PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 176
Taksonomia 18. Klasyfikacja i analiza danych – teoria i zastosowania 2011

Aleksandra Szlachcińska
Wojewódzki Szpital Specjalistyczny im. M. Kopernika w Łodzi

Anna Witaszczyk, Małgorzata Misztal
Uniwersytet Łódzki

O ZASTOSOWANIU METODY WIĄZANIA MODELI DO POPRAWY DOKŁADNOŚCI KLASYFIKACJI PACJENTÓW Z POJEDYNCZYM CIENIEM OKRĄGŁYM PŁUCA

Streszczenie: Metoda wiązania modeli (*bundling*) została zaproponowana przez Hothorna jako modyfikacja metody *bagging*. Polega ona na wykorzystaniu dodatkowych modeli, innych klas niż drzewa klasyfikacyjne, budowanych na podstawie zbioru OOB (*out-of-bag*), zawierającego obserwacje spoza aktualnej próby bootstrapowej. Na podstawie tych modeli dokonuje się predykcji dla obserwacji w próbie bootstrapowej, a następnie wyniki predykcji traktuje się jako dodatkowe zmienne objaśniające przy budowie drzewa klasyfikacyjnego. W artykule przedstawiono wyniki wykorzystania metody wiązania modeli do poprawy dokładności klasyfikacji pacjentów z pojedynczym cieniem okrągłym płuca.

Słowa kluczowe: drzewa klasyfikacyjne, modele zagregowane, *bagging*, *bundling*, diagnostyka medyczna

1. Wstęp

W pracy Szlachcińskiej, Witaszczyk i Misztal [2009] analizowano dane dotyczące pacjentów poddanych leczeniu operacyjnemu z powodu pojedynczego cienia okrągłego płuca, uwidocznionego w badaniu tomograficznym klatki piersiowej. Celem prowadzonych wówczas badań była identyfikacja czynników (wyników badań laboratoryjnych, wielkości i umiejscowienia zmian uwidocznionych na zdjęciach tomograficznych), na podstawie których można sądzić, jeszcze przed badaniem histopatologicznym, że występująca zmiana jest nowotworem złośliwym.

Przeprowadzone analizy potwierdziły przydatność metody rekurencyjnego podziału w procesie identyfikacji rodzaju zmian u pacjentów. Uzyskane w wyniku zastosowania algorytmu CART reguły klasyfikacji pacjentów do wyróżnionych grup (nowotwór złośliwy, nowotwór łagodny) prowadziły do niższych błędów klasyfikacji niż stosowane zwykle w diagnostyce medycznej modele regresji logistycznej.

Pamiętać jednak należy o podstawowej wadzie drzew klasyfikacyjnych, jaką jest brak stabilności. Oznacza to, że mała zmiana wartości cech obiektów w zbiorze uczącym może prowadzić do powstania zupełnie innego modelu, z innym błędem klasyfikacji.

Ponieważ umiejętność rozpoznania typu nowotworu przed podjęciem leczenia operacyjnego może wpływać na sposób leczenia oraz poprawić komfort życia pacjentów, konieczne staje się znalezienie metody pozwalającej klasyfikować pacjentów do wyróżnionych grup ryzyka z możliwie najniższym błędem klasyfikacji.

Poprawę stabilności oraz dokładności predykcji można uzyskać, budując modele złożone (zagregowane lub hybrydowe).

Model zagregowany jest zbiorem pojedynczych klasyfikatorów, których odpowiedzi są zagregowane do jednej odpowiedzi całego systemu, zwykle z wykorzystaniem zasady majoryzacji. Najczęściej stosowane metody agregacji to *bagging* [Breiman 1996], *boosting* [Freund i Schapire 1997] i *Random Forests* [Breiman 2001].

Model hybrydowy natomiast integruje w fazie uczenia przynajmniej dwa różne modele (np. drzewo klasyfikacyjne i regresję logistyczną czy liniowe funkcje dyskryminacyjne). Uzasadnieniem budowy takich modeli jest twierdzenie NFL (*No Free Lunch* [Wolpert i Macready 1997]), z którego wynika, że dla każdego algorytmu uczącego istnieje pewna klasa problemów, dla których jest skuteczny. Dla pozostałych problemów algorytm może być mniej skuteczny od innych algorytmów (por. [Stefanowski 2001]).

Metodą łączącą elementy modelu zagregowanego i hybrydowego jest zaproponowana przez Hothorna [2003] metoda wiązania modeli (*bundling*).

W pracy przedstawiono wyniki zastosowania metody wiązania modeli do poprawy dokładności klasyfikacji pacjentów z pojedynczym cieniem okrągłym płuca.

2. Metoda wiązania modeli

Metoda wiązania modeli (*bundling*) została zaproponowana przez Hothorna [2003] jako rozszerzenie metody agregacji bootstrapowej (*bagging*).

Załóżmy, że rozważamy pewien zbiór obiektów U (zbiór uczący). Każdy element tego zbioru jest scharakteryzowany przez wektor $p + 1$ cech: $[\boldsymbol{x}, y]$, gdzie

$$\boldsymbol{x} = [x_1, x_2, ..., x_p]^T.$$

W metodzie *bagging* dokonuje się losowania ze zwracaniem obiektów ze zbioru uczącego do N-elementowych prób uczących $U_1, \ldots, U_v$, przy czym waga każdego obiektu jest jednakowa i równa $w_i = \frac{1}{N}$.

Prawdopodobieństwo znalezienia się obserwacji w próbie bootstrapowej wynosi: $1 - \frac{1}{e} \approx 0{,}632$. Zatem około 37% obserwacji ze zbioru uczącego U nie znajdzie się w żadnej próbie uczącej. Obserwacje te tworzą tzw. zbiór OOB (*Out–Of–Bag*), który często wykorzystuje się jako zbiór testowy.

Inne zastosowanie zbioru OOB zaproponowali w 2003 roku Hothorn i Lausen. Zbiór OOB jest wykorzystywany do budowy K dodatkowych modeli: $D_1^*,\ldots, D_K^*$, innych niż drzewa klasyfikacyjne. Na podstawie tych modeli dokonuje się predykcji dla obserwacji w próbie bootstrapowej, a wyniki predykcji (przewidywane klasy, prawdopodobieństwa warunkowe lub wartości funkcji dyskryminacyjnych) traktowane są jako dodatkowe zmienne objaśniające przy budowie drzewa klasyfikacyjnego: $X_1^*,\ldots, X_K^*$.

Zatem drzewo klasyfikacyjne budowane jest z wykorzystaniem rozszerzonej próby uczącej, w której obiekty scharakteryzowane są przez wektor $p + K + 1$ cech $[\boldsymbol{x}^*, y]$, gdzie

$$\boldsymbol{x} = [x_1, x_2, \ldots, x_p, x_1^*, \ldots, x_K^*]^T.$$

Metoda rekurencyjnego podziału wybiera najlepsze zmienne spośród zestawu zmiennych objaśniających i dodatkowych zmiennych. W tym sensie drzewo klasyfikacyjne „wiąże" dodatkowe modele klasyfikacyjne.

Procedurę powtarza się V razy, a wyniki predykcji dla obserwacji są oparte na głosowaniu większościowym.

W pierwszej wersji Hothorn i Lausen [2003] zaproponowali łączenie drzewa klasyfikacyjnego CART oraz liniowego modelu dyskryminacyjnego LDA (tzw. metoda *double – bagging*). Zbiór OOB był wykorzystywany do szacowania wartości współczynników liniowej funkcji dyskryminacyjnej. Następnie, wartości zmiennych dyskryminacyjnych dla obserwacji w próbie bootstrapowej uzupełniały zbiór zmiennych objaśniających.

Propozycja Hothorna [2003] rozszerza możliwości łączenia modeli (inne modele dodatkowe niż LDA – np. regresja logistyczna, metoda k-NN czy metoda wektorów nośnych SVM, większa liczba modeli dodatkowych itp).

Algorytm metody wiązania modeli można opisać w następujący sposób [Gatnar 2008]:

I. Określić liczbę modeli bazowych V oraz liczbę modeli dodatkowych K.

II. Dla $v = 1, 2, \ldots, V$:

1. Wylosować N-elementową próbę bootstrapową U_v ze zbioru uczącego U.

2. Zbudować dodatkowe modele $D_1^*, \ldots, D_K^*$ dla zbioru $OOB_v = U \setminus U_v$.

3. Dokonać predykcji modelu D_k^* ($k = 1, 2, \ldots, K$) dla obserwacji $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ z próby uczącej U_v, uzyskując w ten sposób realizację dodatkowej zmiennej objaśniającej X_k^*.

4. Dodać wyniki predykcji do próby uczącej, tworząc zbiór $p + K$ zmiennych objaśniających. Każdy obiekt jest teraz opisany przez wektor cech $[x_1, \ldots, x_p, x_1^*, \ldots, x_K^*]^T$.

5. Zbudować model bazowy D_v w postaci drzewa klasyfikacyjnego.

III. Dokonać predykcji modelu zagregowanego dla obserwacji $\boldsymbol{x}_i$ z wykorzystaniem modeli bazowych $D_1, \ldots, D_v$, stosując głosowanie większościowe.

3. Materiał i metody

Zebrane dane dotyczą 75 pacjentów zakwalifikowanych do leczenia operacyjnego z powodu pojedynczego cienia okrągłego płuca.

W przeprowadzonym po zabiegu badaniu histopatologicznym nowotwór łagodny stwierdzono u 39 pacjentów, a zmiany o charakterze złośliwym u 36 osób. Zmienną zależną jest zatem wynik badania histopatologicznego.

Zmiennymi niezależnymi są:

– płeć;
– wiek;
– wyniki badań laboratoryjnych: OB, LDH, fibrynogen;
– laboratoryjne markery procesu nowotworowego: SCC, NSE, Ca 19-9, Ca 15-3, CEA, Cyfra 21-1;
– CT lung – wielkość cienia szacowana na podstawie skanu płucnego tomografii komputerowej klatki piersiowej;
– CT mediastinum − wielkość cienia szacowana na podstawie skanu śródpiersiowego tomografii komputerowej klatki piersiowej;
– lokalizacja zmiany w płacie górnym lub płucu prawym.

Drzewa klasyfikacyjne nie wymagają przeprowadzenia wstępnej selekcji zmiennych objaśniających, bo są one dobierane jednocześnie z budową drzewa.

Dla analizowanego zbioru danych zbudowano:

1) drzewo klasyfikacyjne CART;

2) modele zagregowane: *bagging*, *boosting*, *Random forests*;

3) modele wiązane (drzewo klasyfikacyjne + jeden model dodatkowy: liniowy model dyskryminacyjny (LDA), stabilny liniowy model dyskryminacyjny (sLDA), model regresji logistycznej (LR), metoda wektorów nośnych (SVM) oraz drzewo klasyfikacyjne + kilka modeli dodatkowych jednocześnie).

Obliczenia wykonano w środowisku R z wykorzystaniem pakietów: `rpart`, `ipred`, `ada`, `randomForest`, `e1071`.

Ze względu na zbyt małą liczbę obserwacji, nie ma możliwości wyodrębnienia próby testowej, dlatego też do oceny błędu klasyfikacji wykorzystano 10-częściowy sprawdzian krzyżowy (10-fold-CV).

Dodatkowo, dla każdego z uzyskanych modeli obliczono czułość, swoistość, dodatnią zdolność predykcyjną (PPV), ujemną zdolność predykcyjną (NPV) oraz wskaźnik J Youdena.

Czułość to odsetek pacjentów z nowotworem złośliwym, którzy zostali prawidłowo rozpoznani przez klasyfikator. Swoistość to odsetek pacjentów ze zmianą łagodną, którzy zostali prawidłowo zidentyfikowani przez klasyfikator.

Dodatnia wartość predykcyjna to odsetek pacjentów zaklasyfikowanych przez model do grupy zagrożonej nowotworem złośliwym, u których rzeczywiście stwierdzono zmianę złośliwą, a ujemna wartość predykcyjna to odsetek pacjentów

zaklasyfikowanych przez model do grupy ze zmianą łagodną, u których faktycznie zdiagnozowano nowotwór łagodny.

Wskaźnik J Youdena, obliczany jako J = czułość + swoistość − 1, jest miarą efektywności reguły decyzyjnej.

4. Wyniki i wnioski

W sposób najbardziej ogólny jakość reguł decyzyjnych, uzyskanych w wyniku zastosowania analizowanych metod, można ocenić, podając wartości wskaźnika J Youdena oraz odsetek poprawnych klasyfikacji.

Wyniki uzyskane z wykorzystaniem metody sprawdzania krzyżowego (10-CV) przedstawiono w tab. 1. Metody uporządkowano według wartości wskaźnika J Youdena.

Tabela 1. Ranking metod klasyfikacji pacjentów według wskaźnika J Youdena oraz odsetka poprawnych klasyfikacji

Metoda	Wskaźnik J Youdena	Odsetek poprawnych klasyfikacji
Bundling LDA + SVM (model wiązany: drzewo klasyfikacyjne + liniowy model dyskryminacyjny + metoda wektorów nośnych)	0,438	72,0
Bundling LDA (model wiązany: drzewo klasyfikacyjne + liniowy model dyskryminacyjny)	0,436	72,0
Bundling LR (model wiązany: drzewo klasyfikacyjne + regresja logistyczna)	0,417	70,7
Bundling LDA+LR (model wiązany: drzewo klasyfikacyjne + liniowy model dyskryminacyjny + regresja logistyczna)	0,333	66,7
Bundling LDA+ SVM +LR (model wiązany: drzewo klasyfikacyjne + liniowy model dyskryminacyjny + metoda wektorów nośnych + regresja logistyczna)	0,333	66,7
Boosting (wzmacnianie dokładności predykcji modelu zagregowanego)	0,331	66,7
Bundling sLDA (model wiązany: drzewo klasyfikacyjne + stabilny liniowy model dyskryminacyjny)	0,329	66,7
Random forests (lasy losowe)	0,327	66,7
Bagging (agregacja bootstrapowa)	0,276	64,0
Bundling SVM (model wiązany: drzewo klasyfikacyjne + metoda wektorów nośnych)	0,252	62,7
CART (pojedyncze drzewo klasyfikacyjne)	0,192	60,0

Źródło: obliczenia własne.

W diagnostyce medycznej istotna jest również poprawna klasyfikacja pacjentów do każdej z wyróżnionych grup ryzyka. Do oceny jakości badanych metod klasyfikacji wykorzystać można czułość, swoistość oraz dodatnią i ujemną zdolność predykcyjną. Uzyskane wyniki przedstawiono w tab. 2. Metody uporządkowano według ich czułości.

Tabela. 2. Ocena metod klasyfikacji pacjentów według miar jakości reguł predykcyjnych (w %)

Metoda	Czułość	Swoistość	Zdolność predykcyjna	
			dodatnia	ujemna
Bundling LR (model wiązany: drzewo klasyfikacyjne + regresja logistyczna)	75,0	66,7	67,5	25,7
Bundling LDA+SVM (model wiązany: drzewo klasyfikacyjne + liniowy model dyskryminacyjny + metoda wektorów nośnych)	69,4	74,4	71,4	27,5
Bundling LDA (model wiązany: drzewo klasyfikacyjne + liniowy model dyskryminacyjny)	66,7	76,9	72,7	28,6
Bundling LDA+LR (model wiązany: drzewo klasyfikacyjne + liniowy model dyskryminacyjny + regresja logistyczna)	66,7	66,7	64,9	31,6
Bundling LDA+SVM+LR (model wiązany: drzewo klasyfikacyjne + liniowy model dyskryminacyjny + metoda wektorów nośnych + regresja logistyczna)	66,7	66,7	64,9	31,6
Boosting (wzmacnianie dokładności predykcji modelu zagregowanego)	63,9	69,2	65,7	32,5
Bundling sLDA (model wiązany: drzewo klasyfikacyjne + stabilny liniowy model dyskryminacyjny)	61,1	71,8	66,7	33,3
Bundling SVM (model wiązany: drzewo klasyfikacyjne + metoda wektorów nośnych)	61,1	64,1	61,1	35,9
Random forests (lasy losowe)	58,3	74,4	67,7	34,1
Bagging (agregacja bootstrapowa)	58,3	69,2	63,6	35,7
CART (pojedyncze drzewo klasyfikacyjne)	50,0	69,2	60,0	40,0

Źródło: obliczenia własne.

Analizując wartości wskaźnika J Youdena oraz odsetki poprawnych klasyfikacji przedstawione w tabeli 1, stwierdzamy poprawę dokładności predykcji po zastosowaniu metody wiązania modeli zarówno w porównaniu z pojedynczym drzewem, jak i dwomaz metodami agregacji – *Random forests* i *bagging*. Najlepsze wyniki dają modele wiązane, uwzględniające liniowy model dyskryminacyjny (LDA) (także w połączeniu z metodą wektorów nośnych).

Pacjentów z nowotworem złośliwym najlepiej klasyfikuje reguła decyzyjna oparta na modelu wiązanym z jednym modelem dodatkowym – funkcją regresji logistycznej − czułość jest tutaj najwyższa i wynosi 75%. Niestety, reguła ta charakteryzuje się niższą swoistością.

Pacjentów z nowotworem łagodnym najlepiej rozpoznaje reguła oparta na modelu wiązanym uwzględniającym dwa modele dodatkowe – liniowy model dyskryminacyjny oraz metodę wektorów nośnych – swoistość wynosi 74,4%.

Wszystkie metody mają dość wysoką dodatnią zdolność predykcyjną (najwyższą – 72,7% obserwujemy w przypadku modelu wiązanego z liniową funkcją dyskryminacyjną).Z drugiej strony w każdym przypadku występuje bardzo niska ujemna zdolność predykcyjna, co jest zjawiskiem niekorzystnym.

Za wadę modeli wiązanych należy uznać brak możliwości zapisu reguł klasyfikacyjnych, każdy model to swego rodzaju czarna skrzynka, którą można wykorzystać wyłącznie do klasyfikacji obiektów. Kolejną wadą jest długi czas obliczeń przy większych zbiorach danych.

Last but not least, budując modele wiązane należy pamiętać o twierdzeniu NFL – nie ma sprecyzowanych zasad, co wybrać jako model dodatkowy; w zależności od badanego problemu najlepszym modelem może być liniowa funkcja dyskryminacyjna, regresja logistyczna, metoda wektorów nośnych czy kombinacja kilku metod.

Literatura

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

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

Efron B., Tibshirani R., *Improvements on Cross-Validation*: *The .632+ Bootstrap Method*, „Journal of the American Statistical Association" 1997, vol. 92, no. 438, s. 548-560.

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., *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*, Wyd. Naukowe PWN, Warszawa 2008.

Hothorn T., *Bundling classifiers with an application to glaucoma diagnosis*, Dissertation, Department of Statistics, University of Dortmund, Germany, 2003, http://eldorado.uni-dortmund.de:8080/bitstream/2003/2790/1/hothornunt.pdf.

Hothorn T., Lausen B., *Double-bagging: Combining classifiers by bootstrap aggregation*, „Pattern Recognition" 2003, 36, s. 1303-1309.

Kohavi R., *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*, „Proceedings of the Fourteenth International Joint Conference on artificial Intelligence", Montreal − Quebec 1995, vol. 2, s. 1137–1145.

Stefanowski J., *Algorytmy indukcji reguł decyzyjnych w odkrywaniu wiedzy*, Rozprawy nr 361, Wyd. Politechniki Poznańskiej, Poznań 2001.

Szlachcińska A., Witaszczyk A., Misztal M., *Zastosowanie drzew klasyfikacyjnych do identyfikacji rodzaju zmian u pacjentów z pojedynczym cieniem okrągłym płuca*, [w:] K. Jajuga, M. Walesiak (red.), *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 47, Taksonomia 16, Wydawnictwo AE we Wrocławiu, Wrocław 2009, s. 442-450.

Wolpert D.H., Macready W.G., *No Free Lunch Theorems for Optimization*, „IEEE Transactions on Evolutionary Computation" 1997, 1 (1), s. 62-68.

ON THE USE OF BUNDLING TO IMPROVE ACCURACY OF CLASSIFICATION OF PATIENTS WITH SOLITARY PULMONARY NODULES

Summary: *Bundling* was proposed by Hothorn as modification of *bagging*. The main idea is to use the *out-of-bag* (OOB) observations of a bootstrap sample to build classifiers of arbitrary type (i.e. LDA or SVM). The predictions of those classifiers are computed for the observations in the bootstrap sample and are used as predictors offered to a classification tree in addition to the original predictors. In the study *bundling* was applied to improve accuracy of classification of patients with solitary pulmonary nodules.