembles, Cross Validation, and Active Learning*, [w:] G. Tesauro, D. Touretzky, T. Leen (red.), *Advances in Neural Information Processing Systems 7*, Mass.: MIT Press, Cambrige.

Opitz D. (1999), *Feature Selection for Ensambles*, Proceedings of 16th National Conference on Artificial Intelligence (AAAI).

Ruta D., Gabrys B. (2001) *Analysis of the Correlation between Majority Voting Error and the Diversity Measures in Multiple Classifier Systems*, Proceedings of the SOCO/ISFI'2001 Conference, Paisley, UK.

Ruta D., Gabrys B. (2002), *Set Analysis of Coincident Errors and its Applications for Combining Classifiers*, "Pattern Recognition and String Matching, Combinatorial Optimisation" 13, Kluwer Academic Publishers.

Ueda N., Nakano R. (1996), *Generalization Error of Ensemble Estimators*, Proceedings of International Conference on Neural Networks.

ANALYSIS OF RELATIONSHIP BETWEEN SINGLE MODELS DIVERSITY END ERROR OF AGGREGATED MODEL

Summary

Ensembles become more and more popular in classification and regression problems. Very often they have better empirical and theoretical properties than single models. One factor that is needed in order to get high performance of aggregated model is recognized as models diversity. The paper discusses the problem of ensemble diversity in regression using the idea of ensemble error decomposition introduced by Ueda, Nakano (1995) and Krogh, Vedelsby (1996). We show also results of experiments concerning the relationship between ensemble error and ensemble diversity, measured by means of the two presented decompositions.