PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 47 2009
TAKSONOMIA 16
Klasyfikacja i analiza danych – teoria i zastosowania

Joanna Trzęsiok

Akademia Ekonomiczna w Katowicach

OCENA WPŁYWU WYMIARU PRZESTRZENI ZMIENNYCH NA JAKOŚĆ PREDYKCJI WYBRANYCH NIEPARAMETRYCZNYCH MODELI REGRESJI

1. Wstęp

Wyniki przeprowadzonych badań empirycznych pokazują, że modele zbudowane na podstawie nieparametrycznych metod regresji charakteryzują się lepszymi własnościami statystycznymi niż modele uzyskane za pomocą metod klasycznych. Modele te są stabilne i odporne na występowanie w zbiorze uczącym szumu oraz wartości oddalonych. Charakteryzują się relatywnie wysoką dokładnością predykcji oraz przede wszystkim dają mniejsze błędy prognoz niż np. modele zbudowane na podstawie klasycznej metody najmniejszych kwadratów (zob. [Meyer, Leisch, Hornik 2002; Trzęsiok 2006; 2008]). Powstaje pytanie, na ile nieparametryczne metody regresji nadają się do analizy danych opisywanych przez dużą liczbę zmiennych.

W wielu przypadkach w pierwszym etapie analiz statystycznych badacz musi zdecydować, czy dokonać selekcji zmiennych objaśniających i do modelu regresyjnego wprowadzić tylko część z nich, czy zbudować model, opierając się na oryginalnym zestawie zmiennych. Niewątpliwą korzyścią z dokonania selekcji zmiennych jest ograniczenie złożoności modelu. Istnieją jednak obawy, że poprzez zastosowanie takiego podejścia utracimy część istotnych informacji, a co za tym idzie – obniżona zostanie dokładność predykcji modelu.

Celem artykułu było zbadanie, czy w przypadku nieparametrycznych metod regresji przeprowadzenie procedury eliminacji zmiennych i wprowadzenie do modelu tylko części z nich znacznie zmienia wartość błędu predykcji tego modelu. Do analizy wykorzystano modele zbudowane z wykorzystaniem wybranych nieparametrycznych metod regresji: POLYMARS oraz PPR. Wszystkie obliczenia wykonano za pomocą programu statystycznego **R**.

2. Wybrane nieparametryczne metody regresji

Nieparametryczne metody regresji stanowią grupę zróżnicowanych i dynamicznie rozwijających się metod. Do analizy, która ma na celu odpowiedź na pytanie, jaki wpływ ma wymiar przestrzeni zmiennych na jakość predykcji modeli regresji, wybrane zostały:

- wielowymiarowa metoda krzywych sklejanych POLYMARS (*Multivariate Adaptive Polynomial Spline Regression*),
- metoda rzutowania PPR (*Projection Pursuit Regression*).

2.1. Wielowymiarowa metoda krzywych sklejanych POLYMARS

Wielowymiarowa metoda POLYMARS została zaproponowana w 1997 r. przez Kooperberga, Bose'a i Stone'a. Jest ona modyfikacją metody MARS przedstawionej przez Friedmana w 1991 r.

Model regresyjny w metodzie POLYMARS można przedstawić w postaci addytywnej:

$$f(\mathbf{X}) = \alpha_0 + \sum_{k=1}^{K} \alpha_k h_k(\mathbf{X}), \tag{1}$$

w której $\mathbf{X} = (X_1, X_2, \ldots, X_m)$ jest wektorem zmiennych objaśniających, natomiast funkcje składowe h_k, przedstawione we wzorze (1), mają postać iloczynów tensorowych:

$$h_k(\mathbf{X}) = u_k \prod_{l=1}^{L_k} (X_{v(k,l)} - \xi_{v(k,l)})_+ , \tag{2}$$

gdzie $u_k \in \{-1,1\}$, $L_k \in \{1,2\}$ oznacza liczbę czynników (funkcji sklejanych pierwszego rzędu[1]) tworzących k-ty składnik modelu, $v(k,l) \in \{1,\ldots,m\}$ zaś wskazuje numer zmiennej tworzącej l-ty czynnik k -tego składnika modelu.

W metodzie POLYMARS algorytm budowy modelu regresyjnego składa się z dwóch głównych etapów: dołączania zmiennych do modelu oraz ich eliminacji.

Etap dołączania zmiennych polega na systematycznym wprowadzaniu do modelu takich funkcji składowych, które w największym stopniu redukują błąd średniokwadratowy oraz zachowują postać iloczynu tensorowego (2). Procedura ta powtarzana jest tak długo, aż osiągnięta zostanie zadana, maksymalna liczba funkcji bazowych modelu.

Wynikiem opisanej procedury dołączania zmiennych jest model, który ma bardzo dobre dopasowanie do danych, ale też wysoką złożoność. Następnym etapem jest więc eliminacja zmiennych. W każdym kroku algorytmu z modelu zostaje usunięta jedna funkcja składowa, mianowicie ta, której eliminacja powoduje naj-

[1] Funkcja sklejana pierwszego rzędu ma postać: $(X-\xi)_+ = \begin{cases} X-\xi, & \text{dla } X \geq \xi, \\ 0, & \text{dla } X < \xi. \end{cases}$ Zob. [Trzęsiok 2004a].

mniejszy wzrost błędu średniokwadratowego. W efekcie uzyskujemy ciąg modeli, z którego wybieramy ten, który jest najlepszy w sensie przyjętego kryterium. Ostatecznie procedura eliminacji zmiennych nieco obniża jakość końcowego modelu, ale znacznie zmniejsza jego złożoność.

Szczegółowo etapy procedur dołączania zmiennych oraz ich eliminacji przedstawione zostały w pracach: [Friedman 1991; Trzęsiok 2004a].

2.2. Metoda rzutowania PPR

Metoda rzutowania PPR (zob. [Friedman, Stuetzle 1981]) oparta jest na transformacji danych z przestrzeni wielowymiarowej w przestrzeń o dużo niższym wymiarze. Transformacja ta odbywa się przez rzutowanie wektora zmiennych objaśniających $\mathbf{X}$ w kierunkach $\boldsymbol{\alpha}_k$. W ten sposób uzyskuje się nowe zmienne:

$$Z_k = \boldsymbol{\alpha}_k^T \cdot \mathbf{X} \text{ dla } k = 1,...,K, \tag{3}$$

gdzie $\boldsymbol{\alpha}_k \in \mathbf{R}^n$ są unormowanymi wektorami, nazywanymi kierunkami rzutowania.

Celem tej transformacji jest próba odkrycia i zaobserwowania pewnych własności badanego zbioru, które nie są „widoczne" w przestrzeni o wyższym wymiarze.

Model zbudowany za pomocą metody rzutowania ma postać addytywną:

$$Y = f(\mathbf{X}) = \alpha_0 + \sum_{k=1}^{K} g_k\left(\boldsymbol{\alpha}_k^T \cdot \mathbf{X}, \boldsymbol{\beta}_k\right), \tag{4}$$

gdzie $\boldsymbol{\alpha}_k$ są kierunkami rzutowania, zaś funkcje składowe g_k to funkcje jednej zmiennej o parametrach zapisanych w postaci wektorów $\boldsymbol{\beta}_k$ (dla $k = 1,\ldots,K$).

Estymatory współrzędnych wektorów parametrów $\boldsymbol{\beta}_k$ i kierunków rzutowania $\boldsymbol{\alpha}_k$ otrzymuje się przez rozwiązanie zadania minimalizacji błędu empirycznego:

$$R_{emp}\left(\boldsymbol{\alpha}, \boldsymbol{\beta}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(\mathbf{x}_i)\right)^2, \tag{5}$$

gdzie $\boldsymbol{\alpha} = \left(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_K\right)$ oraz $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_K\right)$.

Rozwiązanie zadania minimalizacji wyrażenia opisanego wzorem (5) otrzymuje się iteracyjnie, w kolejnych krokach algorytmu, z wykorzystaniem odpowiedniej dekompozycji błędu średniokwadratowego. Szczegółowo algorytm ten został przedstawiony w pracach: [Cherkassky, Mulier 1998, s. 255-259; Trzęsiok 2004b].

3. Procedura eliminacji zmiennych z modelu

W tej części artykułu przedstawiono procedurę realizującą cel pracy. Określenie, jaki wpływ na jakość predykcji wybranych nieparametrycznych modeli regresji

ma wymiar przestrzeni zmiennych, oparte zostało na metodzie eliminacji zmiennych wykorzystującej strategię wspinaczki. Zaproponowana metoda polegała na usuwaniu z modelu, w każdym kroku algorytmu, jednej zmiennej. Eliminacja odbywała się według ustalonego *a priori* kryterium, którym w tym przypadku był minimalny błąd średniokwadratowy liczony metodą sprawdzania krzyżowego. W wyniku zastosowania opisanej procedury otrzymano rozwiązanie optymalne jedynie w sensie lokalnym. Zaletą tego podejścia była jednak stosunkowo niska złożoność algorytmu.

Procedurę eliminacji zmiennych z modelu można przedstawić w następujących krokach:

1) Za pomocą wybranej nieparametrycznej metody regresji zbuduj model regresyjny f_0, wykorzystując kompletny zbiór zmiennych objaśniających:

$$V_0 = \{X_1, X_2, \ldots, X_m\}.$$

2) Dla $j = 1, \ldots, m-1$ wykonaj kroki:

a) ze zbioru zmiennych objaśniających V_{j-1} usuń tymczasowo jedną zmienną, wykonując tę czynność kolejno dla każdej ze zmiennych, i zbuduj $(m-j+1)$ modeli regresyjnych;

b) dla wszystkich zbudowanych w poprzednim kroku modeli oblicz, metodą sprawdzania krzyżowego z podziałem zbioru danych na pięć części, błąd średniokwadratowy;

c) ostatecznie w kroku j wyeliminuj tę zmienną, której usunięcie w najmniejszym stopniu zmieniło dokładność predykcji modelu, a więc tę, dla której obliczony błąd średniokwadratowy jest najmniejszy. Zredukowany zbiór zmiennych oznacz przez V_j, natomiast uzyskany najmniejszy błąd średniokwadratowy zapamiętaj jako MSE_j;

d) przyjmij jako model f_j ten model regresyjny, który zbudowany był na zbiorze zmiennych oznaczonym przez V_j i któremu odpowiada błąd średniokwadratowy MSE_j.

3) Z otrzymanego ciągu modeli regresyjnych $\{f_j\}_{j=0,\ldots,m-1}$ (z malejącą liczbą zmiennych) wybierz ten model, dla którego błąd średniokwadratowy MSE_j jest najmniejszy. Jest to model końcowy, zbudowany za pomocą wybranej nieparametrycznej metody regresji z wykorzystaniem procedury eliminacji zmiennych.

4. Wyniki analizy

Do analizy wybrano dwie opisane wcześniej nieparametryczne metody regresji: POLYMARS oraz PPR. Badanie przeprowadzono na zbiorach danych standardowo wykorzystywanych do badania własności różnych nieparametrycznych metod regresji. Wybrane charakterystyki tych zbiorów przedstawione zostały w tab. 1.

Tabela 1. Charakterystyki zbiorów danych wykorzystanych w analizie

Zbiór danych	Liczba zmiennych objaśniających	Liczba obserwacji
Triazines	58	186
Peak	50	200
Bank	32	1000
Boston	13	506

Źródło: opracowanie własne.

Zgodnie z zaproponowanym algorytmem na każdym zbiorze danych zbudowany został model na podstawie wybranej nieparametrycznej metody regresji. Następnie w modelach tych systematycznie dokonywano eliminacji zmiennych. Uzyskane wyniki przedstawione zostały w tab. 2-5. Ze względu na ograniczenia objętości artykułu szczegółowo przedstawiono etapy omawianej procedury tylko dla części uzyskanych modeli, natomiast podsumowanie wyników wszystkich przeprowadzonych analiz przedstawiono w tab. 6.

Tabela 2. Wyniki procedury eliminacji zmiennych dla zbioru *Peak* dla metody POLYMARS

Etap	Numer wyrzuconej zmiennej	*MSE* modelu	Etap	Numer wyrzuconej zmiennej	*MSE* modelu
0	komplet zm.	25,6441	25	47	17,3772
1	18	23,7334	26	48	17,3772
2	7	22,3758	27	49	**17,3772**
3	46	21,1698	28	23	18,6653
4	32	20,5098	29	28	18,9453
5	38	19,5457	30	13	20,4511
6	11	18,6882	31	4	20,1714
7	17	18,4745	32	6	20,9837
8	1	18,4066	33	50	20,9837
9	45	17,9689	34	24	21,4040
10	8	17,9689	35	19	22,6129
11	9	17,9689	36	31	22,6129
12	14	17,9689	37	37	22,6272
13	16	17,9689	38	36	22,8831
14	20	17,9689	39	12	24,0297
15	22	17,9689	40	29	26,8058
16	10	17,8519	41	41	28,8960
17	26	17,8519	42	35	27,4632
18	27	17,8519	43	2	28,5568
19	30	17,3772	44	42	28,3959
20	33	17,3772	45	5	28,6930
21	34	17,3772	46	21	31,6917
22	40	17,3772	47	3	28,1755
23	43	17,3772	48	15	
24	44	17,3772			

Źródło: opracowanie własne.

Tabela 3. Wyniki procedury eliminacji zmiennych dla zbioru *Bank* dla metody rzutowania PPR

Etap	Numer wyrzuconej zmiennej	*MSE* modelu	Etap	Numer wyrzuconej zmiennej	*MSE* modelu
0	komplet zm.	0,0129	17	27	0,0083
1	11	0,0094	18	2	0,0081
2	1	0,0091	19	20	0,0083
3	19	0,0094	20	29	0,0083
4	31	0,0090	21	3	0,0083
5	5	0,0089	22	10	0,0082
6	30	0,0088	23	14	0,0079
7	17	0,0087	24	21	0,0082
8	4	0,0088	25	7	0,0081
9	28	0,0086	26	26	0,0082
10	8	0,0087	27	9	**0,0079**
11	25	0,0082	28	24	0,0081
12	22	0,0086	29	23	0,0086
13	16	0,0084	30	18	0,0100
14	13	0,0085	31	6	0,0113
15	15	0,0083	32	12	
16	32	0,0083			

Źródło: opracowanie własne.

Tabela 4. Wyniki procedury eliminacji zmiennych dla zbioru *Triazines* dla metody rzutowania PPR

Etap	Numer wyrzuconej zmiennej	*MSE* modelu	Etap	Numer wyrzuconej zmiennej	*MSE* modelu
0	komplet zm.	0,0644	30	58	0,0260
1	17	0,0395	31	41	0,0317
2	23	0,0435	32	21	0,0279
3	40	0,0451	33	34	0,0254
4	18	0,0426	34	38	0,0258
5	26	0,0427	35	50	0,0231
6	9	0,0407	36	47	0,0230
7	31	0,0427	37	19	0,0227
8	37	0,0399	38	32	0,0233
9	44	0,0370	39	30	0,0202
10	8	0,0320	40	45	0,0201
11	1	0,0351	41	15	0,0195
12	51	0,0310	42	54	0,0200
13	13	0,0291	43	39	0,0194
14	28	0,0364	44	4	0,0191
15	53	0,0384	45	12	0,0194
16	46	0,0292	46	48	0,0193
17	36	0,0332	47	29	0,0188
18	3	0,0418	48	42	0,0193
19	14	0,0395	49	10	0,0189
20	2	0,0351	50	20	**0,0184**
21	5	0,0338	51	43	0,0185
22	55	0,0316	52	27	0,0185
23	56	0,0335	53	22	0,0185
24	7	0,0374	54	24	0,0186
25	6	0,0329	55	25	0,0185
26	49	0,0287	56	57	0,0192
27	16	0,0292	57	11	0,0239
28	35	0,0263	58	33	
29	52	0,0282			

Źródło: opracowanie własne.

Tabela 5. Wyniki procedury eliminacji zmiennych dla zbioru *Boston* dla metody POLYMARS

Etap	Numer wyrzuconej zmiennej	*MSE* modelu	Etap	Numer wyrzuconej zmiennej	*MSE* modelu
0	komplet zm.	13,7237	7	1	11,9015
1	2	12,8769	8	7	12,3357
2	4	12,8618	9	11	12,4137
3	9	11,8943	10	8	14,3339
4	5	11,5780	11	10	19,5492
5	3	**11,5780**	12	6	26,8435
6	12	11,7152	13	13	

Źródło: opracowanie własne.

W pierwszej oraz czwartej kolumnie każdej z tabel 2-5 znajduje się numer wykonanego etapu algorytmu. W kolumnach drugiej i piątej umieszczony został numer kolejnej zmiennej wyeliminowanej z modelu w kroku j, natomiast w pozostałych kolumnach przestawiono wartości błędu średniokwadratowego modelu otrzymanego w tym kroku.

Tabela 6. Błędy średniokwadratowe modeli zbudowanych za pomocą wybranych nieparametrycznych metod regresji dla każdego z analizowanych zbiorów danych

Zbiór danych	Oryginalna liczba zmiennych	POLYMARS			PPR		
		Liczb zmiennych wyelimi-nowanych	*MSE* począt-kowego modelu	*MSE* końcowego modelu	Liczba zmiennych wyelimi-nowanych	*MSE* począt-kowego modelu	*MSE* końcowego modelu
Triazines	58	55	0,0240	0,0198	50	0,0644	0,0184
Peak	50	27	25,6441	17,3772	29	59,6437	32,9196
Bank	32	12	1,8178	0,0078	27	0,0129	0,0079
Boston	13	5	13,7237	11,5780	3	14,0885	11,3309

Źródło: opracowanie własne.

W tabeli 6 przedstawiono wyniki analizy dla modeli otrzymanych za pomocą obu opisanych nieparametrycznych metod regresji: POLYMARS oraz PPR. W kolumnach czwartej i siódmej przedstawiono wartości błędu średniokwadratowego uzyskanego dla modeli zbudowanych na oryginalnym komplecie zmiennych. Kolumny piąta i ósma zawierają informację o najniższych wartościach *MSE* uzyskanego dla modeli, w których zastosowano opisany algorytm eliminacji zmiennych. Liczby zmiennych ostatecznie wyeliminowanych z modeli przedstawione są w kolumnach 3 i 6. Zaprezentowane wyniki pokazują, że w każdym przypadku procedura eliminacji zmiennych prowadzi do uzyskania modelu regresyjnego charakteryzującego się niższym błędem średniokwadratowym.

5. Podsumowanie

W algorytmach obu przedstawionych nieparametrycznych metod regresji zawarty jest mechanizm doboru zmiennych do modelu. W metodzie POLYMARS

ujęty jest on w etapie eliminacji zmiennych, zaś w metodzie rzutowania PPR odbywa się przez transformację zmiennych – rzutowanie. Pomimo to, zastosowanie opisanej prostej procedury eliminacji zmiennych, w każdym z omawianych przypadków, zwiększyło dokładność predykcji modelu. Przedstawione wyniki analizy pokazują, że modele zbudowane na zredukowanym zbiorze zmiennych dają mniejsze wartości błędu średniokwadratowego.

Poza poprawą dokładności predykcji, procedura eliminacji zmiennych pozwala również na uzyskanie modelu o mniejszej złożoności niż model zbudowany na oryginalnym zestawie zmiennych objaśniających. Wadą tego podejścia jest jednak znaczne zwiększenie czasu obliczeń.

Literatura

Cherkassky V., Mulier F. (1998), *Learning from data – concepts, theory, and methods*, Wiley, New York.

Friedman J.H. (1991), *Multivariate adaptive regression splines*, „Annals of Statistics" nr 19, s. 1-141.

Friedman J.H., Stuetzle W. (1981), *Projection pursuit regression*, „Journal of the American Statistical Association" nr 76, s. 817-823.

Kooperberg C., Bose S., Stone C.J., *Polychotomous regression*, „Journal of the American Statistical Association" 1997 nr 92, s. 117-127.

Meyer D., Leisch F., Hornik K. (2002), *Benchmarking support vector machines*, Report no 78, Vienna University of Economics and Business Administration, http://www.wuwien.ac.at/am/Download/report78.pdf.

Trzęsiok J. (2004a), *Wybrane nieparametryczne metody regresji i ich zastosowania*, Taksonomia 11, *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1022, AE, Wrocław, s. 107-115.

Trzęsiok J. (2004b), *Metoda rzutowania w budowie modelu regresyjnego*, [w:] *Postępy ekonometrii*, red. A.S. Barczak, AE, Wrocław, s. 121-130.

Trzęsiok J. (2006), *Analiza wybranych własności metody MART*, [w:] Taksonomia 13, *Klasyfikacja i analiza danych*, red. K. Jajuga, M. Walesiak, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1126, AE, Wrocław, s. 510-518.

Trzęsiok J. (2008), *Ocena zasadności łączenia wybranych nieparametrycznych modeli regresji*, [w:] Taksonomia 15, *Klasyfikacja i analiza danych*, red. K. Jajuga, M. Walesiak, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 1207, UE, Wrocław, s. 346-353.

THE IMPACT OF THE NUMBER OF VARIABLES ON THE PREDICTION ACCURACY IN SELECTED NONPARAMETRIC REGRESSION MODELS

Summary

The paper presents the procedure for variable selection for regression models built with the use of two nonparametric methods: POLYMARS and projection pursuit regression. The results obtained on the benchmark data sets show that using the procedure for the reduction of the number of predictors yields models with smaller mean squared errors than models built on the complete set of the input variables.