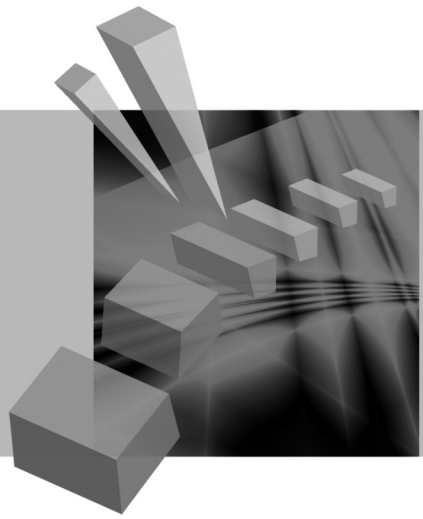
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

242

Taksonomia 19.

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pociecha,
Mirosław Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
oraz w The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon http://kangur.uek.krakow.pl/
bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM
Nakład: 320 egz.

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 242
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS 2012

Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Mariusz Kubus
Politechnika Opolska

ZASTOSOWANIE WSTĘPNEGO UWARUNKOWANIA ZMIENNEJ OBJAŚNIANEJ DO SELEKCJI ZMIENNYCH

Streszczenie: Paul i in. [2008] zaproponowali metodę wstępnego uwarunkowania zmiennej objaśnianej, którą realizuje się w dwóch krokach. W pierwszym przewidywane są wartości zmiennej objaśnianej za pomocą metody głównych składowych z nauczycielem. W drugim budowany jest jeszcze raz model regresji, ale na zbiorze uczącym, w którym rolę zmiennej objaśnianej odgrywają jej wartości teoretyczne estymowane w kroku pierwszym. Tu wykorzystuje się LASSO, które pozwala wyeliminować zmienne nieistotne. W artykule zaproponowano modyfikację oryginalnej metody oraz za pomocą symulacji oceniono ich przydatność w zadaniu selekcji zmiennych w przypadku modelu liniowego z interakcjami.

Słowa kluczowe: selekcja zmiennych, metoda wstępnego uwarunkowania, interakcje.

1. Wstęp

Zagadnienie selekcji zmiennych zajmuje obecnie znaczące miejsce w literaturze z zakresu statystycznych metod uczenia czy *data mining*. Zdolności predykcyjne modeli zależą w dużym stopniu od jakości danych, te natomiast zawierają często wiele zmiennych nieistotnych. Modelowanie w tych warunkach prowadzi do nadmiernego dopasowania do danych, a co za tym idzie, mniej dokładnych predykcji dla nowych obiektów, które nie brały udziału w etapie uczenia. Identyfikacja zmiennych istotnie wpływających na badane zjawisko ma walory interpretacyjne oraz może się przyczynić do zmniejszenia kosztów gromadzenia danych w przyszłości. W ostatnich latach pojawił się nowy obszar zastosowań dla selekcji zmiennych, który stanowi duże wyzwanie dla statystyków. Mowa o biologii obliczeniowej, a dokładniej o analizie macierzy ekspresji genów pozyskiwanych z DNA. Jednym z podstawowych zadań jest tu znajdywanie genów reagujących zmianą ekspresji na zmiany środowiskowe (np. podanie leku). Analizowane zbiory danych liczą nawet dziesiątki tysięcy zmiennych (genów) i jednocześnie nieliczne próby uczące (ze względu na kosztowność badań), zwykle nie więcej niż 200, a czasem tylko kilkadziesiąt obserwacji. Modelowanie predykcyjne jest wtedy szczególnie trudne ze względu na: brak stopni swobody, problem nadmiernego dopasowania do danych, niestabilność esty-

mowanych współczynników oraz problem ze współliniowością. Takim zbiorom danych szczególnie dedykowana jest dyskutowana w tym artykule metoda wstępnego uwarunkowania zmiennej objaśnianej (*pre-condidtioning*) zaproponowana przez Paula i in. [2008].

Prezentowane w literaturze metody redukcji wymiaru przestrzeni cech można podzielić na dwa główne podejścia. Pierwsze polega na selekcji oryginalnych zmiennych wejściowych na podstawie wybranego kryterium. Istotną rolę odgrywa tu technika przeszukiwania przestrzeni wszystkich podzbiorów zmiennych. Wszystkie metody w tym nurcie klasyfikowane są obecnie do trzech grup (zob. np. [Guyon i in. 2006]): dobór zmiennych przed zastosowaniem algorytmu uczącego (*filters*), wyszukiwanie optymalnego podzbioru związane z oceną jakości modelu (*wrappers*) lub selekcja wewnątrz algorytmu uczącego (*embedded methods*). Drugie podejście do problemu redukcji wymiaru polega na transformacji oryginalnych zmiennych do przestrzeni o mniejszym wymiarze. Obok popularnej metody głównych składowych zaproponowano też jej modyfikację: metodę głównych składowych z nauczycielem (*supervised principal components*) [Bair i in. 2006]. Innym praktykowanym sposobem jest taksonomia cech (zob. np. [Li, Hong 2001; Hastie i in. 2001]). Dyskutowana w artykule metoda *pre-conditioning* wykracza jednak poza ramy podanej klasyfikacji.

W artykule proponowana jest modyfikacja metody wstępnego uwarunkowania zmiennej objaśnianej. Za pomocą symulacji dokonano oceny przydatności w zadaniu selekcji zmiennych zarówno metody oryginalnej, jak i proponowanej modyfikacji. Badania przeprowadzono dla modelu liniowego z interakcjami, a między predyktorami wprowadzano zależności liniowe. Wzięto również pod uwagę przypadek, gdy liczba zmiennych jest znacznie większa od liczby obiektów.

2. Metoda wstępnego uwarunkowania zmiennej objaśnianej

Metoda wstępnego uwarunkowania zmiennej objaśnianej (*pre-conditioning*) [Paul i in. 2008] realizowana jest w dwóch krokach. W pierwszym przewidywane są wartości zmiennej objaśnianej za pomocą metody głównych składowych z nauczycielem. Następnie empiryczne wartości zmiennej objaśnianej zamieniane są wartościami teoretycznymi estymowanymi przez metodę głównych składowych z nauczycielem. W drugim kroku budowany jest jeszcze raz model regresji, ale na zbiorze uczącym, w którym rolę zmiennej objaśnianej odgrywają jej wartości teoretyczne estymowane w kroku pierwszym. Tu proponuje się metodę LASSO [Tibshirani 1996], która pozwala wyeliminować zmienne nieistotne. Główną ideą całej tej procedury jest wyeliminowanie szumu w kroku pierwszym, by uzyskać lepsze rezultaty selekcji zmiennych w kroku drugim.

2.1. Metoda głównych składowych z nauczycielem

Zaproponowana przez autorów pracy [Bair i in. 2006] metoda głównych składowych z nauczycielem (*supervised principal components*) jest odpowiedzią na problem

klasycznej metody głównych składowych stosowanej w regresji, gdzie kierunki maksymalizujące wariancję nie muszą być współliniowe ze zmienną objaśnianą. Główną ideą metody jest wyznaczenie głównych składowych tylko dla podzbioru zmiennych – najbardziej skorelowanych ze zmienną objaśnianą. Szczegółowo algorytm wygląda następująco. W pierwszym kroku oceniana jest ważność predyktorów, a więc budowany jest ranking. Oryginalnie funkcję kryterium pełni moduł współczynnika regresji prostej. Następnie dla różnych wartości progowych wyznaczanych jest *m* głównych składowych, ale tylko dla zmiennych, które w rankingu przekroczyły ustalony próg. Wykonywana jest regresja względem nich oraz wybierana optymalna wartość progowa i liczba głównych składowych za pomocą sprawdzania krzyżowego.

2.2. Propozycja modyfikacji

Oryginalnie zaproponowana metoda *pre-conditioning* pozostawia możliwości modyfikacji, polegające na zastosowaniu innych metod regresji w poszczególnych krokach. W artykule proponuje się wykorzystanie metody *elastic net* [Zou, Hastie 2005] do selekcji zmiennych. Z kolei do wstępnego uwarunkowania proponowana jest metoda POLYMARS [Kooperberg, Bose, Stone 1997], ze względu na jej odporność na zmienne nieistotne oraz walory predykcyjne. W przeprowadzonych wstępnie badaniach symulacyjnych dla modelu liniowego z interakcjami (2) metoda ta wykazywała zdecydowaną przewagę nad innymi nieparametrycznymi metodami regresji ze względu na dokładność predykcji oraz czas obliczeń.

3. Selekcja zmiennych za pomocą regularyzowanej regresji liniowej

Selekcja zmiennych za pomocą regularyzowanej regresji liniowej jest szczególnie rekomendowana w przypadku dużych wymiarów przestrzeni cech. Ogromną popularnością cieszy się zaproponowana przez Tibshiraniego [1996] metoda LASSO oraz algorytm LARS [Efron i in. 2004] oferujący również rozwiązanie przybliżone LASSO. LARS jest szybkim algorytmem, który wprowadza iteracyjnie do modelu zmienne niosące najwięcej informacji o niewyjaśnionej części zmienności zmiennej objaśnianej (zob. także [Kubus 2011]). Maksymalna liczba iteracji jest równa liczbie zmiennych. Ostatecznie decyzję o liczbie zmiennych wprowadzanych do modelu podejmuje się za pomocą kryteriów jakości modelu. W regularyzowanej regresji liniowej parametry modelu $\boldsymbol{\beta}$ estymowane są przez minimalizację sumy funkcji straty (często kwadratowej) oraz komponentu kary *P*, który zależy od wartości współczynników oraz parametrów regularyzacji. Zou i Hastie [2005] zaproponowali (*elastic net*):

$$P(\lambda,\alpha,\boldsymbol{\beta}) = \lambda \cdot \sum_{j=1}^{p}\left(\alpha\beta_j^2 + (1-\alpha)\left|\beta_j\right|\right), \tag{1}$$

gdzie λ decyduje o rozmiarze kary za wielkość współczynników β_j, p jest liczbą zmiennych objaśniających, natomiast parametr α nadaje wagi wyrażeniom zaczerpniętym z regresji grzbietowej oraz LASSO. Metoda ta cechuje się tym, że jeśli pewna zmienna objaśniająca jest istotna, to zmienne współliniowe z nią są włączane do modelu. Tej cechy nie ma LASSO. Wbrew powszechnemu postulatowi o niewprowadzaniu do modelu zmiennych redundantnych, własność ta może być w pewnych sytuacjach korzystna, np. można ją wykorzystać do modelowania interakcji, co zilustrowane będzie w przeprowadzonym dalej eksperymencie.

Zmodyfikowaną wersję LASSO zaproponował Meinshausen [2007]. W metodzie *relaxed* LASSO estymatory uzyskuje się w dwukrokowej procedurze. Najpierw algorytm LARS dokonuje selekcji zmiennych, a następnie za pomocą drugiego parametru regularyzacji wartości bezwzględne współczynników jeszcze raz są zmniejszane (*shrinkage*). Metoda ta jest szczególnie rekomendowana dla wysokich wymiarów przestrzeni cech i dużej liczby zmiennych nieistotnych. Wyniki, jakie uzyskał Meinshausen [2007], dotyczą jednak przypadku zmiennych niezależnych.

4. Eksperyment

Ponieważ w wielu zastosowaniach praktycznych można się spodziewać zależności między zmiennymi objaśniającymi oraz zachodzących między nimi interakcji, badania przeprowadzono właśnie w takich warunkach. Rozważany będzie model:

$$y = \beta_0 + \sum_{j=1}^{10} \beta_j x_j + \beta_{11} x_1 x_2 + \beta_{12} x_2 x_3 + \beta_{13} x_4 x_5 + \beta_{14} x_6 x_7 x_8 + \sigma\varepsilon, \tag{2}$$

gdzie $\varepsilon \sim N(0;1)$. Motywacją wyboru takiego modelu był też fakt, że symulacje prezentowane w artykułach źródłowych ([Bair i in. 2006; Meinshausen 2007; Paul i in. 2008; Zou, Hastie 2005]) przeprowadzane są na ogół dla klasycznego modelu liniowego względem zmiennych, czasem tylko dla przypadku zmiennych ortogonalnych. Liczebności generowanych zbiorów uczących oraz testowych ustalono odpowiednio na 200 oraz 500. Szum gaussowski wprowadzano tylko w zbiorach uczących, a jego poziom losowano według formuły: $\sigma = m \cdot sd(y)$, gdzie $m \in \{0{,}5;\ 0{,}6;\ 0{,}7;\ 0{,}8;0{,}9;1\}$. Rozważono dwa przypadki ze względu na liczbę zmiennych p. Po wprowadzeniu dodatkowych zmiennych nieistotnych do zbiorów danych uzyskano: $p = 100$ lub $p = 1000$. Wprowadzano też zależności liniowe między zmiennymi w ten sposób, by co piąty predyktor był kombinacją liniową poprzedzających go czterech:

$$x_{5+k*5} = \alpha_{k1} x_{1+k*5} + \alpha_{k2} x_{2+k*5} + \alpha_{k3} x_{3+k*5} + \alpha_{k4} x_{4+k*5} + \sigma_k \varepsilon, \tag{3}$$

dla $k \in \{0,1, ..., p/5-1\}$. Tym razem dla uzyskania silnej zależności poziom szumu losowano nieco mniejszy: $\sigma_k = m_k \cdot sd(x_{5+k*5})$, gdzie $m_k \in \{0{,}1;\ 0{,}2;\ 0{,}3;\ 0{,}4\}$.

Zwróćmy uwagę, że zależności (3) wprowadzono zarówno dla zmiennych nieistotnych, jak i dla zmiennych z modelu. Wszystkie współczynniki (modelu (2) oraz w zależności (3)), a także realizacje zmiennych objaśniających (z wyjątkiem x_{5+k*5}) generowano losowo z jednowymiarowego, standaryzowanego rozkładu normalnego.

W badaniu wzięto pod uwagę następujące metody:

- metodę głównych składowych z nauczycielem (spc),
- oryginalną metodę wstępnego uwarunkowania zmiennej objaśnianej (spc+lars),
- metodę głównych składowych z nauczycielem do wstępnego uwarunkowania oraz *elastic net* do selekcji zmiennych (spc+en),
- metodę POLYMARS do wstępnego uwarunkowania oraz *elastic net* do selekcji zmiennych (polymars+en).

Liczby wprowadzanych do modelu zmiennych obrazuje tab. 1.

Tabela 1. Średnie liczby wybieranych zmiennych nieistotnych/zmiennych z modelu (z odchyleniami standardowymi). W nawiasach podano mediany. Wyniki dla 100 symulacji

	spc	spc+lars	spc+en	polymars+en
$p = 100$	10,6±18,8 / 5,8±1,9 (2) / (6)	13,7±22,2 / 6,55±2,1 (3) / (6,5)	7,8±9,2 / 6,7±1,8 (5) / (7)	6,7±6,65 / 6,7±1,8 (4) / (7)
$p = 1000$	47,0±138,7 / 4,9±1,5 (6) / (5)	42,7±69,8 / 5,45±1,5 (7) / (6)	5,65±6,1 / 4,0±1,7 (5) / (4)	5,9±4,8 / 4,65±1,8 (4) / (4)

Źródło: obliczenia własne.

W przypadku $p = 100$ zaproponowane modyfikacje identyfikują więcej zmiennych z modelu. Mediany wprowadzanych zmiennych nieistotnych dla spc i spc+lars są wprawdzie mniejsze, ale średnie i odchylenia standardowe wysokie, co świadczy o niestabilności tych metod. W przypadku $p = 1000$ zaproponowane modyfikacje wprowadzają mniej zmiennych nieistotnych, a ponadto wyniki są bardziej stabilne. Na przykład metoda spc wprowadzała czasem ponad 600 zmiennych nieistotnych do modelu, natomiast oryginalna metoda *pre-conditioning* (spc+lars) dość często ok. 200 zmiennych nieistotnych. Zaproponowane modyfikacje identyfikowały jednak mniej zmiennych z modelu. Uzyskane wyniki należałoby więc zweryfikować błędami predykcji.

W tym miejscu należy podkreślić, że dyskutowane metody stosowano wyłącznie w celu wstępnej selekcji zmiennych. Celem jest zbudowanie modelu liniowego z interakcjami, więc w drugim kroku analizy wprowadzone będą wyrażenia interakcyjne dla uzyskanego podzbioru zmiennych. Takie postępowanie wiąże się z szybkim wzrostem wymiaru przestrzeni cech i nie zawsze jest praktyczne. W przypadku zaproponowanych w artykule modyfikacji wzrost ten nie jest znaczny. Dla $p = 1000$

wymiar nowej przestrzeni w żadnej ze 100 symulacji nie przekraczał 200, a więc był przynajmniej pięciokrotnie mniejszy od wymiaru przestrzeni pierwotnej. Do selekcji zmiennych w drugim etapie analizy zastosowano metodę *relaxed* LASSO. Błędy predykcji estymowane na zbiorach testowych dla tak uzyskanych modeli liniowych z interakcjami ilustrują tab. 2-3. Dla porównania przedstawiono też błędy dla metody POLYMARS, którą stosowano na pełnym zestawie zmiennych. Dla zbadania istotności różnic między średnimi zastosowano test rangowanych znaków Wilcoxona dla obserwacji zestawionych w pary (zob. np. [Aczel 2005]). W przypadku $p = 100$ średnie w zaproponowanych modyfikacjach okazały się istotnie mniejsze od średniej uzyskanej w oryginalnej metodzie *pre-conditioning* (p-wartości w porównaniach z spc+en oraz polymars+en wynosiły odpowiednio 0,0022 oraz 0,0013). Z kolei nie ma istotnych różnic w błędach predykcji pomiędzy proponowanymi modyfikacjami a metodą POLYMARS stosowaną dla pełnego zestawu zmiennych wejściowych.

Tabela 2. Średnie (ze 100 symulacji) błędy predykcji estymowane na zbiorach testowych wraz z odchyleniami standardowymi w przypadku 100 zmiennych objaśniających

	POLYMARS	spc+lars	spc+en	polymars+en
błąd predykcji	2,39 ± 0,93	2,73 ± 0,87	2,39 ± 0,90	2,38 ± 1,04

Źródło: obliczenia własne.

W przypadku $p = 1000$ (tab. 3) średnie błędy predykcji dla modeli liniowych z selekcją zmiennych za pomocą spc+en lub polymars+en są istotnie mniejsze od średniej uzyskanej metodą POLYMARS (p-wartości wynosiły odpowiednio 0,0016 oraz 0,0435).

Tabela 3. Średnie (ze 100 symulacji) błędy predykcji estymowane na zbiorach testowych wraz z odchyleniami standardowymi w przypadku 1000 zmiennych objaśniających

	POLYMARS	spc+en	polymars+en
błąd predykcji	3,68 ± 1,92	3,18 ± 1,30	3,59 ± 2,01

Źródło: obliczenia własne.

5. Podsumowanie

W artykule zaproponowano modyfikację metody wstępnego uwarunkowania zmiennej objaśnianej (*pre-conditioning*) polegającą na zastosowaniu innych metod regresji w wymaganych krokach. Badanie przeprowadzono dla modelu liniowego z interakcjami, gdzie między predyktorami zachodziły zależności liniowe. Dzięki zaproponowanym modyfikacjom uzyskano lepsze rezultaty w eliminacji zmiennych nieistotnych, co pozwoliło na wprowadzanie wyrażeń interakcyjnych bez nadmiernego

wzrostu wymiaru badanej przestrzeni. Po ponownej selekcji zmiennych metodą *relaxed* LASSO uzyskano modele liniowe o mniejszych błędach predykcji w porównaniu z oryginalną metodą *pre-conditioning*. W przypadku gdy liczba zmiennych była znacznie większa od liczby obserwacji, uzyskano też nieco mniejsze błędy od metody POLYMARS stosowanej na pełnym zestawie zmiennych wejściowych.

Literatura

Aczel A.D., *Statystyka w zarządzaniu,* PWN, Warszawa 2005.

Bair E., Hastie T., Paul D., Tibshirani R., *Prediction by supervised principal components,* „J. Amer. Statist. Assoc." 2006, no 101.

Efron B., Hastie T., Johnstone I., Tibshirani R., *Least angle regression*, „Annals of Statistics" 2004, no 32 (2).

Guyon I., Gunn S., Nikravesh M., Zadeh L., *Feature Extraction: Foundations and Applications,* Springer, New York 2006.

Hastie T., Tibshirani R., Botstein D., Brown P., *Supervised harvesting of expression trees*, „Genome Biol." 2001, no 2.

Kooperberg C., Bose S., Stone C.J., *Polychotomous regression*, „Journal of the American Statistical Association" 1997, no 92.

Kubus M., *Analiza metody LARS w problemie selekcji zmiennych w regresji,* [w:] K. Jajuga, M. Walesiak (red.), Taksonomia 18, *Klasyfikacja i analiza danych – teoria i zastosowania,* Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 176, UE, Wrocław 2011.

Li H., Hong F., *Cluster-Rasch models for microarray gene expression data,* „Genome Biol." 2001, no 2.

Meinshausen N., *Lasso with relaxation*, „Computational Statistics and Data Analysis" 2007, no. 52(1).

Paul D., Bair E., Hastie T., Tibshirani R., *"Pre-conditioning" for feature selection and regression in high-dimensional problems,* „Annals of Statistics" 2008, no 36(4).

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

Zou H., Hastie T., *Regularization and variable selection via the elastic net,* „Journal of the Royal Statistical Society Series B." 2005, no 67(2).

THE APPLICATION OF PRE-CONDITIONING OF EXPLANATORY VARIABLE FOR FEATURE SELECTION

Summary: Paul et al. [2008] proposed pre-conditioning of explanatory variable that consists of two steps. Firstly, supervised principal components are applied for prediction and then regression model is constructed again. This time the response is replaced with the previously obtained predictions. The second regression model is LASSO, which eliminates irrelevant variables. The modification of original pre-conditioning is proposed in this paper. The usefulness of original pre-conditioning as well as proposed modifications are assessed using simulations. The experiment was carried out for linear model with interactions.

Keywords: feature selection, pre-conditioning, interactions.