ZESZYTY NAUKOWE POLITECHNIKI BIAŁOSTOCKIEJ 2007
Informatyka – Zeszyt 2

Małgorzata Krętowska[1]

LASY LOSOWE – OCENA JAKOŚCI PROGNOSTYCZNEJ CECH

Streszczenie: W pracy bezwzględny błąd predykcji jest wykorzystywany do oceny jakości prognostycznej poszczególnych cech. Narzędzie prognostyczne – lasy losowe – jest konstruowane w celu uzyskania estymatora funkcji przeżycia. Jest on następnie porównywany z estymatorem funkcji przeżycia Kaplana-Meiera, utworzonym przy założeniu jednorodności populacji. Elementem składowym lasów są dipolowe drzewa przeżycia. Zastosowanie dipolowej funkcji kryterialnej pozwala wykorzystać niepełną informację o czasie zajścia porażki, pochodzącą z obserwacji obciętych.

Słowa kluczowe: lasy losowe, analiza przeżyć, bezwzględny błąd predykcji

1. Wstęp

Ocena stanu zdrowia pacjenta oparta jest z reguły na weryfikacji pewnych cech (np. testów laboratoryjnych), odpowiedzialnych za występowanie określonego schorzenia. Dzisiejszy stan wiedzy o wielu chorobach pozwala dosyć precyzyjnie ocenić obecność czy też nasilenie choroby. Mamy jednak bardzo wiele poważnych schorzeń, których źródła nie znamy. W takich sytuacjach postawienie właściwej diagnozy jest zadaniem bardzo trudnym. Pojawia się wówczas potrzeba weryfikacji posiadanej wiedzy i ocena poszczególnych cech pacjenta pod kątem ich wpływu na analizowaną chorobę.

W analizie przeżyć chcemy zbadać wpływ cech na przeżycie pacjentów. Przez pojęcie przeżycie rozumiemy tu czas do zajścia porażki. Mianem porażki określamy zgon pacjenta, czy też nawrót choroby, przy czym zdarzeniem początkowym, od którego rozpoczynamy odliczanie czasu, jest z reguły rozpoznanie choroby lub operacja. Analizę tego typu danych utrudnia specyfika tych danych. Nie posiadamy dokładnych informacji o czasie wystąpienia porażki u wszystkich pacjentów. Są to tzw. obserwacje obcięte. Zmienna czasowa t występująca w opi-

[1] Wydział Informatyki, Politechnika Białostocka, Białystok

sie pacjenta oznacza tutaj jedynie czas, do którego porażka nie miała miejsca. W przypadku obserwacji pełnej zmienna t określa dokładny czas porażki.

Występowanie obserwacji obciętych uniemożliwia w praktyce stosowanie standardowych modeli regresyjnych, jak też ocenę dopasowania modeli do danych empirycznych poprzez analizę np. współczynnika determinacji. Modelem, który najczęściej jest wykorzystywany do analizy tego typu danych, jest model proporcjonalnych hazardów Cox'a [8]. Dosyć restrykcyjne założenia tego modelu wymuszają rozwój nowych, alternatywnych metod analizy danych przeżycia. Wykorzystywane są tutaj zarówno sieci neuronowe [1, 9], jak też drzewa regresyjne [7] czy modele złożone, takie jak np. lasy losowe [6, 12]. Rozwój tych ostatnich ma na celu stabilizację wyników uzyskanych przy użyciu drzew regresyjnych. Algorytmy indukcji drzew są często algorytmami heurystycznymi, co determinuje możliwość uzyskania różnych drzew dla tych samych danych. Stabilność otrzymanych w ten sposób wyników jest często dyskusyjna, stąd potrzeba stworzenia modeli, które umożliwiłyby uzyskanie rozwiązań powtarzalnych. Pozwoli to również wyodrębnić cechy, które mają rzeczywisty wpływ na przeżycie.

Wybór metody oceny jakości otrzymanego modelu jest zdeterminowany w dużym stopniu sposobem prezentacji wyników: dokładny czas porażki, funkcja przeżycia czy też hazardu. Metodami bezpośrednimi porównuje się wartości empiryczne z teoretycznymi, uzyskanymi jako wynik działania określonego narzędzia prognostycznego. Wśród tych metod możemy wyróżnić współczynniki korelacji rangowej Kendall'a i Somer'a [15], jak też współczynnik zaproponowany przez Schemper'a i Hendersona [19]. W przypadku pośrednim porównuje się nie wartości występujące w zbiorze, ale pewne funkcje, np. funkcję przeżycia, otrzymaną na bazie wartości ze zbioru uczącego, z funkcją przeżycia – rezultatem działania analizowanej metody. Przykładem tego typu rozwiązania jest współczynnik Briera zaproponowany przez Graf i in. [11], czy też miary dokładności predykcji rozwijane przez Schempera [18].

W pracy dokładność predykcji mierzona bezwzględnym błędem predykcji [18] jest wykorzystywana do oceny jakości prognostycznej poszczególnych cech. Dodatkowo weryfikacji podlega stworzone narzędzie – lasy losowe [12], budowane na bazie dipolowych drzew przeżycia [16]. Zastosowanie dipolowej funkcji kryterialnej pozwala wykorzystywać informacje niepełne, pochodzące z danych obciętych. Wynikiem działania lasów losowych jest sumaryczna funkcja przeżycia Kaplana-Meiera, budowana dla nowego pacjenta opisanego pewnym wektorem cech $\boldsymbol{x}$.

Praca składa się jest z sześciu rozdziałów. W rozdziale drugim przedstawiono ogólną charakterystykę danych analizy przeżyć, jak również opisano metodę estymacji funkcji przeżycia. Rozdział trzeci przedstawia schemat budowy lasów losowych wraz z algorytmem indukcji pojedynczego drzewa przeżycia. Opis metod

oceny jakości predykcji zawarty jest w rozdziale czwartym. Rozdział piąty zawiera wyniki eksperymentów, wykonanych na bazie dwóch zbiorów danych: *Primary Biliary Cirrhosis,* zawierający informacje o pacjentach z pierwotną marskością żółciową wątroby, oraz zbiór „*VA lung cancer*" z opisem pacjentów chorych na raka. Rozdział szósty to podsumowanie uzyskanych rezultatów.

2. Dane analizy przeżyć

Niech T oznacza nieujemną zmienną losową reprezentującą czas przeżycia pacjentów, natomiast C - zmienną losową reprezentującą czas obcięcia. Dane analizy przeżyć są reprezentowane przez zmienną $O=(\boldsymbol{X}, T, \Lambda)$, gdzie $\boldsymbol{X}=(X_1, X_2 \ ..., X_N)$ jest zbiorem N zmiennych z pewnej przestrzeni cech, natomiast $\Lambda=\mathrm{I}(T \leq C)$ jest wskaźnikiem przeżycia, gdzie I(*wyrażenie*) przyjmuje wartość 1, jeżeli wyrażenie jest prawdziwe i 0, gdy nie jest prawdziwe. Załóżmy, że mamy dany n elementowy zbiór uczący $L=[\boldsymbol{x}_i, t_i, \delta_i]$, $i=1,2,...,n$, gdzie $\boldsymbol{x}_i$ jest N-wymiarowym wektorem cech, t_i jest czasem przeżycia, natomiast δ_i – oznacza wskaźnik przeżycia, który w przypadku danych obciętych przyjmuje wartość równą 0, a dla danych pełnych jego wartość wynosi 1.

Rozkład zmiennej losowej T może być opisany m. in. przy użyciu funkcji przeżycia, określonej dla danego czasu t jako prawdopodobieństwo tego, że porażka nie miała miejsca wcześniej: $S(t)=P(T>t)$. Najczęściej wykorzystywanym estymatorem funkcji przeżycia jest estymator Kaplana-Meiera [14], który oznaczamy jako $\hat{S}(t)$:

$$\hat{S}(t) = \prod_{j | t_{(j)} \leq t} \left(\frac{m_j - d_j}{m_j} \right) \tag{1}$$

gdzie $t_{(1)}<t_{(2)}<...<t_{(D)}$ są uporządkowanymi rosnąco czasami porażki pacjentów ze zbioru uczącego L, d_j jest liczbą porażek, które miały miejsce w czasie $t_{(j)}$, m_j jest liczbą obserwacji, dla których czas porażki jest nie mniejszy niż $t_{(j)}$. Funkcję przeżycia (warunkową) wyliczoną dla nowego pacjenta opisanego wektorem cech $\boldsymbol{x}$ będziemy oznaczać symbolem $\hat{S}(t \mid x)$.

3. Modele złożone w analizie przeżyć

Wykorzystywane w pracy lasy losowe są zbiorem drzew regresyjnych indukowanych na podstawie danych przeżycia. W odróżnieniu od klasycznych drzew regre-

syjnych [15] poszczególne liście reprezentują krzywe przeżycia Kaplana-Meiera wyznaczone na podstawie obserwacji ze zbioru uczącego, które dany liść osiągnęły.

3.1. Indukcja dipolowego drzewa przeżycia

Większość algorytmów indukcji drzew rozpoczyna swoje działanie od korzenia drzewa. W każdym węźle wyliczany jest test optymalizujący zadane kryterium. Test ten w drzewach wielowymiarowych przyjmuje postać hiperpłaszczyzny $H(\boldsymbol{w}, \theta) = \{\boldsymbol{x}: <\boldsymbol{w}, \boldsymbol{x}>=\theta\}$. Odpowiednio dobrana funkcja decyzyjna dzieli zbiór uczący na dwa rozłączne podzbiory (drzewo binarne). W przypadku danych analizy przeżyć powinny powstać podzbiory różniące się długością czasu przeżycia. Proces podziału jest powtarzany dla każdego nowo utworzonego węzła-potomka do momentu, aż zostanie spełnione kryterium stopu i węzeł zostanie uznany za liść, czyli powstałe podzbiory są jednorodne z punktu widzenia czasu przeżycia.

Celem algorytmu indukcji drzewa jest wyznaczenie odpowiedniej liczby węzłów wewnętrznych drzewa, jak też położenia hiperpłaszczyzn $H(\boldsymbol{w}, \theta)$ w każdym węźle [4]. Proponowana w pracy metoda podziału (budowy testów w węzłach drzewa regresyjnego) bazuje na wykorzystaniu dipolowej funkcji kryterialnej [3], której budowa opiera się na pojęciu dipola. Dipolem nazywany parę wektorów cech $\{\boldsymbol{x}_i, \boldsymbol{x}_j\}$. Wyróżniamy dwa rodzaje dipoli - mieszane i czyste. Dipol mieszany tworzymy pomiędzy parą wektorów cech, które powinny zostać rozdzielone, dipol czysty pomiędzy wektorami cech jednorodnymi w punktu widzenia analizowanego kryterium. W przypadku analizy przeżyć dipole czyste tworzymy pomiędzy parami wektorów cech, dla których różnica czasu przeżycia jest odpowiednio mała, dipole mieszane pomiędzy tymi parami, dla których różnica czasu przeżycia jest odpowiednio duża. Uwzględniając zróżnicowanie obserwacji (pełne, obcięte) można sformułować następujące reguły konstrukcji dipoli [17]:

Para wektorów cech $\{\boldsymbol{x}_i, \boldsymbol{x}_j\}$ tworzy dipol czysty, gdy

- $\delta_i = \delta_j = 1 \wedge | t_i - t_j | < \eta$

Para wektorów cech $\{\boldsymbol{x}_i, \boldsymbol{x}_j\}$ tworzy dipol mieszany, gdy

- $\delta_i = \delta_j = 1 \wedge | t_i - t_j | > \zeta$
- $\delta_i = 0, \delta_j = 1 \wedge t_i - t_j > \zeta$ lub $\delta_i = 1, \delta_j = 0 \wedge t_j - t_i > \zeta$

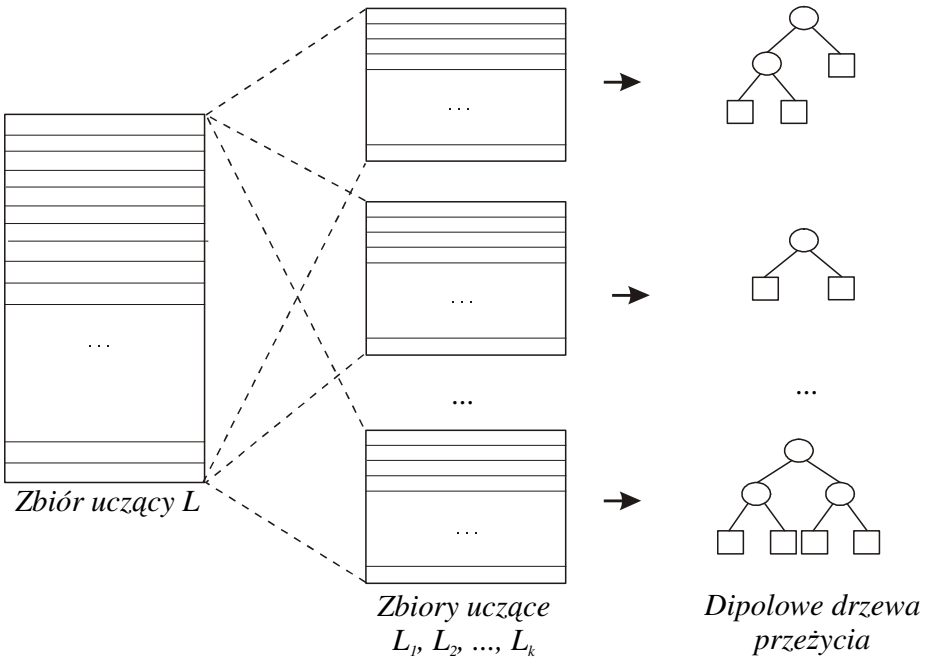
Parametry η oraz ζ przyjmują wartości równe kwantylom wartości bezwzględnych różnic czasu porażki obserwacji pełnych. Parametr η ustalany jest jako kwantyl rzędu 0.1-0.3, ζ jako kwantyl rzędu 0.6-0.9.

Z każdym dipolem związana jest odcinkowo-liniowa funkcja kary. Z dipolami mieszanymi wiążemy funkcje, których minimalizacja pozwoli znaleźć hiperpłaszczyznę $H(\boldsymbol{w}, \theta)$, przecinającej te dipole. Z dipolami czystymi natomiast wiążemy takie funkcje, aby hiperpłaszczyzna $H(\boldsymbol{w}, \theta)$ ich nie przecięła. Suma funkcji kary

po wszystkich dipolach tworzy dipolową funkcję kryterialną. W wyniku jej minimalizacji, przeprowadzanej przy użyciu algorytmów wymiany rozwiązań bazowych [2], otrzymujemy wartości współczynników $\boldsymbol{w}$ i θ hiperpłaszczyzny $H(\boldsymbol{w}, \theta)$ w kolejnych węzłach drzewa.

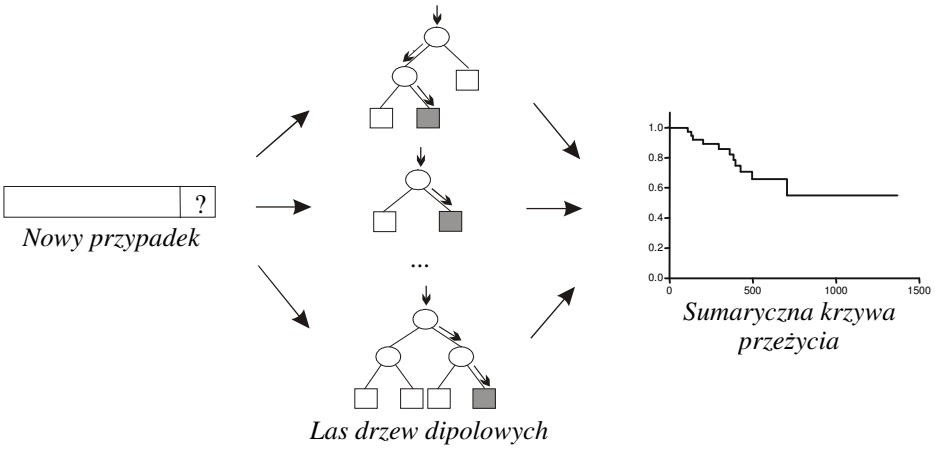
3.2. Algorytm budowy lasów losowych

Metoda lasów losowych [12] pozwala na estymację warunkowej funkcji przeżycia $\hat{S}(t \mid x_{n'})$. Jest ona budowana na bazie k zbiorów uczących ($L_1, L_2, ..., L_k$) losowanych ze zwracaniem ze zbioru uczącego L. Dla każdego zbioru uczącego L_i (i=1,2,.., k) wyznaczany jest zbiór obserwacji $L_i(\boldsymbol{x}_{n'})$ bliskich wektorowi $\boldsymbol{x}_{n'}$.



Rys. **1**. Konstrukcja lasu losowego

Dipolowe drzewo przeżycia jest indukowane dla każdego zbioru L_i, i=1,2,.., k (rys. 1). Wektor cech $\boldsymbol{x}_j$ jest włączany do zbioru $L_i(\boldsymbol{x}_{n'})$, jeżeli należy do tego samego liścia co wektor $\boldsymbol{x}_n$. Mając dane k zbiorów $L_i(\boldsymbol{x}_{n'})$, tworzy się rodzinę zbiorów $L_A(\boldsymbol{x}_{n'})$: $L_A(\boldsymbol{x}_{n'})=[L_1(\boldsymbol{x}_{n'}); L_2(\boldsymbol{x}_{n'}); ... ; L_k(\boldsymbol{x}_{n'})]$. Sumaryczna, warunkowa funkcja przeżycia Kaplana-Meiera, wyznaczana na bazie otrzymanego zbioru $L_A(\boldsymbol{x}_{n'})$ będzie określana mianem $\hat{S}_A(t \mid x_{n'})$.



Rys. **2**. Wykorzystanie lasu losowego

Algorytm budowy lasów losowych możemy zatem przedstawić w następujących punktach:

- Wylosowanie ze zwracaniem k n-elementowych zbiorów (L_1, L_2, ..., L_k) ze zbioru uczącego L
- Indukcja dipolowych drzew przeżycia $T(L_i)$ na bazie kolejnych zbiorów L_i, i=1,2,.., k
- Tworzenie rodziny zbiorów danych $L_A(\boldsymbol{x}_{n'})=[L_1(\boldsymbol{x}_{n'}); L_2(\boldsymbol{x}_{n'}); \ldots ; L_k(\boldsymbol{x}_{n'})]$ (dla nowego pacjenta, opisanego wektorem cech $\boldsymbol{x}_{n'}$).
- Estymacja sumarycznej funkcji przeżycia Kaplana-Meiera $\hat{S}_A(t \mid x_{n'})$ dla nowej obserwacji $\boldsymbol{x}_{n'}$ (rys. 2).

4. Ocena jakości predykcji

Ocena jakości prognostycznej danego narzędzia, a przez to poszczególnych cech tworzących to narzędzie, wykonana jest przy użyciu miary zaproponowanej przez Schempera [18]. Bezwzględny błąd predykcji jest wyliczany na bazie wartości funkcji przeżycia Kaplana-Meiera $\hat{S}(t)$, wyliczanej przy założeniu jednorodności populacji (bez uwzględniania wartości zmiennych) i warunkowej funkcji przeżycia otrzymanej w wyniku działania analizowanego narzędzia $\hat{S}(t \mid x)$.

Estymator bezwzględnego błędu predykcji, wyliczany dla dowolnego czasu $t_{(j)}$, jest zdefiniowany jako:

$$\tilde{M}(t_{(j)}) = 2\hat{S}(t_{(j)})(1-\hat{S}(t_{(j)})) \tag{2}$$

oraz

$$\tilde{M}(t_{(j)} \mid x) = 2n^{-1}\sum_i \hat{S}(t_{(j)} \mid x_i)(1-\hat{S}(t_{(j)} \mid x_i)) \tag{3}$$

Ponieważ z reguły zainteresowani jesteśmy przeżyciem przez określony czas $(0, \tau)$, uogólnione estymatory bezwzględnego błędu predykcji wylicza się jako ważoną średnią estymatorów (2) i (3) po wszystkich czasach porażki ze zbioru uczącego. Wyliczanie wartości wag pozwala na uwzględnienie występowania obserwacji obciętych w zbiorze. Otrzymujemy następujące estymatory:

$$\tilde{D}_s = w^{-1}\sum_j \hat{G}(t_{(j)})^{-1} d_j \tilde{M}(t_{(j)}) \tag{4}$$

$$\tilde{D}_{s,x} = w^{-1}\sum_j \hat{G}(t_{(j)})^{-1} d_j \tilde{M}(t_{(j)} \mid x) \tag{5}$$

gdzie $w = \sum_j \hat{G}(t_{(j)})^{-1} d_j$, d_j jest liczbą porażek, które miały miejsce w czasie $t_{(j)}$ natomiast $\hat{G}$ oznacza estymator Kaplana-Meiera, przy czym jako porażkę traktuje się obcięcie obserwacji.
Wariancja wyjaśniona modelem jest zdefiniowana jako:

$$\tilde{V}_s = \frac{(\tilde{D}_s - \tilde{D}_{s,x})}{\tilde{D}_s} \tag{6}$$

5. Wyniki eksperymentów

Analizę wykonano na bazie dwóch zbiorów danych. Pierwszy zbiór, *Primary Biliary Cirrhosis – PBC,* zawiera dane z kliniki w Mayo [10]. Obserwacji poddani byli pacjenci z pierwotną marskością żółciową wątroby. 312 pacjentów brało udział w randomizowanym badaniu klinicznym, mającym na celu ocenę leczenia z wykorzystaniem D-penicillaminy. Okres obserwacji trwał 10 lat: od 1974 do 1984 roku. Czas przeżycia liczony był od momentu rejestracji do zgonu, transplantacji wątroby lub zakończenia okresu obserwacji. Do analizy wzięto następujące cechy: wiek (w latach) - WIEK, występowanie obrzęku (tak, nie), stężenie albuminy we krwi [g/dl] - ALBUMINA, log(poziom bilirubiny we krwi [mg/dl]) -LOGBIL oraz log(czas protrombinowy [sek]) -LOGPRO. Zbiór zawiera 60% obserwacji obciętych.

Stosując model regresji wielokrotnej Cox'a uzyskano następujące zmienne znaczące: LOGBIL ($p<0.00001$), ALBUMINA ($p<0.0001$), WIEK ($p=0.0001$), LOGPRO ($p=0.001$) [18].

Tabela 1

Bezwzględny błąd predykcji i wariancja wyjaśniona modelem dla zbioru *PBC*

Model	Bezwzględny błąd predykcji (średnia ± odch. std.)	Wariancja wyjaśniona (średnia ± odch. std.)
Estymator KM	0,37	-
Model regresji Cox'a	0,23	0,40
Lasy losowe		
wszystkie zmienne	0,233 ± 0,005	0,37 ± 0,013
LOGBIL	0,244 ± 0,005	0,34 ± 0,014
ALBUMINA	0,289 ± 0,007	0,22 ± 0,02
WIEK	0,335 ± 0,013	0,095 ± 0,035
LOGPRO	0,3 ± 0,003	0,189 ± 0,009

W skład wykorzystywanych w analizie lasów losowych wchodzi 100 drzew dipolowych. Dla każdego zestawu danych analiza była powtórzona 20 razy. Uzyskane wartości średnie bezwzględnego błędu predykcji oraz wariancji wyjaśnionej modelem wraz z odchyleniem standardowym przedstawione są w tabeli 1. Możemy zauważyć, że bezwzględny błąd predykcji jest porównywalny z błędem uzyskanym przy modelu regresji Cox'a (0,23), tym samym procent wariancji wyjaśnionej jest podobny (40% i 37%, odpowiednio dla modelu Cox'a i lasów losowych). Analizowano również wpływ poszczególnych pojedynczych cech. Najlepsze własności predykcji dla pacjentów z pierwotną marskością żółciową wątroby ma poziom bilirubiny we krwi (logarytm). Wprowadzenie tej zmiennej redukuje błąd predykcji średnio o 12,6. Wariancja wyjaśniona modelem wynosi 34%. Zmiennymi mającymi mniejszy wpływ są wiek i czas protrombinowy, dla których bezwzględny błąd predykcji jest równy odpowiednio 0,335 i 0,3.

Drugi zbiór „*VA lung cancer*" (ang. *Veteran's Administration lung cancer*) [13] zawiera informacje o 137 pacjentach (9 obserwacji obciętych), u których wykryto raka płuc. Pacjenci, mężczyźni, zostali poddani terapii standardowej (69 przypadków) oraz chemoterapii (68). Dodatkowo są oni opisani następującymi zmiennymi: indeks KPS (stan pacjenta w trakcie randomizacji: 10-30 – w trakcie pobytu w szpitalu, 40-60 – okresowo hospitalizowany i okresowo pod opieką poradni ambulatoryjnej, 70-90 – pod opieką ambulatoryjną), czas trwania choroby w miesiącach, wiek, wcześniejsza terapia (tak, nie) oraz typ komórek rakowych

(0 – rak płaskonabłonkowy, 1 – rak drobnokomórkowy, 2 – gruczolakorak, 3 – rak wielkokomórkowy).

Tabela 2

Bezwzględny błąd predykcji i wariancja wyjaśniona modelem dla zbioru „*VA lung cancer"*

Model	Bezwzględny błąd predykcji (średnia ± odch. std.)	Wariancja wyjaśniona (średnia ± odch. std.)
Estymator KM	0,335	-
Lasy losowe		
wszystkie zmienne	0,214 ± 0,008	0,357 ± 0,03
WIEK	0,314 ± 0,012	0,061 ± 0,037
CZAS_CHOROBY	0,315 ± 0,008	0,061 ± 0,023
KPS	0,253 ± 0,008	0,245 ± 0,023
TYP_KOMÓREK	0,297 ± 0,006	0,115 ± 0,018

Wyniki uzyskane dla zbioru „*VA lung cancer*" zawiera tabela 2. Bezwzględny błąd predykcji dla estymatora funkcji przeżycia Kaplana-Meiera wynosi 0,335. W przypadku lasów losowych, brano pod uwagę model wykorzystujący wszystkie wymienione wyżej cechy oraz modele budowane na podstawie pojedynczych cech: wiek pacjentów (WIEK), indeks KPS (KPS), czas trwania choroby w miesiącach (CZAS_CHOROBY) oraz typ komórek rakowych (TYP_KOMÓREK). Można zaobserwować, że najlepszą jakość predykcji ma zmienna KPS (średni błąd wynosi 0,253, co daje nam 24,5% wariancji wyjaśnionej modelem), następnie typ komórek rakowych (0,297 i 11,5%, odpowiednio błąd i wariancja wyjaśniona modelem). Wpływ wieku i czasu choroby na przeżycie jest podobny. Dla tych zmiennych błąd predykcji wynosi odpowiednio 0,314 i 0,315, co wyjaśnia tylko 6,1% wariancji.

6. Podsumowanie

W pracy przedstawiono możliwość wykorzystania bezwzględnego błędu predykcji oraz wariancji wyjaśnionej modelem do oceny jakości prognostycznej cech. Miary dokładności predykcji budowane były przez porównanie, uzyskanych na bazie lasów losowych, sumarycznych estymatorów funkcji przeżycia Kaplana-Meiera z funkcją przeżycia, utworzoną przy założeniu jednorodności populacji. Dodatkowo jakość narzędzia prognostycznego – lasów losowych – została zweryfikowana przez porównanie z modelem Cox'a. Uzyskane wyniki potwierdzają przydatność lasów losowych do analizy danych przeżycia, a tym samym do oceny jakości prognostycznej poszczególnych zmiennych.

Literatura

[1] Biganzoli, E., Boracchi, P., Mariani, L., Marubini, E., *Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach*, Statistics in Medicine 17(10), 1998, str. 1169-1186

[2] Bobrowski, L., *Design of piecewise linear classifiers from formal neurons by some basis exchange technique*, Pattern Recognition 24(9), 1991, str. 863-870.

[3] Bobrowski, L., Krętowska, M., Krętowski, M., *Design of neural classifying networks by using dipolar criterions*, Proceedings of the 3rd Conference on Neural Networks and their Applications, Częstochowa, 1997, str. 689-694.

[4] Bobrowski, L., Krętowski, M., *Induction of multivariate decision trees by using dipolar criteria*, Zighed D.A., Komorowski J., Żytkow J. (Eds.): PKDD 2000, LNAI 1910, Springer-Verlag, 2000, str. 331-336

[5] Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J., *Classification and Regression Trees*, Wadsworth, 1984.

[6] Breiman, L., *How to use survival forest*. [http://stat-www.berkeley.edu/users/breiman]

[7] Ciampi, A., Negassa, A., Lou, Z., *Tree-structured prediction for censored survival data and the Cox model,* Journal of Clinical Epidemiology 48(5), 1995, str. 675-689

[8] Cox, D.R., *Regression models and life tables (with discussion),* Journal of the Royal Statistical Society B 34, 1972, str. 187-220

[9] Faraggi, D., Simon, R., *A neural network model for survival data*, Statistics in Medicine 14, 1995, str. 73-82

[10] Fleming, T. R., Harrington, D. P., *Counting Processes and Survival Analysis*, John Wiley & Sons, Inc., 1991

[11] Graf, E., Schmoor, C., Sauerbrei, W., Schumacher, M., *Assessment and comparison of prognostic classification schemes for survival data*, Statistics in Medicine 18, 1999, str. 2529-2545

[12] Hothorn, T., Lausen, B., Benner, A., Radespiel-Troger, M., *Bagging survival trees*, Statistics in medicine 23, 2004, str. 77-91

[13] Kalbfleisch, J.D., Prentice, R.L., *The statistical analysis of failure time data*, Wiley, New York, 1980.

[14] Kaplan, E.L., Meier, P., *Nonparametric estimation from incomplete observations*, Journal of the American Statistical Association 5, 1958, str. 457-481

[15] Korn, E. L., Simon, R., *Measures of explained variation for survival data*, Statistics in medicine 9, 1990, str. 487-503

[16] Krętowska, M.: *Random forest of dipolar trees for survival prediction*, Lecture Notes in Computer Science 4029, Lecture Notes in Artificial Intelligence, 2006, str. 909-918.
[17] Krętowska, M., *Dipolar regression trees in survival analysis,* Biocybernetics and biomedical engineering 24(3), 2004, str. 25-33
[18] Schemper, M., *Predictive accuracy and explained variation*, Statistics in medicine 22, 2003, str. 2299-2308
[19] Schemper, M., Henderson, R., *Predictive accuracy and explained variation in Cox regression*, Biometrics 56(1), 2000, str. 494-255

RANDOM FORESTS – EVALUATION OF PREDICTIVE ACCURACY

Abstract: In the paper, predictive accuracy measured as the absolute predictive error is used to evaluate the quality of covariates. The prognostic tool – random forests – is built to receive the aggregated survival function. The function is compared to Kaplan-Meier estimator of survival function with assumption that the population is homogenous. The induction of individual dipolar survival tree is based on minimization of a piece-wise linear function – dipolar criterion. The algorithm allows using the information from censored observations for which the exact survival time is unknown.

Keywords: random forest, survival analysis, predictive accuracy, explained variation

Artykuł zrealizowano w ramach pracy badawczej W/WI/4/05.