PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 107
Taksonomia 17 2010

Mariusz Kubus
Politechnika Opolska

DYSKRYMINACJA ZA POMOCĄ MODELU REGUŁ ŁĄCZONYCH

Streszczenie: Podejście wielomodelowe okazało się jednym z najskuteczniejszych narzędzi dyskryminacji. Friedman i Popescu [2005] zaproponowali wykorzystanie w charakterze funkcji bazowych reguły klasyfikacji postaci „jeśli koniunkcja warunków, to klasa". Metoda zaimplementowana w algorytmie RuleFit łączy idee podejścia wielomodelowego, indukcji reguł oraz regularyzowanej regresji liniowej.

Celem artykułu jest zweryfikowanie jakości klasyfikacji na danych rzeczywistych oraz zbadanie wybranych własności algorytmu RuleFit.

1. Wstęp

Podejście wielomodelowe okazało się jednym z najskuteczniejszych narzędzi dyskryminacji. Zapewnia znaczną poprawę dokładności klasyfikacji oraz większą stabilność wyników. Ogólnie rzecz ujmując, polega ono na budowie M modeli bazowych na różnych podpróbach uczących oraz agregowaniu ich wyników klasyfikacji. Różne podpróby uczące otrzymuje się przez losowy dobór obiektów lub system nadawania wag, lub wybór tylko niektórych zmiennych.

Model zagregowany można przedstawić w postaci liniowej:

$$F(\boldsymbol{x}) = a_0 + \sum_{m=1}^{M} a_m f_m(\boldsymbol{x}) , \tag{1}$$

gdzie: f_m są funkcjami bazowymi, współczynniki a_m interpretuje się jako wagi nadawane predykcjom poszczególnych funkcji bazowych, a wyraz wolny interpretować można jako domyślną regułę klasyfikacji (np. klasyfikacja na podstawie estymatorów prawdopodobieństw *a priori*). Bogata literatura w tym temacie relacjonuje głównie wyniki uzyskane w przypadku, gdy modele bazowe są drzewami klasyfikacyjnymi (zob. np. [Gatnar 2008]). Friedman i Popescu [2005] zaproponowali wykorzystanie w charakterze funkcji bazowych f_m reguł klasyfikacji, co jest kontynuacją badań w nurcie indukcji reguł rozwijającym się równolegle z drzewami klasyfikacyjnymi.

Przez reguły klasyfikacji rozumie się w artykule implikacje: *jeśli spełniona jest koniunkcja warunków, to przypisz obiekt do klasy*. Analogicznie do węzłów drzew

klasyfikacyjnych warunki mają na ogół postać równości dla cech nominalnych oraz nierówności dla pozostałych skal pomiaru. W regule nie musi też występować zestaw wszystkich zmiennych objaśniających. Znaczna część metod budująca modele w postaci zbioru reguł wykorzystuje schemat separuj-i-zwyciężaj (zob. np. [Fürnkranz 1999; Kubus 2009]). Zbiory reguł na ogół nie mają hierarchicznej struktury drzewa, co stwarza możliwość odkrycia innych wzorców w danych. Inne podejście do indukcji reguł przyjął Quinlan [1993] w algorytmie C4.5RULES, budując najpierw drzewo klasyfikacyjne, a następnie dekomponując je do postaci zbioru reguł (reprezentujących ścieżki od korzenia do liści) w celu uproszczenia modelu (*pruning*).

Ponieważ indukcja reguł jest metodą mającą wiele cech wspólnych z drzewami klasyfikacyjnymi (metoda eksploracyjna, adaptacyjna, dla zmiennych metrycznych i niemetrycznych, podobne miary jakości i techniki upraszczania modelu, a nawet postać modelu), również tu można wskazać wiele propozycji nawiązujących do podejścia wielomodelowego. Cohen i Singer [1999] proponują do konstrukcji reguł zastosować boosting. Inny system nadawania wag wraz z pewnymi ograniczeniami dotyczącymi struktury modelu zaproponowali Weiss i Indurkhya [2000] w algorytmie LRI. Sekwencyjną budowę reguł z nadawaniem im rangi i skracaniem (*shrinkage*) zaproponowali też Dembczyński, Kotłowski i Słowiński [2008].

W artykule zbadana będzie metoda łączenia reguł pochodząca od Friedmana i Popescu [2005], która łączy idee podejścia wielomodelowego, indukcji reguł oraz regularyzowanej regresji liniowej. Autorzy pokazali empirycznie na 100 sztucznie generowanych zbiorach danych obiecujące rezultaty swojej metody w porównaniu z tak uznawanymi metodami, jak: MART [Friedman 2001] czy Random Forests [Breiman 2001]. Celem artykułu jest zweryfikowanie jakości klasyfikacji tej metody na zbiorach danych rzeczywistych oraz zbadanie jej wybranych własności. Do analizy porównawczej wykorzystane będą zbiory danych standardowo stosowane w tym celu oraz inne metody dyskryminacji bazujące na podejściu wielomodelowym.

2. Model reguł łączonych

Przez model reguł łączonych (*rule ensembles*) rozumie się sumę reguł wygenerowanych przez kolejne modele bazowe również w postaci zbioru reguł. Ze względu na własności geometryczne (niekoniecznie rozłączne regiony decyzyjne w indukcji reguł) postać modelu reguł łączonych nie różni się właściwie od pojedynczego modelu bazowego. Cała różnica polega na sposobie uczenia, które w przypadku podejścia wielomodelowego można podsumować jako wielokrotne wykorzystanie informacji tkwiącej w próbie uczącej. Ponadto regułom przypisywane są wagi, co można wykorzystać do celów interpretacyjnych.

2.1. Algorytm RuleFit

Podejście Friedmana i Popescu [2005] do generowania reguł opiera się na dekompozycji drzew klasyfikacyjnych. Uzyskane z drzew reguły budują model o strukturze liniowej, którego parametry są następnie estymowane za pomocą regularyzowanej regresji liniowej. Autorzy zaimplementowali swą metodę w algorytmie RuleFit, który jest dostępny na stronie www-stat.stanford.edu/~jhf/R-RuleFit.html. W zagadnieniu dyskryminacji oprogramowany jest przypadek klasyfikacji binarnej.

Dane wejściowe to zbiór uczący:

$$\{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_N, y_N) : \boldsymbol{x}_i \in \boldsymbol{X} = (X_1, ..., X_p), y_i \in Y, i \in \{1, ..., N\}\}, \tag{2}$$

gdzie zmienne objaśniające mogą być metryczne lub niemetryczne, natomiast o Y będziemy zakładać dalej, że jest zmienną binarną zakodowaną wartościami –1 i 1. Parametrami procedury RuleFit są: liczba drzew klasyfikacyjnych M oraz średnia liczba liści $\bar{L}$.

Pierwszy krok polega na zbudowaniu M drzew klasyfikacyjnych $\{f_m(x)\}_1^M$ na różnych próbach uczących, które losowane są ze zbioru uczącego bez zwracania. Liczebność każdej wynosi $\min\{N/2;\ \ 100 + 6\sqrt{N}\}$. Z góry ograniczany jest też rozmiar drzew (liczba liści). W ten sposób kontroluje się złożoność modelu oraz skraca czas obliczeń. W przypadku pojedynczego drzewa klasyfikacyjnego zabieg taki nie gwarantuje wysokiej dokładności klasyfikacji. Daje natomiast dobre rezultaty w modelach agregowanych (zob. np. [Friedman, Hastie, Tibshirani 2000]), gdzie autorzy ograniczają głębokość drzew). W omawianej metodzie autorzy prezentują podejście polegające na uzmiennieniu liczby liści w kolejnych modelach bazowych. Jest ona zmienną losową:

$$t_m = 2 + [\gamma], \tag{3}$$

gdzie γ jest losowane z rozkładu wykładniczego z prawdopodobieństwem:

$$\Pr(\gamma) = \frac{\exp(\frac{-\gamma}{\bar{L} - 2})}{\bar{L} - 2}, \tag{4}$$

a $\bar{L}$ jest średnią liczbą liści. Takie podejście sprawia, że reguły różnią się długością i niektóre z nich są w stanie opisać interakcje wyższego rzędu.

Drugi krok polega na dekompozycji drzew na zbiór reguł. Część warunkowa reguły utożsamiana jest z koniunkcją warunków na ścieżce od korzenia do węzła (niekoniecznie końcowego). Zatem liczba reguł uzyskanych z dekompozycji drzew $\{f_m(x)\}_1^M$ wynosi:

$$K = \sum_{m=1}^{M} 2(t_m - 1), \tag{5}$$

gdzie t_m oznacza liczbę liści w m-tym drzewie. Konstruowany jest zatem model:

$$F(\boldsymbol{x}) = a_0 + \sum_{k=1}^{K} a_k r_k(\boldsymbol{x}) , \tag{6}$$

gdzie $r_k(\boldsymbol{x})$ zwracają wartość 1 lub 0 w zależności od tego, czy obiekt $\boldsymbol{x}$ jest opisany przez regułę czy nie.

W trzecim kroku parametry modelu (6) są estymowane przez minimalizację wyrażenia:

$$\{\hat{a}_k\}_0^K = \arg\min_{\{a_k\}_0^K} \left(\sum_{i=1}^{N} L\left(y_i, a_0 + \sum_{k=1}^{K} a_k r_k(\boldsymbol{x}_i) \right) + \lambda \sum_{k=1}^{K} |a_k| \right). \tag{7}$$

Oprócz minimalizacji ryzyka empirycznego uwzględniana jest kara za duże (bezwzględne) wartości parametrów. Funkcja straty ma postać:

$$L(y, F(\boldsymbol{x})) = [y - \max(-1; \min(1, F(\boldsymbol{x})))]^2 \tag{8}$$

i jest rekomendowana przez autorów jako odporna na błędnie sklasyfikowane obiekty w zbiorze uczącym (*mislabeled cases*). Funkcja kary wykorzystuje operator LASSO (*Least Absolute Shrinkage and Selection Operator*) [Tibshirani 1996], który ma tu korzystną własność ze względu na możliwości interpretacji. Prowadzi mianowicie do wyzerowania ok. 80-90% parametrów. Wpływ funkcji kary regulowany jest parametrem λ, którego oszacowanie odbywa się przez sprawdzanie krzyżowe. Zagadnienie (7) można przeformułować na problem programowania kwadratowego z ograniczeniami w postaci liniowych nierówności, lecz ze względu na dużą liczbę funkcji bazowych rozwiązanie w zamkniętej postaci nie jest możliwe. Dobrą aproksymację rozwiązania można uzyskać, stosując metody gradientowe (zob. [Friedman, Popescu 2004]) lub algorytm sekwencyjnego doboru reguł przypominający krokowe modelowanie addytywne (*forward stagewise additive modeling*) ze skracaniem (*shrinkage*) (zob. np. [Hastie, Tibshirani, Friedman 2001]).

2.2. Podejście hybrydowe

Friedman i Popescu [2005] zaproponowali też model, w którym oprócz reguł modelem bazowym jest również funkcja liniowa. W celu zapewnienia odporności na obserwacje oddalone oryginalne zmienne (metryczne) są transformowane według formuły:

$$l_j(X_j) = \min(\delta_j^+, \max(\delta_j^-, X_j)) \quad \text{dla} \quad j \in \{1, ..., p\} , \tag{9}$$

gdzie: δ_j^-, δ_j^+ są odpowiednio β i $(1-\beta)$ kwantylami rozkładów empirycznych zmiennych X_j. Parametr β ustalany jest arbitralnie, przy czym autorzy sugerują $\beta = 0,025$ (opcja domyślna w procedurze RuleFit). Model ma teraz postać:

$$F(\boldsymbol{x}) = a_0 + \sum_{k=1}^{K} a_k r_k(\boldsymbol{x}) + \sum_{j=1}^{p} b_j l_j(X_j), \tag{10}$$

gdzie parametry wyznaczane są przez minimalizację:

$$\left(\{\hat{a}_k\}_0^K, \{\hat{b}_j\}_1^p\right) =$$

$$= \underset{\{a_k\}_0^K, \{b_j\}_1^p}{\arg\min} \left(\sum_{i=1}^{N} L\left(y_i, a_0 + \sum_{k=1}^{K} a_k r_k(\boldsymbol{x_i}) + \sum_{j=1}^{p} b_j l_j(x_{ij}) \right) + \lambda \left(\sum_{k=1}^{K} |a_k| + \sum_{j=1}^{p} |b_j| \right) \right). \tag{11}$$

3. Badania empiryczne

W badaniach wykorzystano zbiory danych z *UCI Repository of Machine Learning* (www.ics.uci.edu/~mlearn/MLRepository.html). W celu estymacji błędu klasyfikacji zbiory podzielono losowo na próbę uczącą i testową (1/3 zbioru uczącego).

3.1. Porównanie błędów klasyfikacji

Tabela 1 przedstawia porównanie błędów klasyfikacji z kilkoma metodami reprezentującymi podejście wielomodelowe. Algorytm SLIPPER [Cohen, Singer 1999] wykorzystuje metodę *boosting* w indukcji reguł, natomiast AdaBoost [Freund, Schapire 1996] oraz Random Forests [Breiman 2001] budują model zagregowany na podstawie modeli bazowych w postaci drzew klasyfikacyjnych i są uważane powszechnie za jedne z najskuteczniejszych metod dyskryminacji ze względu na dokładność klasyfikacji.

Tabela 1. Błąd klasyfikacji (w %) szacowany na zbiorze testowym

Zbiory	Rule Fit	SLIPPER	AdaBoost	Random Forests
breast cancer	3,68	5,26	2,11	2,74
credit german	26,73	23,72	20,72	21,63
heart-disease C	19,80	22,77	20,79	19,70
ionosphere	8,55	9,40	7,69	8,66
pima	22,27	24,22	24,22	23,24
sonar	21,43	25,71	18,57	26,86

Źródło: obliczenia własne.

W algorytmie SLIPPER liczba iteracji ustalana była za pomocą sprawdzania krzyżowego, przy czym maksymalna to 100. W algorytmie AdaBoost ustalono liczbę drzew na 100 (zob. [Breiman 1998]). W metodzie Random Forests liczbę generowanych drzew ustalono na 500, a liczbę zmiennych losowanych w celu wyznaczenia podziałów w węzłach ustalano jako przybliżenie pierwiastka z liczby zmiennych objaśniających [Breiman 2001]. Ponadto ze względu na losowość będą-

cą u podstaw tej metody błąd klasyfikacji liczony był jako średnia z trzydziestu błędów uzyskanych przez trzydzieści wygenerowanych modeli (na różnych próbach uczących). Wyniki z tabeli 1 podsumowano w postaci bilansów „wygrana-przegrana-remis" (tab. 2). „Wygrana" oznacza tu liczbę zbiorów, dla których RuleFit dał mniejszy błąd klasyfikacji od algorytmu w kolumnie tabeli.

Tabela 2. Bilanse „wygrana-przegrana-remis" podsumowujące wyniki z tab. 1

	SLIPPER	AdaBoost	Random Forests
Rule Fit	5-1-0	2-4-0	3-3-0

Źródło: obliczenia własne.

3.2. Porównanie dwóch postaci modelu

W kolejnym badaniu porównano błędy klasyfikacji dla modeli: w postaci reguł (6) oraz z dołączoną liniową funkcją dyskryminacyjną (10) (tab. 3). Nie stwierdzono istotnych różnic wielkości błędów. W procedurze RuleFit wybór postaci modelu dokonuje się przez ustawienie parametru model.type.

Tabela 3. Błąd klasyfikacji (w %) szacowany na zbiorze testowym dla dwóch postaci modelu

Zbiory	Model (6)	Model (10)	Zbiory	Model (6)	Model (10)
breast cancer	3,68	3,68	*ionosphere*	8,55	8,55
credit german	26,73	26,43	*pima*	22,27	22,66
heart-disease C	19,80	19,80	*sonar*	21,43	21,43

Źródło: obliczenia własne.

3.3. Stabilność wyników

Celem kolejnego badania była ocena stabilności wyników uzyskiwanych metodą reguł łączonych. Znany jest fakt, że modele agregujące drzewa klasyfikacyjne poprawiają stabilność (zob. np. [Gatnar 2008]). Z kolei SLIPPER wykorzystujący metodę boosting nie gwarantuje poprawy stabilności w porównaniu z klasyczną indukcją reguł schematem separuj-i-zwyciężaj (zob. [Kubus 2008]). Ponieważ RuleFit łączy w swej idei drzewa i reguły, interesujące wydaje się sprawdzenie tej własności. Badanie przeprowadzono według następującego schematu. Oryginalny zbiór danych dzielono losowo (bez zwracania) na dwa (w przybliżeniu) równoliczne podzbiory U_{k1} oraz U_{k2}. Następnie budowano model na podstawie zbioru U_{k1} i estymowano błąd klasyfikacji $e(U_{k2})$ na zbiorze U_{k2} oraz odwrotnie. Powyższe czynności wykonano 30 razy, uzyskując w ten sposób próbę różnic błędów: $\Delta e_k = e(U_{k2}) - e(U_{k1})$ ($k \in \{1, ..., 30\}$), na której testowano hipotezę o wartości przeciętnej równej zeru. Obliczeń dokonano dla wybranych trzech zbiorów (tab. 5).

Tabela 4. Wyniki testu średniej względem stałej wartości równej zeru

Zbiory	Średnia	Odchylenie standardowe	Błąd standardowy	*t*	*p*
Pima	0,416667	2,879321	0,525690	0,792609	0,4344
ionosphere	–0,066991	4,189771	0,764944	–0,087577	0,9308
credit german	–0,227586	2,357192	0,437720	–0,519936	0,6072

Źródło: obliczenia własne.

W żadnym przypadku nie było podstaw do odrzucenia hipotezy zerowej o rozkładzie normalnym w teście Shapiro-Wilka. Uzyskano następujące wyniki: *Pima* (W = 0,96872; p = 0,5048); *ionosphere* (W = 0,95501; p = 0,2298); *credit german* (W = 0,97738; p = 0,7680). Wobec spełnionych założeń przeprowadzono test średniej względem stałej wartości równej zeru (tab. 4) i uzyskano dość wysokie istotności, co świadczy o stabilności wyników uzyskiwanych algorytmem RuleFit.

3.4. Odporność na zmienne nieistotne

W następnym badaniu sprawdzano odporność metody reguł łączonych na obecność w danych zmiennych nieistotnych. W eksperymencie wykorzystano zbiór *Pima*, gdzie dołączono 30 zmiennych z generowanych losowo rozkładów (w każdej klasie według tego samego schematu): (a) standaryzowanego rozkładu normalnego (zbiór *Pima (a)*); (b) mieszanki rozkładów normalnych: 1/4 obserwacji z N(0,1) oraz pozostałe z N(5,1) (zbiór *Pima (b)*). W tabeli 5 podano błędy klasyfikacji uzyskane algorytmem RuleFit oraz dla porównania błędy uzyskane algorytmami: AdaBoost, SLIPPER i CART. Jak widać, RuleFit jest znacznie bardziej czuły na występowanie zmiennych nieistotnych.

Tabela 5. Błędy klasyfikacji (w %) szacowane na zbiorze testowym dla zbiorów *Pima* ze zmiennymi nieistotnymi

Zbiory	Rule Fit	AdaBoost	SLIPPER	CART
Pima (a)	27,34	22,27	22,27	21,87
Pima (b)	29,30	21,87	23,83	21,87

Źródło: obliczenia własne.

4. Wnioski

Algorytm RuleFit będący kontynuacją badań nad podejściem wielomodelowym w indukcji reguł wykazuje przewagę nad algorytmem SLIPPER w dokładności klasyfikacji i stabilności modelu. Jest jednak bardziej wrażliwy na zmienne nieistotne, co ogranicza jego automatyczne stosowanie w typowych zadaniach *data mining*, gdzie zmienne często nie są dobierane do modelu merytorycznie, a raczej badacz poszukuje nieoczekiwanych, systematycznych relacji w danych. W porównaniu

z metodami agregującymi drzewa klasyfikacyjne RuleFit daje porównywalne błędy klasyfikacji z Random Forests, lecz większe od AdaBoost.

Na podstawie przeprowadzonych badań na rzeczywistych zbiorach danych należy stwierdzić, że próba podejścia hybrydowego łączącego reguły z liniową funkcją dyskryminacyjną nie przyniosła zamierzonych przez autorów rezultatów.

Literatura

Breiman L., *Arcing classifiers*, „Annals of Statistics" 1998 no 26.

Breiman L., *Random forests*, „Machine Learning" 2001 no 45.

Cohen W.W., Singer Y., *A Simple, Fast, and Effective Rule Learner*, Proceedings of Ann. Conf. of American Association for Artif. Intelligence, 1999.

Dembczyński K., Kotłowski W., Słowiński R., *A General Framework for Learning an Ensemble of Decision Rules*, [w:] *LeGo '08: From Local Patterns to Global Models, ECML/PKDD 2008 Workshop,* Antwerp, Belgium 2008.

Freund Y., Schapire R.E., *Experiments with a New Boosting Algorithm,* Proceedings of the 13th Intern. Conf. on Machine Learning, Morgan Kaufmann, 1996.

Friedman J.H., *Greedy function approximation: a gradient boosting machine,* „Annals of Statistics" 2001 no 29.

Friedman J.H., Hastie T., Tibshirani R., *Additive logistic regression: a statistical view of boosting*, „Annals of Statistics" 2000 no 28(2).

Friedman J.H., Popescu B.E., *Gradient Directed Regularization for Linear Regression and Classification*, (Technical Report). Dept. of Statistics, Stanford University, 2004.

Friedman J.H., Popescu B.E., *Predictive Learning Via Rule Ensembles*, (Technical Report). Dept. of Statistics, Stanford University, 2005.

Fürnkranz J., *Separate-and-conquer rule learning*, „Artif. Intelligence Review" 1999 no 13(1).

Gatnar E., *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*, PWN, Warszawa 2008.

Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning: Data Mining, Inferance, and Prediction*, Springer, New York 2001.

Kubus M., *Porównanie indukcji reguł z wybranymi metodami dyskryminacji*, [w:] *Klasyfikacja i analiza danych – teoria i zastosowania*, Taksonomia 16, K. Jajuga, M. Walesiak (red.), Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 47, UE, Wrocław 2009.

Kubus M., *The Analysis of Some Properties of SLIPPER Algorithm*, 27th Conference on Multivariate Statistical Analysis, University of Łódź (w druku), 2008.

Quinlan J.R., *C4.5 Programs for Machine Learning*, Morgan Kaufmann, San Mateo 1993.

Tibshirani R., *Regression shrinkage and selection via the lasso*, „J.Royal. Statist. Soc. B." 1996 no 58.

Weiss S., Indurkhya N., *Lightweight Rule Induction*, Proceedings of the 17th International Conference on Machine Learning, Morgan Kaufmann, 2000.

DISCRIMINATION USING RULE ENSEMBLES

Summary: Ensembles have turned out to be one of the most effective tools of discrimination. Friedman and Popescu [2005] proposed to use the classification rules in the form "*if conjunction of conditions then class*" as a base functions. Their method implemented in RuleFit combines the ideas of ensembles, rules induction and regularized linear regression.

The goal of this paper is to verify the classification accuracy of RuleFit on real world data and to test some of its properties.