PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 7 (1207) 2008
TAKSONOMIA 15
Klasyfikacja i analiza danych – teoria i zastosowania

Małgorzata Misztal
Uniwersytet Łódzki

ZAGREGOWANE MODELE DYSKRYMINACYJNE I REGRESYJNE W PROGNOZOWANIU CZASU POBYTU NA OIOM PACJENTÓW Z CHOROBĄ WIEŃCOWĄ*

1. Wstęp

W pracy [Misztal 2007] podjęto próbę analizy i prognozowania czasu pobytu na Oddziale Intensywnej Opieki Medycznej (OIOM) 4642 pacjentów poddanych operacyjnemu leczeniu choroby niedokrwiennej serca w 12 klinikach kardiochirurgicznych w Polsce w latach 2003-2005.

Celem referatu było porównanie ośrodków z punktu widzenia czasu pobytu pacjenta na OIOM oraz wyodrębnienie zestawu przedoperacyjnych czynników ryzyka mających wpływ na przedłużony czas pobytu na OIOM.

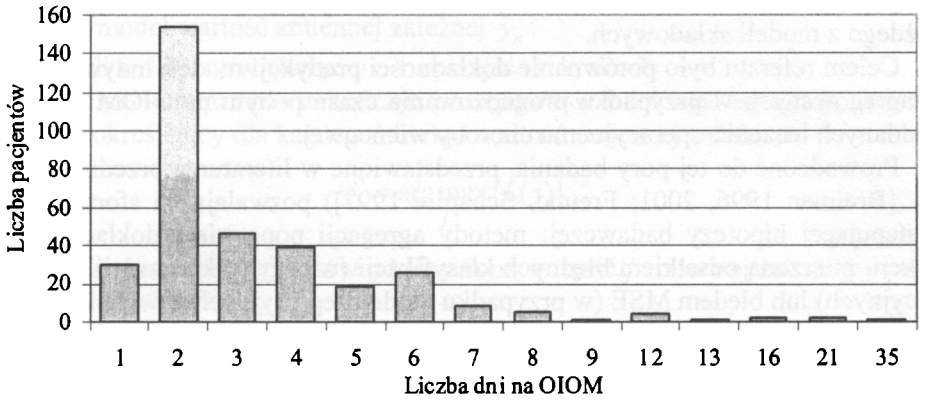
Uzyskane wyniki dla pacjentów ogółem prowadziły do konkluzji, że badane kliniki różnią się istotnie z punktu widzenia czasu pobytu pacjenta na OIOM, a zatem szczegółowe analizy należy wykonać dla każdego ośrodka oddzielnie.

W kolejnych badaniach skupiono się na prognozowaniu czasu pobytu na OIOM na przykładzie ośrodka łódzkiego. W latach 2003-2005 w łódzkiej klinice operowano 337 pacjentów. Rozkład czasu pobytu na OIOM w Łodzi przedstawiono na rys. 1. Jak widać, rozkład ten charakteryzuje się występowaniem silnej asymetrii prawostronnej.

Jako zmienną zależną w budowanych modelach rozpatrywano czas pobytu na OIOM w dniach, zmienną binarną OIOM_kat (OIOM_kat=0, gdy czas pobytu na OIOM <= 2 dni oraz OIOM_kat=1, gdy czas pobytu na OIOM > 2 dni) i zmienną transformowaną logOIOM będącą logarytmem naturalnym z czasu pobytu na OIOM.

W wyniku przeprowadzonych analiz jednowymiarowych z 66 zmiennych opisujących stan pacjenta przed operacją wyodrębniono 14 czynników ryzyka istotnie wpływających na przedłużony czas pobytu na OIOM (por. [Misztal 2007]).

* Pracę wykonano w ramach realizacji tematu badawczego nr 505/596 pt. *Metody agregacji modeli dyskryminacyjnych i regresyjnych i ich zastosowania*, finansowanego z dotacji na badania własne w UŁ w 2007 r.



Rys. 1. Czas pobytu na OIOM pacjentów ośrodka łódzkiego
Źródło: opracowanie własne.

Spośród wielu analizowanych metod statystyki wielowymiarowej zdecydowanie najlepsze wyniki, w sensie dokładności klasyfikacji, uzyskano dla modelu regresji logistycznej, liniowej funkcji dyskryminacyjnej Fishera, algorytmu CRUISE budującego drzewa klasyfikacyjne (por. [Kim, Loh 2001]) i algorytmu LOTUS budującego drzewa regresji logistycznej (por. [Chan, Loh 2004]) – zob. tab. 1.

Tabela 1. Wyniki klasyfikacji pacjentów (binarna zmienna zależna)

Metoda	Grupa	% poprawnych klasyfikacji
Regresja logistyczna	OIOM ≤ 2 dni	79,55
	OIOM > 2 dni	50,00
Analiza dyskryminacyjna	OIOM ≤ 2 dni	75,82
	OIOM > 2 dni	50,00
CRUISE – drzewo klasyfikacyjne	OIOM ≤ 2 dni	91,30
	OIOM > 2 dni	60,13
LOTUS – drzewo regresji logistycznej	OIOM ≤ 2 dni	85,33
	OIOM > 2 dni	42,48

Źródło: [Misztal 2007].

W przypadku algorytmu GUIDE, tworzącego drzewa regresyjne (por. [Loh 2002]), najmniejszą wartość błędu MSE uzyskano dla zmiennej objaśnianej transformowanej, będącej logarytmem naturalnym z czasu pobytu na OIOM (MSE = 0,3554).

Jak łatwo zauważyć, uzyskane wyniki nie są satysfakcjonujące, zwłaszcza w przypadku pacjentów o dłuższym czasie pobytu na OIOM. Przyczyną może tu być między innymi brak stabilności utworzonych modeli. Rozwiązaniem tego problemu może być agregacja indywidualnych modeli w jeden model zagregowany, co

powinno prowadzić do uzyskania ogólnego błędu predykcji mniejszego niż błąd każdego z modeli składowych.

Celem referatu było porównanie dokładności predykcji modeli indywidualnych i zagregowanych w przypadku prognozowania czasu pobytu na OIOM pacjentów poddanych leczeniu operacyjnemu choroby wieńcowej.

Prowadzone do tej pory badania, przedstawione w literaturze przedmiotu (por. np. [Breiman 1996, 2001; Freund, Schapire 1997]) pozwalają na sformułowanie następującej hipotezy badawczej: metody agregacji poprawiają dokładność predykcji, mierzoną odsetkiem błędnych klasyfikacji (w przypadku modeli dyskryminacyjnych) lub błędem MSE (w przypadku modeli regresyjnych).

2. Metody agregacji

Agregacja modeli dotyczyć może zarówno drzew klasyfikacyjnych jak i regresyjnych. Najogólniej polega ona na wyodrębnieniu ze zbioru uczącego V prób uczących, na podstawie których budowane są modele drzew $D_1(\mathbf{x}), \ldots, D_V(\mathbf{x})$, łączone następnie w jeden model zagregowany $D^*(\mathbf{x})$ (por. np. [*Metody statystycznej*... 2004]).

W modelu zagregowanym wartość zmiennej zależnej y określa się dla klasyfikacji na podstawie zasady majoryzacji – wybierana jest ta, która najczęściej została wskazana przez modele składowe:

$$\hat{D}^*(\mathbf{x}) = \arg\max_y \left\{ \sum_{v=1}^{V} I(\hat{D}_v(\mathbf{x}) = y) \right\} \tag{1}$$

lub – dla regresji – uśredniając wyniki z modeli składowych:

$$\hat{D}^*(\mathbf{x}) = \frac{1}{V} \sum_{v=1}^{V} \hat{D}_v(\mathbf{x}) . \tag{2}$$

Wśród metod agregacji można wyróżnić metody oparte na losowaniu ze zbioru uczącego kolejnych prób uczących (m.in. *bagging* [Breiman 1996; 1998] i *boosting* [Freund, Shapire 1997]) oraz metody oparte na losowym wyborze zmiennych (m.in. *random forests* [Breiman 2001]).

W metodzie *bagging* (*bootstrap aggregation*) 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 wynosi: $w_i = \frac{1}{N}$. Algorytm metody *bagging* dla klasyfikacji można opisać następująco (por. [Gatnar 2003]):

1. Ustalić liczbę prób uczących V.
2. Przyjąć $v = 1$.
3. Wylosować ze zwracaniem ze zbioru uczącego N obiektów tworzących próbę U_v.

4. Zbudować drzewo klasyfikacyjne D_v i zapamiętać dla każdego obiektu ustaloną przez model wartość zmiennej zależnej $\hat{y}_n$.

5. Jeżeli $v < V$, to zwiększyć v o 1 i przejść do kroku 3.

6. W przeciwnym razie zakończyć postępowanie i dokonać agregacji, tworząc model D^* określający dla każdego obiektu wartość zmiennej zależnej:

$$j^* = \arg\max_j \{L(j)\}, \tag{3}$$

gdzie $L(j)$ oznacza liczbę głosów oddanych na wartość j zmiennej zależnej y przez modele składowe.

7. Przydzielić obiekt do klasy, która była najczęściej wybierana przez kolejne modele składowe.

W metodzie *boosting* z kolei dokonywane jest ważenie obiektów ze zbioru uczącego, przy czym wagi określają prawdopodobieństwo wyboru obiektu do próby uczącej U_v, zmieniające się w zależności od wyników klasyfikacji dla poprzedniej próby U_{v-1}. Algorytm metody *boosting* dla klasyfikacji można opisać następująco (por. [Gatnar 2003]):

1. Ustalić liczbę prób uczących V.

2. Przyjąć $v = 1$ i ustalić początkowe wagi obiektów:

$$\forall_{n=1,\ldots,M} w_1^{(n)} = \frac{1}{N}. \tag{4}$$

3. Wylosować ze zwracaniem ze zbioru uczącego N obiektów tworzących próbę U_v.

4. Zbudować drzewo klasyfikacyjne D_v.

5. Obliczyć błąd klasyfikacji dla modelu D_v jako:

$$e(D_v) = \sum_{n=1}^{N} w_v^{(n)} I(\hat{y}_n \neq y_n). \tag{5}$$

6. Jeżeli $e(D_v) = 0$, to zakończyć postępowanie i dokonać agregacji, tworząc model D^*.

7. W przeciwnym wypadku zmodyfikować wagi każdego z obiektów:

$$w_{v+1}^{(n)} = \begin{cases} \dfrac{w_v^{(n)}}{2e(D_v)} & \text{dla } \hat{y}_n \neq y_n \\ \dfrac{w_v^{(n)}}{2(1-e(D_v))} & \text{dla } \hat{y}_n = y_n \end{cases}. \tag{6}$$

8. Zwiększyć v o 1 i przejść do kroku 3.

Agregacja modeli cząstkowych odbywa się także na zasadzie majoryzacji, przy czym liczba głosów obliczana jest jako:

$$L(j) = \frac{1}{2}\sum_{v} \log\left(\frac{1-e(D_v)}{e(D_v)}\right) I(\hat{y}_v = j). \tag{7}$$

Algorytm metody *random forest* można zapisać w sposób następujący (por. [Koronacki, Ćwik 2005]):

1. Ustalić liczbę prób uczących V.
2. Przyjąć $v = 1$.
3. Wylosować ze zwracaniem ze zbioru uczącego N obiektów tworzących próbę U_v, na podstawie której będzie budowane drzewo klasyfikacyjne.
4. W każdym węźle budowanego drzewa wylosować niezależnie m spośród p cech opisujących obiekty (przy czym $m << p$; zwykle przyjmuje się $m = \sqrt{p}$); podział dokonywany jest na podstawie wylosowanych m cech, spośród których wybierana jest najlepsza zmienna do podziału.
5. Zbudować drzewo klasyfikacyjne D_v bez przycinania i, jeśli to możliwe, uzyskując liście zawierające obiekty tylko z jednej klasy. Zapamiętać dla każdego obiektu ustaloną przez model wartość zmiennej zależnej $\hat{y}_n$.
6. Jeżeli $v < V$, to zwiększyć v o 1 i przejść do kroku 3.
7. W przeciwnym razie zakończyć postępowanie i dokonać agregacji, tworząc model D^*.
8. Przydzielić obiekt do klasy, która była najczęściej wybierana przez kolejne modele składowe.

Jak zatem widać, metody *bagging* i *random forest* wykorzystują dużą liczbę niezależnych drzew, a redukcja błędu predykcji następuje poprzez uśrednianie błędów predykcji pojedynczych drzew. W metodzie *boosting* z kolei tworzona jest sekwencja pojedynczych drzew, z których każde wyjaśnia zmienność niewyjaśnioną przez wcześniejsze drzewa.

Ideę podobną do *boostingu* zastosowali również twórcy algorytmu BART (*Bayesian Additive Regression Trees*) – por. [Chipman, George, McCulloch 2006]. Model zagregowany można tu ogólnie zapisać jako:

$$\begin{aligned} Y &= f(x) + \varepsilon \\ f(x) &= g_1(x) + g_2(x) + \ldots + g_v(x) \end{aligned}, \tag{8}$$

a zatem mamy do czynienia z sumą V drzew regresyjnych $g_i(\mathbf{x})$, z których każde wyjaśnia małą i inną część łącznej wariancji. Szczegółowy opis procedury przedstawiają [Chipman, George, McCulloch 2006].

Omówione pokrótce algorytmy wykorzystano do prognozowania czasu pobytu na OIOM pacjentów z chorobą wieńcową leczonych operacyjnie.

3. Wyniki i wnioski

Niezbędne obliczenia wykonano w środowisku R, wykorzystując następujące pakiety:

1) „*ada*" (algorytm *Real AdaBoost* – [Friedman, Hastie, Tibshirani 2000]);
2) „*randomForest*" [Breiman 2001];
3) „*BayesTree*" (algorytm BART – [Chipan, George, McCulloch 2006]);
4) „*ipred*" (*bagging* – [Breiman 1996, 1998]);
5) „*mboost*" (algorytm *gamboost – Gradient Boosting with Component-wise Smoothing Splines* – [Buhlmann, Yu 2003]).

W każdym z modeli dokonano agregacji 100 drzew. Wyniki przedstawiają tab. 2 i 3.

Tabela 2. Wyniki klasyfikacji pacjentów (zagregowane modele dyskryminacyjne)

Metoda	Grupa	% poprawnych klasyfikacji
Boosting	OIOM ≤ 2 dni	87,57
	OIOM > 2 dni	86,58
Bagging	OIOM ≤ 2 dni	98,91
	OIOM > 2 dni	98,04
Random Forest	OIOM ≤ 2 dni	100,00
	OIOM > 2 dni	97,39

Źródło: obliczenia własne.

Tabela 3. Błąd predykcji MSE (zagregowane modele regresyjne)

Metoda	MSE
Boosting	0,2992
Bagging	0,2133
Random Forest	0,0869
BART	0,2662

Źródło: obliczenia własne.

Analiza uzyskanych wyników prowadzi do konkluzji, że metody agregacji drzew decyzyjnych poprawiają dokładność predykcji czasu pobytu na OIOM w stosunku do pojedynczych drzew.

Zaobserwować można znaczną redukcję odsetka błędnych klasyfikacji dla zagregowanych modeli dyskryminacyjnych, zwłaszcza w przypadku grupy pacjentów z przedłużonym czasem pobytu na oddziale intensywnej opieki medycznej. Podobnie agregacja drzew regresyjnych prowadzi do znacznie mniejszego błędu MSE niż dla najlepszego z modeli pojedynczych. W obu przypadkach najlepsze wyniki dały modele zagregowane metodą *random forest*.

Reasumując, zagregowane modele dyskryminacyjne i regresyjne można z powodzeniem stosować do prognozowania czasu pobytu na OIOM pacjentów z chorobą wieńcową leczonych operacyjnie. Należy jednak zwrócić uwagę, że model zagregowany jest swego rodzaju „czarną skrzynką", a zatem brakuje możliwości zapisu reguł klasyfikacyjnych i identyfikacji czynników wpływających na przedłużony czas pobytu na OIOM. Jest to pewną wadą tych modeli w porównaniu z pojedynczymi drzewami.

Literatura

Breiman L. (1996), *Bagging Predictors*, „Machine Learning", 24(2), s. 23-140.

Breiman L (1998), *Arcing Classifiers*, „The Annals of Statistics", 26(3), s. 801-824.

Breiman L. (2001), *Random Forests*, „Machine Learning" 45(1), s. 5-32.

Buhlmann P., Yu B. (2003), *Boosting with the L_2 Loss: Regression and Classification*, „Journal of the American Statistical Association", 98, s. 324-339.

Chan K.-Y., Loh W.-Y. (2004), *LOTUS: An Algorithm for Building Accurate and Comprehensible Logistic Regression Trees*, „Journal of Computational and Graphical Statistics", vol. 13, issue 4, s. 826-852.

Chipman H.A., George E.I., McCulloch R.E. (2006), *BART: Bayesian Additive Regression Trees*, Technical report, University of Chicago.

Freund Y., Schapire R.E. (1997), *A Decision-theoretic Generalization of On-line Learning and an Application to Boosting*, „Journal of Computer and System Sciences" 55, s. 119-139.

Friedman J., Hastie T., Tibshirani R. (2000), *Additive Logistic Regression: A Statistical View of Boosting*, „The Annals of Statistics", vol. 28, nr 2, s. 337-407.

Gatnar E. (2003), *O pewnej metodzie redukcji błędu klasyfikacji*, [w:] Taksonomia 10, red. K. Jajuga, M. Walesiak, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 988, AE, Wrocław, s. 245-253.

Kim H., Loh W.-Y. (2001), *Classification Trees With Unbiased Multiway Splits*, „Journal of the American Statistical Association" 96, s. 598-604.

Koronacki J., Ćwik J. (2005), *Statystyczne systemy uczące się*, WNT, Warszawa.

Loh W.-Y. (2002), *Regression Trees with Unbiased Variable Selection and Interaction Detection*, „Statistica Sinica", vol. 12, s. 361-386.

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

Misztal M. (2007), *Wybrane metody analizy i prognozowania czasu pobytu na OIOM pacjentów z chorobą wieńcową*, [w:] Taksonomia 14, red. K. Jajuga, M. Walesiak, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1169, AE, Wrocław, s. 288-296.

AGGREGATION OF DISCRIMINANT AND REGRESSION MODELS FOR PREDICTION OF ICU LENGTH OF STAY AMONG PATIENTS WITH CAD

Summary

In the work [Misztal 2007] the classification and regression trees were applied to indicate preoperative risk factors associated with prolonged ICU stay among patients with coronary artery disease treated surgically. The results were not satisfying. The reason could be that single tree was not a stable classifier. The aggregation of classification or regression trees into the ensembles of classifiers can lead to a more accurate prediction.

In the paper aggregated tree-based models (*bagging*, *boosting*, *random forest*, *BART*) are applied to improve the accuracy of the prediction of ICU length of stay among patients with CAD.