PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 86 — 2009
Wizualizacja wyników badań marketingowych – podejścia, metody i zastosowania

Eugeniusz Gatnar
Akademia Ekonomiczna w Katowicach

PODZIAŁ WIELOWYMIAROWEJ PRZESTRZENI ZMIENNYCH A MODELE HYBRYDOWE

1. Wstęp

Modele hybrydowe to modele wykorzystujące rekurencyjny podział wielowymiarowej przestrzeni zmiennych na podprzestrzenie (segmenty). W każdym z tych segmentów jest budowany model lokalny, np. liniowy, a następnie modele te są łączone w jeden model globalny.

Taki sposób budowy modelu odbiega od rozwiązania przyjętego w klasycznej analizie regresji, gdzie zwykle tworzony jest jeden model globalny, którym najczęściej jest uogólniony model liniowy (GLM).

Celem artykułu jest omówienie wyników zastosowania różnych typów modeli hybrydowych w analizie regresji oraz porównanie dokładności ich dopasowania do danych.

Do obliczeń został wykorzystany program komputerowy **R**, tj. środowisko służące do prowadzenia zaawansowanych analiz statystycznych, wraz z wieloma pakietami oraz ogólnodostępne zbiory danych. Dla zwiększenia czytelności wywodów przedstawiono także wykorzystane procedury w języku **R**.

2. Metoda rekurencyjnego podziału przestrzeni zmiennych

Metoda rekurencyjnego podziału dzieli sekwencyjnie wielowymiarową przestrzeń zmiennych $\mathbf{X}^L$ na podprzestrzenie R_k (segmenty, regiony) aż do chwili, gdy zmienna zależna Y osiągnie w każdej z nich minimalny poziom zróżnicowania (mierzony za pomocą odpowiedniej funkcji straty). Metoda ta była stosowana w statystyce już przez Morgana i Sonquista [1963]. Jej wykorzystanie w analizie dyskryminacyjnej i regresji przedstawili Breiman i in. [1984]. W języku polskim wyczerpującą monografią poświęconą zagadnieniom budowy modeli w postaci drzew klasyfikacyjnych i regresyjnych jest praca Gatnara [2001].

Algorytm metody rekurencyjnego podziału składa się z kilku kroków:

1. Sprawdź, czy wszystkie obiekty w przestrzeni zmiennych $\mathbf{X}^L$ są jednorodne ze względu na zmienną Y. Jeżeli tak, to zakończ pracę.

2. W przeciwnym wypadku sprawdź wszystkie możliwe podziały przestrzeni zmiennych $\mathbf{X}^L$, tj. według każdej zmiennej $X_1, ..., X_L$ oraz każdego sposobu ich dyskretyzacji, na rozłączne podprzestrzenie (segmenty) R_k.

3. Dokonaj oceny każdego z tych podziałów i wybierz najlepszy z nich, tj. dołącz zmienną X_l do modelu.

4. Podziel przestrzeń zmiennych $\mathbf{X}^L$ na K podprzestrzeni zgodnie z wybranym najlepszym podziałem, a następnie wykonaj kroki 1-4 dla każdej z podprzestrzeni.

Przebieg procedury rekurencyjnego podziału najlepiej reprezentuje drzewo, tj. graf spójny i bez cykli. Stąd nazwa metody – drzewa klasyfikacyjne (*classification trees*) oraz drzewa regresyjne (*regression trees*). Przykład modelu w postaci drzewa klasyfikacyjnego znajduje się na rys. 1, a drzewa regresyjnego – na rys. 2.

W ramach omawianej metody model jest tworzony nie globalnie, lecz przez złożenie modeli lokalnych, budowanych w każdym z K rozłącznych segmentów, na jakie dzielona jest wielowymiarowa przestrzeń zmiennych $\mathbf{X}^L$:

$$Y = \alpha_0 + \sum_{k=1}^{K} \alpha_k g_k(\mathbf{X}) . \tag{1}$$

W szczególnym przypadku modele lokalne mogą mieć najprostszą postać, tj. stałej, jak to ma miejsce dla drzew klasyfikacyjnych i regresyjnych:

$$g_k(\mathbf{x}_i) = I(\mathbf{x}_i \in R_k), \tag{2}$$

gdzie R_k ($k = 1, ..., K$) to podprzestrzenie (segmenty) przestrzeni $\mathbf{X}^L$, α_k – parametry modelu, I zaś jest funkcją wskaźnikową.

Każdy z obszarów R_k jest definiowany poprzez jego granice w przestrzeni $\mathbf{X}^L$, które dla zmiennych metrycznych $X_1, ..., X_L$ można przedstawić jako:

$$I(\mathbf{x}_i \in R_k) = \prod_{l=1}^{L} I\left(v_{kl}^{(d)} \le x_{il} \le v_{kl}^{(g)}\right), \tag{3}$$

gdzie wartości $v_{kl}^{(d)}$ oraz $v_{kl}^{(g)}$ oznaczają odpowiednio jego górną i dolną granicę w l-tym wymiarze przestrzeni zmiennych.

W przypadku, gdy zmienne $X_1, ..., X_L$ mają charakter niemetryczny, podprzestrzeń R_k można zdefiniować jako:

$$I(\mathbf{x}_i \in R_k) = \prod_{l=1}^{L} I\left(x_{il} \in B_{kl}\right), \tag{4}$$

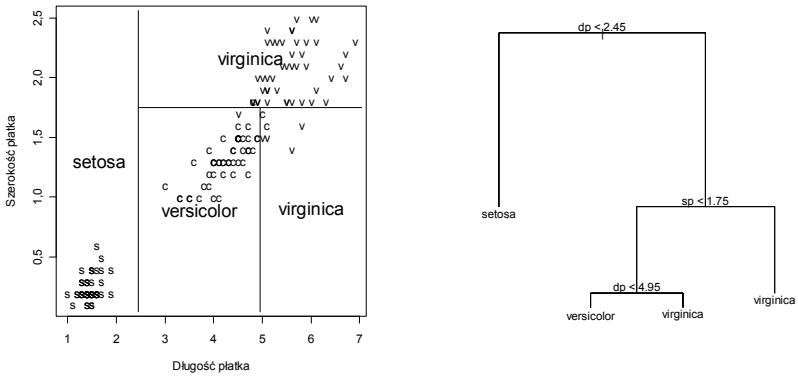
gdzie B_{kl} to podzbiór zbioru kategorii zmiennej X_l, tj. $B_{kl} \subseteq V_l$.

3. Hybrydowe modele dyskryminacyjne

Jeżeli zmienna zależna *Y* w modelu (1) jest zmienną nominalną, to model ten nazywany jest dyskryminacyjnym i reprezentuje go drzewo klasyfikacyjne. Wtedy parametry α_k modelu (2) są wyznaczane jako:

$$\alpha_k = \arg\max_j p(C_j \mid \mathbf{x}_i \in R_k) . \tag{5}$$

Na rysunku 1 pokazano model w postaci drzewa klasyfikacyjnego dla zbioru IRYS, w którym znajduje się 150 obserwacji kwiatów irysa, należących do 3 gatunków: *Setosa* (s), *Virginica* (v) oraz *Versicolor* (c), wykorzystujący dwie zmienne objaśniające: *długość płatka (dp)* i *szerokość płatka (sp)* oraz odpowiadający mu podział przestrzeni dwuwymiarowej na 4 segmenty.



Rys. 1. Rekurencyjny podział i jego interpretacja w postaci drzewa klasyfikacyjnego

Źródło: opracowanie własne.

Procedura w programie **R**, która tworzy rys. 1, składa się z 4 poleceń:

```
drzewo.iris <- tree(klasa~., data=iris)
partition.tree(drzewo.iris)
plot(drzewo.iris)
text(drzewo.iris).
```

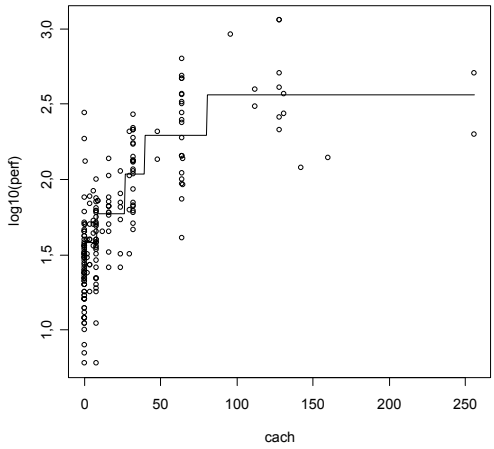
4. Hybrydowe modele regresji

W przypadku, gdy zmienna zależna *Y* jest mierzona na jednej ze skal mocnych, model (1) jest modelem regresji, którego graficzną postacią jest drzewo regresyjne. Jego parametry są wyznaczane za pomocą formuły:

$$\alpha_k = \frac{1}{N(k)} \sum_{\mathbf{x}_i \in R_k} y_i \,, \tag{6}$$

gdzie: $N(k)$ to liczba obserwacji należących do segmentu R_k. Innymi słowy, parametr ten jest wartością przeciętną Y dla obserwacji znajdujących się w podprzestrzeni R_k.

Zbudowano także hybrydowy model regresji dla zbioru CPUS zawierającego dane dotyczące parametrów pracy 209 jednostek centralnych komputerów stacjonarnych. Zmienną zależną jest wskaźnik szybkości pracy jednostki centralnej w porównaniu z komputerem IBM 370/158-3 (*perf*), a zmienną objaśniającą była zmienna *cach* (wielkość tzw. pamięci podręcznej w kilobajtach). Jego graficzna postać znajduje się na rys. 2.



Rys. 2. Hybrydowy model regresji dla jednej zmiennej objaśniającej

Źródło: opracowanie własne.

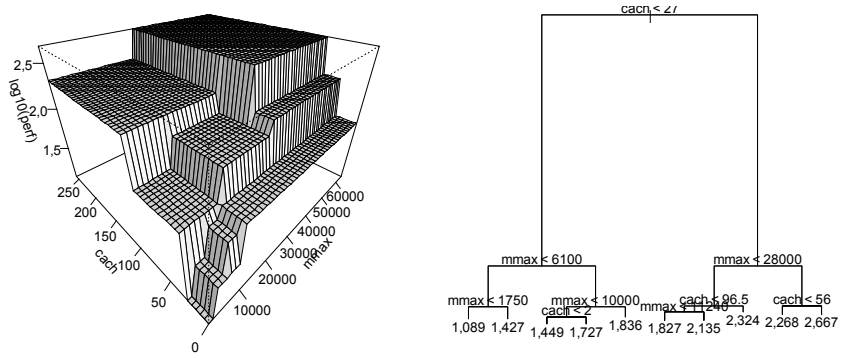
Procedura w programie **R**, która tworzy model na rys. 2, składa się z pięciu poleceń:

```
model<-tree(log10(perf)~cach)
w<-seq(min(cach),max(cach),0.5)
y.pred<-predict(model,list(cach=w))
plot(cach,log10(perf))
lines(w,y.pred).
```

Następnie zbudowano model regresji dla zbioru CPUS, w którym zmienną zależną była ta sama zmienna *perf*, a zmiennymi objaśniającymi były: *mmax* (największa pojemność pamięci operacyjnej w kilobajtach) i *cach*. Ma on postać drzewa, które znajduje się w prawej części rys. 3.

Procedura w programie **R**, która tworzy model na rys. 3, składa się z 4 poleceń:

```
drzewo.cpus <- tree(log10(perf)~ mmax+cach, data=cpus)
wireframe(perf ~ mmax*cach, data=cpusT, scales=list(arrows=FALSE),
 drape=TRUE, colorkey=FALSE, screen=list(z=-40, y=0, x=-60))
plot(drzewo.cpus)
text(drzewo.cpus).
```



Rys. 3. Hybrydowy model regresji dla dwóch zmiennych objaśniających

Źródło: opracowanie własne.

5. Ocena jakości podziału

Ocena jakości podziału, tj. homogeniczności poszczególnych podprzestrzeni, na które dzielona jest przestrzeń zmiennych, dokonywana jest za pomocą pewnej funkcji oceniającej zróżnicowanie obserwacji znajdujących się w R_k. W przypadku gdy model (1) jest modelem dyskryminacyjnym, stosowana jest funkcja entropii:

$$Q(R_k) = -\sum_{j=1}^{J} p(C_j \mid R_k) \log_2 p(C_j \mid R_k), \tag{7}$$

gdy zaś (1) jest modelem regresyjnym, wykorzystywana jest funkcja kwadratowa:

$$Q(R_k) = \frac{1}{N(k)} \sum_{\mathbf{x}_i \in R_k} (y_i - \alpha_k)^2 . \tag{8}$$

6. Modele lokalne klasy GLM

Model lokalny może być jednak także modelem klasy GLM, a szczególnie prostym modelem liniowym:

$$g_k(\mathbf{X}) = \beta_0 + \sum_{l=1}^{L} \beta_l X_l . \tag{9}$$

Tego typu modele hybrydowe łączące podejście nieparametryczne (drzewo) z parametrycznym nazywane są drzewami funkcyjnymi [Gama 2004]. Przykładami takich algorytmów są: QUEST [Loh, Shih 1997], GUIDE [Loh 2002], CRUISE [Kim, Loh 2001] oraz LOTUS [Chan, Loh 2004]].

Na szczególną uwagę zasługuje propozycja Zeileisa, Hothorna i Hornika [2008], którzy zastosowali inne kryterium doboru zmiennych do podziału przestrzeni. Jest nim stabilność parametrów modeli lokalnych (liniowych), które są estymowane metodą najmniejszych kwadratów lub największej wiarygodności.

W odróżnieniu od klasycznego rozwiązania w tym przypadku najpierw wyznaczana jest zmienna, która definiuje podział, a następnie dokonywana jest jej dyskretyzacja. Algorytm rekurencyjnego podziału ma wtedy postać:

1. Zbuduj model lokalny (liniowy) w segmencie R_k.

2. Oceń stabilność parametrów tego modelu. Jeżeli niestabilność występuje, wybierz zmienną X_l, przy której stoi najbardziej niestabilny parametr. W przeciwnym wypadku zakończ pracę.

3. Dokonaj dyskretyzacji zmiennej X_l.

4. Dokonaj podziału przestrzeni na segmenty według zmiennej X_l i przejdź do kroku 1.

Do testowania stabilności parametrów modelu lokalnego w kroku 2 stosowany jest test wahań (fluktuacji) zaproponowany przez Zeileisa, Hothorna i Hornika [2008]. Wykorzystuje on statystykę:

$$W_l(t) = \frac{1}{\sqrt{n \cdot \hat{\mathbf{D}}}} \cdot \sum_{i=1}^{n \cdot t} \varphi(X_{il}), \tag{10}$$

która jest zbieżna z procesem Browna w przypadku prawdziwości hipotezy zerowej o stabilności parametrów modelu. We wzorze (10) n oznacza liczbę obserwacji, $\hat{\mathbf{D}}$ jest estymatorem macierzy wariancji i kowariancji, φ zaś jest funkcją oceny obserwacji, dającą antyrangę (*antirank*) obserwacji X_{il}.

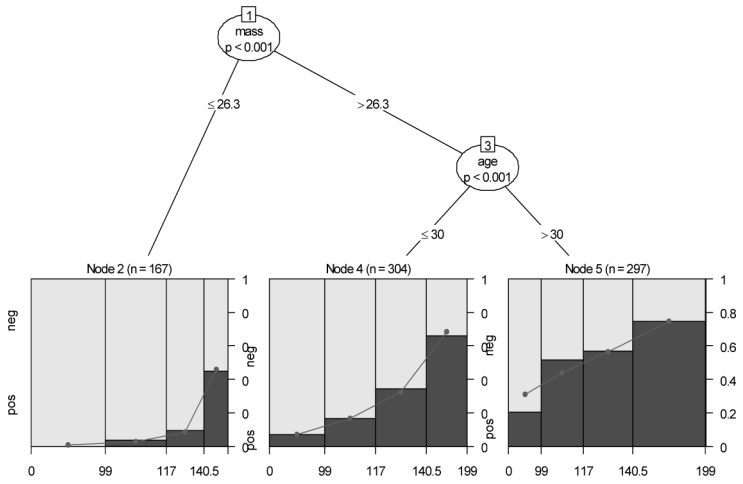
Oczywiście można także zastosować inne, podobne testy stabilności parametrów, znane w ekonometrii: CUSUM [Ploberger, Krämer 1992], MOSUM [Chu, Hornik, Kuan 1995], test Nybloma [1989] itp.

7. Przykłady

Zbudowano model hybrydowy (model regresji logistycznej) dla zbioru DIABETES, który zawiera dane o 768 osobach, które poddano testom na cukrzycę. Wynik tego testu jest zmienną zero-jedynkową (*diabetes*) i oznacza klasę (wartości `pos` i `neg`). Zmiennymi objaśniającymi jest 7 zmiennych charakteryzujących wiek (*age*), ciśnienie krwi (*press*), indeks masy ciała (*mass*) itd.

Zmienną, która definiuje modele lokalne w segmentach, jest *glucose*, będąca wynikiem pomiaru poziomu glukozy we krwi.

Postać graficzna tego modelu hybrydowego znajduje się na rys. 4.



Rys. 4. Hybrydowy model regresji

Źródło: opracowanie własne.

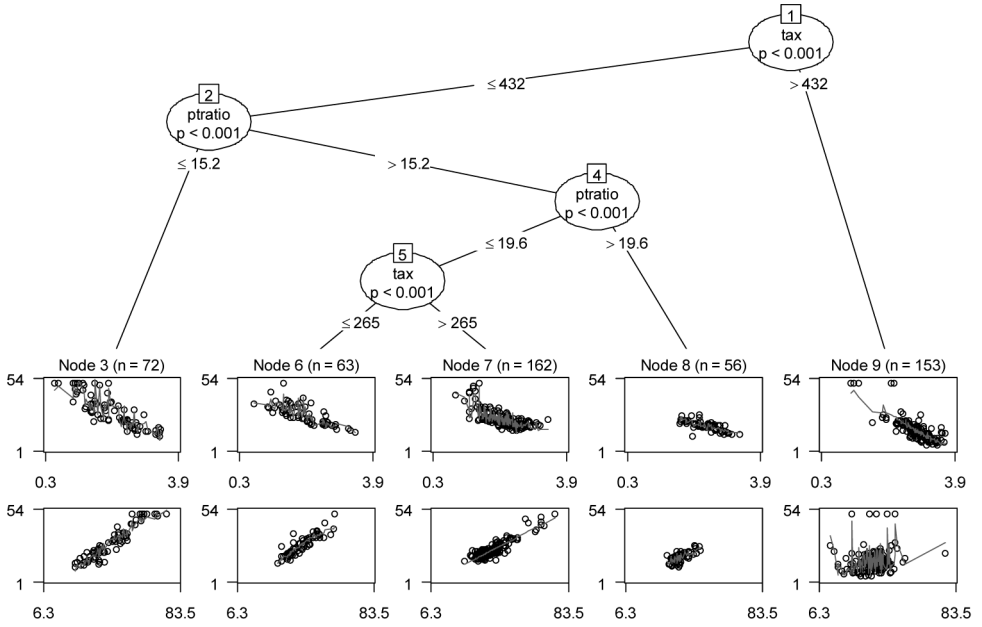
Model hybrydowy, przedstawiony na rys. 4, został przygotowany za pomocą polecenia `mob` z pakietu `party`:

```
model <- mob(diabetes ~ glucose | pregnant + pressure + triceps +
insulin + mass + pedigree + age, data = Diabetes, model = glinearModel,
family = binomial())
```

Jak widać na rys. 4, do podziału przestrzeni wykorzystano indeks masy ciała oraz wiek, co w rezultacie daje 3 podprzestrzenie. Dla każdej z nich dokonano wizualizacji modelu regresji logistycznej za pomocą spinogramu. Pokazuje on, że dla osób o małej wartości indeksu masy ciała ryzyko pojawienia się cukrzycy jest małe. Wzrasta ono jednak wraz z wiekiem i największe ryzyko występuje u osób starszych o wysokiej masie ciała.

Zbudowano także hybrydowy model regresji dla zbioru BOSTON, który zawiera informacje o wpływie różnych czynników na ceny 506 nieruchomości na przedmieściach Bostonu. Zmienną zależną jest mediana wartości domu w tys. dolarów (*medv*), zmiennymi objaśniającymi zaś są: koncentracja tlenku azotu (*nox*), wskaźnik przestępstw (*crim*), wielkość podatku od nieruchomości (*tax*), procent ludności afroamerykańskiej (*b*), procent budynków zbudowanych przed 1940 r. (*age*), dostęp do autostrady (*rad*), wskaźnik uczniów przypadających na jednego nauczyciela (*ptratio*), ważona odległość od pięciu centrów zatrudnienia w Bostonie (*dis*), dostęp do rzeki Charles River (*chas*), odsetek terenów nieprzeznaczonych do celów handlowych (*indus*), odsetek terenów zamieszkałych (*zn*). Modele wewnętrzne zbudowane

są dla zmiennych: procent populacji o niskim statusie społecznym (*lstat*), średnia liczba pokoi w domu (*rm*).



Rys. 5. Hybrydowy model regresji wielorakiej

Źródło: opracowanie własne.

W programie **R** metoda budowy modeli hybrydowych jest zaimplementowana w pakiecie `party`. Przykład wykorzystania znajduje się poniżej, czego efektem jest rys. 5.

```
drzewo <- mob(medv ~ lstat + rm | zn + indus + chas + nox + age +
dis + rad
+ tax + crim + b + ptratio, control = mob_control(minsplit = 40),
data = Boston, model = linearModel).
```

W rezultacie powstaje drzewo regresyjne w postaci:

```
1) tax <= 432; criterion = 1, statistic = 115.364
 2) ptratio <= 15.2; criterion = 1, statistic = 50.482
  3)* weights = 72
Terminal node model
Linear model with coefficients:
(Intercept)    lstat       rm
   9.2349    -4.9391   0.6859
 2) ptratio > 15.2
  4) ptratio <= 19.6; criterion = 1, statistic = 55.266
   5) tax <= 265; criterion = 1, statistic = 43.178
```

```
   6)* weights = 63
Terminal node model
Linear model with coefficients:
(Intercept)      lstat       rm
    3.9637     -2.7663   0.6881
     5) tax > 265
      7)* weights = 162
Terminal node model
Linear model with coefficients:
(Intercept)      lstat       rm
 -1.7984       -0.2677   0.6539
  4) ptratio > 19.6
   8)* weights = 56
Terminal node model
Linear model with coefficients:
(Intercept)      lstat       rm
  17.5865      -4.6190   0.3387
1) tax > 432
 9)* weights = 153
Terminal node model
Linear model with coefficients:
(Intercept)      lstat       rm
    68.2971   -16.3540  -0.1478.
```

Uzyskany model należy interpretować w ten sposób, że przestrzeń zmiennych została podzielona na 5 segmentów za pomocą dwóch zmiennych: wielkość podatku od nieruchomości (*tax*) i wskaźnik uczniów przypadających na jednego nauczyciela (*ptratio*). W każdym z nich zbudowano model liniowy, którego parametry znajdują się na wydruku zamieszczonym powyżej, np. w segmencie nr 3 model lokalny ma postać:

$$g_3(\mathbf{X}) = 9{,}2349 - 4{,}9391 \times lstat + 0{,}6859 \times rm. \tag{11}$$

Ogólnie mówiąc, wzrost liczby mieszkańców o niskim statusie społecznym w sąsiedztwie (*lstat*) powoduje spadek ceny nieruchomości, a wzrost liczby pokoi (*rm*) powoduje wzrost ceny nieruchomości. Wyjątkiem jest jedynie segment nr 9 (największe nieruchomości), w którym parametr przy zmiennej *rm* także jest ujemny (co prawda jest on jedynie nieznacznie większy od zera).

Dla zbioru BOSTON dokonano także porównania oceny dopasowania różnych modeli regresji do danych, wykorzystując współczynnik determinacji (tab. 1).

Tabela 1. Współczynniki determinacji modeli regresji dla zbioru BOSTON

Metoda	Współczynnik determinacji
rpart	80,6%
MOB	90,2%
Randomforest	92,9%

Źródło: opracowanie własne.

Uzyskane wyniki, przedstawione w tab. 1, wskazują co prawda na największą dokładność modeli zagregowanych, zbudowanych za pomocą metody RandomForest [Gatnar 2008], lecz modele hybrydowe zaproponowane przez Zeileisa, Hothorna i Hornika [2008] są dużo bardziej dokładne niż modele w postaci drzew regresyjnych (procedura `rpart`).

8. Konkluzje

Metoda budowy hybrydowych modeli regresyjnych ma następujące własności:

- wykorzystuje modele nieparametryczne i parametryczne,
- oddziela poszukiwanie zmiennej definiującej podział od procesu jej dyskretyzacji,
- stosuje testy stabilności parametrów modeli liniowych,
- daje lepsze dopasowanie modelu globalnego do danych.

Niezwykle ważna jest także możliwość ujęcia w omawianych modelach hybrydowych zmiennych objaśniających mierzonych na skalach zarówno mocnych, jak i na słabych, bez konieczności dokonywania ich transformacji.

Literatura

Breiman L., Friedman J., Olshen R., Stone C., *Classification and Regression Trees*, CRC Press, London 1984.

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

Chu C.-S., Hornik K., Kuan C.-M., *MOSUM Tests for Parameter Constancy*, „Biometrika" 1995, 82 , s. 603-617.

Gama J., *Functional Trees*, „Machine Learning" 2004, 55, s. 219-250.

Gatnar E., *Nieparametryczna metoda dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa 2001.

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

Kim H., Loh W.-Y., *Classification Trees with Unbiased Multiway Splits*, „Journal of the American Statistical Association" 2001, 96, 589-604.

Loh W.-Y., Shih Y.-S., *Split Selection Methods for Classification Trees*, „Statistica Sinica" 1997, 7, s. 815-840.

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

Morgan J.N., Sonquist J.A., *Problems in the Analysis of Survey Data: a Proposal*, „Journal of the American Statistical Association" 1963, 58, s. 417-434.

Nyblom J., *Testing for the Constancy of Parameters Over Time*, „Journal of the American" Statistical Association" 1989, 84, s. 223-230.

Ploberger W., Krämer W., *The CUSUM Test with OLS Residuals*, „Econometrica" 1992, 60, s. 271-285.

Therneau T.M., Atkinson E.J., *An Introduction to Recursive Partitioning Using the RPART Routines*, Mayo Foundation, Rochester 1997.

Zeileis A., *A Unified Approach to Structural Change Tests Based on ML Scores, F Statistics, and OLS Residuals*, „Econometric Reviews" 2005, 24, s. 445-466.
Zeileis A., Hothorn T., Hornik K., *Model-based Recursive Partitioning*, „Journal of Computational and Graphical Statistics" 2008, 17(2), s. 492-514.

MULTIDIMENSIONAL FEATURE SPACE PARTITION AND HYBRID MODELS

Summary

Hybrid models are based on the recursive partitioning of multidimensional feature space into subspaces (regions). Then, in each segment a local model is built (e.g. linear model) and finally all the local models are combined into the global model.

The aim of the paper is to discuss the results of application of different hybrid models in regression. The author compares the goodness of fit of the models to the training data. The paper also presents the portions of code of the **R** statistical package used in the author's experiments.