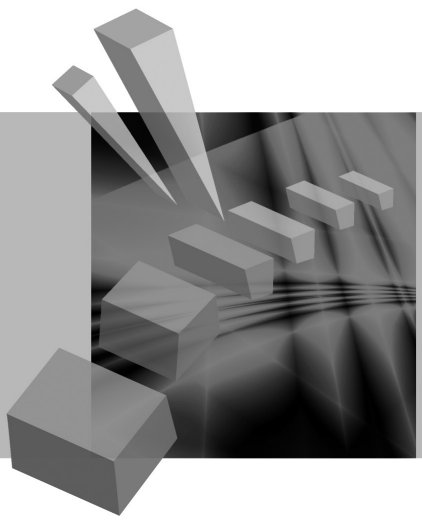
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

279

Taksonomia 21

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2013

Redaktor Wydawnictwa: Aleksandra Śliwka
Redaktor techniczny: Barbara Łopusiewicz
Korektor: Barbara Cibis
Łamanie: Małgorzata Czupryńska
Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
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

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy danych PTS

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

Wersja pierwotna: publikacja drukowana
Druk: Drukarnia TOTEM

Spis treści

Wstęp..... 9

Sabina Denkowska, Kamil Fijorek, Marcin Salamaga, Andrzej Sokołowski: Sejm VI kadencji – maszynka do głosowania 11
Barbara Pawełek, Adam Sagan: Zmienne ukryte w modelach ekonomicznych – respecyfikacja modelu Kleina I 19
Jan Paradysz: Nowe możliwości badania koniunktury na rynku pracy 29
Krzysztof Najman: Samouczące się sieci GNG w grupowaniu dynamicznym zbiorów o wysokim wymiarze 41
Kamila Migdał-Najman: Zastosowanie jednowymiarowej sieci SOM do wyboru cech zmiennych w grupowaniu dynamicznym 48
Aleksandra Matuszewska-Janica, Dorota Witkowska: Zróżnicowanie płac ze względu na płeć: zastosowanie drzew klasyfikacyjnych 58
Iwona Foryś, Ewa Putek-Szeląg: Przestrzenna klasyfikacja gmin ze względu na sprzedaż użytków gruntowych zbywanych przez ANR w województwie zachodniopomorskim 67
Joanna Banaś, Małgorzata Machowska-Szewczyk: Klasyfikacja internetowych rachunków bankowych z uwzględnieniem zmiennych symbolicznych..... 77
Marta Jarocka: Wpływ metody doboru cech diagnostycznych na wynik porządkowania liniowego na przykładzie rankingu polskich uczelni 85
Anna Zamojska: Badanie zgodności rankingów wyznaczonych według różnych wskaźników efektywności zarządzania portfelem na przykładzie funduszy inwestycyjnych..... 95
Dorota Rozmus: Porównanie dokładności taksonomicznej metody propagacji podobieństwa oraz zagregowanych algorytmów taksonomicznych opartych na idei metody *bagging* 106
Ewa Wędrowska: Wrażliwość miar dywergencji jako mierników niepodobieństwa struktur..... 115
Katarzyna Wójcik, Janusz Tuchowski: Wpływ automatycznego tłumaczenia na wyniki automatycznej identyfikacji charakteru opinii konsumenckich ... 124
Małgorzata Misztal: Ocena wpływu wybranych metod imputacji na wyniki klasyfikacji obiektów w modelach drzew klasyfikacyjnych..... 135
Anna Czapkiewicz, Beata Basiura: Badanie wpływu wyboru współczynnika zależności na grupowanie szeregów czasowych 146
Tomasz Szubert: Czynniki różnicujące poziom zadowolenia z życia oraz wartości życiowe osób sprawnych i niepełnosprawnych w świetle badań „Diagnozy społecznej” 154

Marcin Szymkowiak: Konstrukcja estymatorów kalibracyjnych wartości globalnej dla różnych funkcji odległości 164
Wojciech Roszka: Szacowanie łącznych charakterystyk cech nieobserwowanych łącznie 174
Justyna Brzezińska: Metody wizualizacji danych jakościowych w programie **R** 182
Agata Sielska: Regionalne zróżnicowanie potencjału konkurencyjnego polskich gospodarstw rolnych w województwach po akcesji do Unii Europejskiej 191
Mariusz Kubus: Liniowy model prawdopodobieństwa z regularyzacją jako metoda doboru zmiennych 201
Beata Basiura: Metoda Warda w zastosowaniu klasyfikacji województw Polski z różnymi miarami odległości 209
Katarzyna Wardzińska: Wykorzystanie metody obwiedni danych w procesie klasyfikacji przedsiębiorstw 217
Katarzyna Dębkowska: Modelowanie upadłości przedsiębiorstw oparte na próbach niezbilansowanych 226
Danuta Tarka: Wpływ metody doboru cech diagnostycznych na wyniki klasyfikacji obiektów na przykładzie danych dotyczących ochrony środowiska .. 235
Artur Czech: Zastosowanie wybranych metod doboru zmiennych diagnostycznych w badaniach konsumpcji w ujęciu pośrednim 246
Beata Bal-Domańska: Ocena relacji zachodzących między inteligentnym rozwojem a spójnością ekonomiczną w wymiarze regionalnym z wykorzystaniem modeli panelowych 255
Mariola Chrzanowska: *Ordinary kriging* i *inverse distance weighting* jako metody szacowania cen nieruchomości na przykładzie warszawskiego rynku 264
Adam Depta: Zastosowanie analizy wariancji w badaniu jakości życia na podstawie kwestionariusza SF-36v2 272
Maciej Beręsewicz, Tomasz Klimanek: Wykorzystanie estymacji pośredniej uwzględniającej korelację przestrzenną w badaniach cen mieszkań 281
Karolina Paradysz: Benchmarkowa analiza estymacji dla małych obszarów na lokalnych rynkach pracy 291
Anna Gryko-Nikitin: Dobór parametrów w równoległych algorytmach genetycznych dla problemu plecakowego 301
Tomasz Ząbkowski, Piotr Jałowiecki: Zastosowanie reguł asocjacyjnych do analizy danych ankietowych w wybranych obszarach logistyki przedsiębiorstw przetwórstwa rolno-spożywczego 311
Agnieszka Przedborska, Małgorzata Misztal: Zastosowanie metod statystyki wielowymiarowej do oceny wydolności stawów kolanowych u pacjentów z chorobą zwyrodnieniową leczonych operacyjnie 321
Dorota Perło: Rozwój zrównoważony w wymiarze gospodarczym, społecznym i środowiskowym – analiza przestrzenna 331

Ewa Putek-Szeląg, Urszula Gierałtowska, Analiza i diagnoza wielkości produkcji energii odnawialnej w Polsce na tle krajów Unii Europejskiej .. 342

Summaries

Sabina Denkowska, Kamil Fijorek, Marcin Salamaga, Andrzej Sokołowski: VIth-term Sejm – a voting machine 18
Barbara Pawełek, Adam Sagan: Latent variables in econometric models – respecification of Klein I model 28
Jan Paradysz: New possibilities for studying the situation on the labour market 40
Krzysztof Najman: Self-learning neural network of GNG type in the dynamic clustering of high-dimensional data..... 47
Kamila Migdał-Najman: Applying the one-dimensional SOM network to select variables in dynamic clustering 57
Aleksandra Matuszewska-Janica, Dorota Witkowska: Gender wage gap: application of classification trees..... 66
Iwona Foryś, Ewa Putek-Szeląg: Spatial classification of communes by usable land traded by the APA in the Zachodniopomorskie voivodeship... 76
Joanna Banaś, Małgorzata Machowska-Szewczyk: Classification of Internet banking accounts including symbolic variables 84
Marta Jarocka: The impact of the method of the selection of diagnostic variables on the result of linear ordering on the example of ranking of universities in Poland..... 94
Anna Zamojska: Empirical analysis of the consistency of mutual fund ranking for different portfolio performance measures..... 105
Dorota Rozmus: Comparison of accuracy of affinity propagation clustering and cluster ensembles based on bagging idea..... 114
Ewa Wędrowska: Sensitivity of divergence measures as structure dissimilarity measurements 123
Katarzyna Wójcik, Janusz Tuchowski: Machine translation impact on the results of the sentiment analysis 134
Małgorzata Misztal: Assessment of the influence of selected imputation methods on the results of object classification using classification trees ... 145
Anna Czapkiewicz, Beata Basiura: Simulation study of the selection of coefficient depending on the clustering time series..... 153
Tomasz Szubert: Factors differentiating the level of satisfaction with life and the life's values of people with and without disabilities in the light of the "Social Diagnosis" survey 162
Marcin Szymkowiak: Construction of calibration estimators of totals for different distance measures 173

Wojciech Roszka: Joint characteristics' estimation of variables not jointly observed ... 181
Justyna Brzezińska: Visualizing categorical data in **R** ... 190
Agata Sielska: Regional diversity of competitiveness potential of Polish farms after the accession to the European Union ... 200
Mariusz Kubus: Regularized linear probability model as a filter ... 208
Beata Basiura: The Ward method in the application for classification of Polish voivodeships with different distances ... 216
Katarzyna Wardzińska: Application of Data Envelopment Analysis in company classification process ... 225
Katarzyna Dębkowska: Modeling corporate bankruptcy based on unbalanced samples ... 234
Danuta Tarka: Influence of the features selection method on the results of objects classification using environmental data ... 245
Artur Czech: Application of chosen methods for the selection of diagnostic variables in indirect consumption research ... 254
Beata Bal-Domańska: Assessment of relations occurring between smart growth and economic cohesion in regional dimension using panel models 263
Mariola Chrzanowska: Ordinary kriging and inverse distance weighting as methods of estimating prices based on Warsaw real estate market ... 271
Adam Depta: Application of analysis of variance in the study of the quality of life based on questionnaire SF-36v2 ... 280
Maciej Beręsewicz, Tomasz Klimanek: Using indirect estimation with spatial autocorrelation in dwelling price surveys ... 290
Karolina Paradysz: Benchmark analysis of small area estimation on local labor markets ... 300
Anna Gryko-Nikitin: Selection of various parameters of parallel evolutionary algorithm for knapsack problems ... 310
Tomasz Ząbkowski, Piotr Jałowiecki: Application of association rules for the survey of data analysis in the selected areas of logistics in food processing companies ... 320
Agnieszka Przedborska, Małgorzata Misztal: Using multivariate statistical methods to assess the capacity of the knee joint among the patients treated surgically for osteoarthritis ... 330
Dorota Perło: Sustainable development in the economic, social and environmental dimensions – spatial analysis ... 341
Ewa Putek-Szeląg, Urszula Gierałtowska: Analysis and diagnosis of the volume of renewable energy production in Poland compared to EU countries ... 352

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 279 • 2013

Taksonomia 21. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Mariusz Kubus
Politechnika Opolska

LINIOWY MODEL PRAWDOPODOBIEŃSTWA Z REGULARYZACJĄ JAKO METODA DOBORU ZMIENNYCH

Streszczenie: W artykule zaproponowano zastosowanie liniowego modelu prawdopodobieństwa z regularyzacją jako narzędzia doboru zmiennych przed regresją logistyczną. W etapie selekcji zmiennych dodatkowo stosowano sprawdzanie krzyżowe. Takie podejście zapewnia skuteczniejszą eliminację zmiennych nieistotnych od powszechnie stosowanej regularyzowanej regresji logistycznej, a błędy klasyfikacji porównywanych metod nie różnią się w sposób statystycznie istotny. W badaniach empirycznych wykorzystano zbiory z repozytorium Uniwersytetu Kalifornijskiego, a sztucznie wprowadzane zmienne nieistotne generowano z rozkładów zero-jedynkowego lub normalnego.

Słowa kluczowe: selekcja zmiennych, regularyzacja, liniowy model prawdopodobieństwa.

1. Wstęp

W dobie dostępu do dużych baz danych powszechnie stosowane są metody *data mining* w celu wydobycia wiedzy z danych. Badacz przystępujący do analizy często nie ma wiedzy *a priori* na temat badanego zjawiska, dotyczy to zarówno specyfikacji modelu, jak i zmiennych istotnie wpływających na badane zjawisko, które w regresji i dyskryminacji reprezentowane jest przez zmienną objaśnianą. Można wskazać wiele metod cechujących się wysoką dokładnością predykcji czy klasyfikacji nowych obiektów (np. agregowane drzewa), jednak działają one na zasadzie czarnej skrzynki i brak im walorów interpretacyjnych. Między innymi dlatego wciąż atrakcyjne są modele liniowe, często bowiem celem analizy jest nie tylko predykcja czy klasyfikacja, ale odkrycie związków zachodzących między cechami statystycznymi. Pracownik sieci telefonii komórkowej (czy firmy ubezpieczeniowej) chciałby poznać przyczyny odejścia niektórych swych klientów, by odpowiednio wcześnie poczynić kroki w kierunku ich utrzymania. Inwestor (lub pracownik banku) jest zainteresowany poznaniem reguł przewidujących bankructwo firmy. Z kolei lekarzowi zależy na wczesnym zdiagnozowaniu choroby, by wybrać odpowiedni sposób leczenia (przy dzisiejszych technologiach dokonuje tego nieraz na podstawie danych zawierających ekspresje tysięcy genów pozyskiwanych z DNA).

W przypadku dużych zbiorów danych i braku przesłanek co do ważności zmiennych (wiedzy eksperta) stosowanie metod regresji czy dyskryminacji bez selekcji zmiennych prowadzi do modeli niestabilnych, nadmiernie dopasowanych do danych, a w efekcie do słabej jakości rozpoznawania dla nowych obiektów (spoza próby uczącej). Znajduje to teoretyczne odzwierciedlenie w tzw. kompromisie między obciążeniem a wariancją (*bias-variance trade-off*). Modele zbyt złożone, które są nadmiernie dopasowane do danych, cechują się małym obciążeniem i dużą wariancją błędu. Z drugiej strony modele zbyt proste, które nie wydobywają całej informacji z danych, charakteryzują się dużym obciążeniem i małą wariancją. Obrazowym przykładem może być domyślna reguła klasyfikacji, która przypisuje obiekt do klasy z większym prawdopodobieństwem *a priori*, szacowanym na zbiorze uczącym. Reguła taka nie wykorzystuje w ogóle informacji niesionej przez zmienne objaśniające. Istotą skutecznego modelowania w regresji i dyskryminacji jest wybór modelu o odpowiedniej złożoności, pośredniego między dwiema wspomnianymi skrajnościami. W przypadku modeli liniowych złożoność jest najczęściej definiowana jako liczba parametrów modelu. Jeśli nie uwzględniamy wyrażeń interakcyjnych czy ogólnie dodatkowych zmiennych będących funkcjami zmiennych oryginalnych, to złożoność jest tożsama z liczbą zmiennych. W tym ujęciu selekcja zmiennych jest nie tylko zadaniem odpowiadającym na potrzeby interpretacyjne, ale też sposobem na konstrukcję modelu o jak najlepszych zdolnościach przewidywania wartości zmiennej objaśnianej dla obiektów, które pojawią się w przyszłości.

Celem artykułu jest zaproponowanie procedury doboru zmiennych do modelu regresji logistycznej. W tym celu wykorzystany będzie liniowy model prawdopodobieństwa (LMP) z regularyzacją, a do uzyskania większej stabilności wyników dodatkowo przeprowadzane zostanie sprawdzanie krzyżowe. Takie podejście zapewnia skuteczniejszą eliminację zmiennych nieistotnych od regularyzowanej regresji logistycznej, mimo że LMP nie jest atrakcyjną metodą dyskryminacji. Teza ta zostanie zweryfikowana empirycznie za pomocą symulacji.

2. Dyskryminacja jako przeformułowanie zadania regresji

Metody statystycznego uczenia z nauczycielem polegają na budowaniu modelu na podstawie zbioru uczącego:

$$\{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_N, y_N) : \boldsymbol{x}_i \in \boldsymbol{X} = (X_1, ..., X_p), y_i \in Y, i \in \{1, ..., N\}\}, \qquad (1)$$

gdzie Y (tzw. zmienna objaśniana) reprezentuje zjawisko, które chcemy wyjaśnić na podstawie obserwowanych cech $X_1, ..., X_p$. Jeśli Y jest zmienną ilościową, to mamy do czynienia z modelem regresji, w przypadku nominalnej zmiennej Y z modelem dyskryminacji. Model taki jest następnie wykorzystywany do przewidywania nieznanej wartości Y dla nowych obiektów, dla których zaobserwowano cechy $X_1, ..., X_p$. Ze względu na prostotę i możliwości interpretacji chyba najpopularniejszy jest liniowy model regresji wielorakiej:

$$y = b_0 + b_1 x_1 + ... + b_p x_p + \varepsilon \,. \tag{2}$$

Model taki łatwo przeformułować na zagadnienie dyskryminacji dwóch klas. Kategorie dychotomicznej zmiennej objaśnianej kodowane są wówczas wartościami {0; 1}. Interpretuje się je jako prawdopodobieństwa, że obiekt należy do klasy zakodowanej przez 1. Model taki znany jest z literatury jako liniowy model prawdopodobieństwa (LMP). Jego atrakcyjną cechą jest fakt, że estymacji parametrów strukturalnych można dokonać metodą najmniejszych kwadratów. Niedogodność takiego podejścia polega na tym, że wartości teoretyczne $\hat{y}_i$, które są oszacowaniami prawdopodobieństw *a posteriori*, mogą być ujemne lub większe od jedności. Wady tej nie ma model regresji logistycznej. Zakłada się tu, że logarytm ilorazu wiarygodności jest liniową funkcją zmiennych objaśniających:

$$\ln \frac{P(Y=1|\mathbf{x})}{1-P(Y=1|\mathbf{x})} = b_0 + b_1 x_1 + ... + b_p x_p \,. \tag{3}$$

Estymacji jego parametrów dokonuje się metodą największej wiarygodności, a logarytm funkcji wiarygodności można przedstawić następująco:

$$\ln L(\mathbf{b}) = \sum_{i=1}^{N} \left[y_i \cdot (b_0 + b_1 x_1 + ... + b_p x_p) - \ln(1 + \exp(b_0 + b_1 x_1 + ... + b_p x_p)) \right]. \tag{4}$$

gdzie y_i są zaobserwowanymi wartościami zmiennej Y (a więc zerami lub jedynkami). Ponieważ pochodne cząstkowe logarytmu wiarygodności są nieliniowymi funkcjami parametrów, do rozwiązania problemu estymacji stosuje się metody numeryczne. Najpopularniejszy jest algorytm Newtona-Raphsona, który można tu przeformułować na metodę najmniejszych kwadratów z iteracyjnie aktualizowanymi wagami (IRLS – *iteratively reweighted least squares*). Numeryczne rozwiązanie problemu estymacji jest dość kosztowne obliczeniowo dla dużej liczby zmiennych, co stanowi jeszcze jeden argument przemawiający za selekcją zmiennych.

3. Selekcja zmiennych przez regularyzację

Metody selekcji zmiennych dzieli się obecnie na trzy główne podejścia (zob. np. [Guyon i in. 2006]): dobór zmiennych na podstawie wybranego kryterium przed zastosowaniem algorytmu uczącego (*filters*), wyszukiwanie optymalnego podzbioru zmiennych sterowane oceną jakości modelu (*wrappers*) lub selekcja zmiennych wewnątrz algorytmu uczącego (*embedded methods*). Ogromną popularnością cieszy się trzecie z nich, a przykładami są regularyzowane wersje metod omówionych w poprzednim punkcie.

Główną ideą regularyzacji jest możliwość sterowania złożonością modelu. Uzyskuje się to poprzez nałożenie kary $P(\boldsymbol{b})$ za duże wartości bezwzględne parametrów w kryterium wykorzystywanym do estymacji:

(regresja liniowa lub LMP) $$\hat{\boldsymbol{b}} = \arg\min_{b} \left(\sum_{i=1}^{N} \left(y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij} \right)^2 + \lambda \cdot P(\boldsymbol{b}) \right), \tag{5}$$

(regresja logistyczna) $$\hat{\boldsymbol{b}} = \arg\min_{b} \left(dev + \lambda \cdot P(\boldsymbol{b}) \right), \tag{6}$$

gdzie: $dev = -2\ln L(\boldsymbol{b})$ jest tzw. odchyleniem modelu. Pierwszy składnik takiego kryterium odzwierciedla stopień dopasowania modelu do danych. W regresji liniowej (oraz LMP) jest to zwykle kwadratowa funkcja straty, a w regresji logistycznej odchylenie. Regularyzacja powoduje zmniejszanie wartości bezwzględnych parametrów, a czasem ich zerowanie, co jest równoznaczne z selekcją zmiennych i decyduje o atrakcyjności tych metod. Nawiązując do kompromisu obciążeniowo-wariancyjnego, regularyzacja daje możliwość uzyskania estymatorów o mniejszej wariancji, choć obciążonych. Różne metody regularyzacji różnią się przede wszystkim postacią komponentu kary. Były one pierwotnie proponowane dla regresji liniowej, ale mogą też być stosowane w regresji logistycznej. Historycznie pierwsza była regresja grzbietowa [Hoerl i Kennard 1970]:

$$P(\boldsymbol{b}) = \sum_{j=1}^{p} b_j^2. \tag{7}$$

Następnie Tibshirani [1996] zaproponował LASSO:

$$P(\boldsymbol{b}) = \sum_{j=1}^{p} |b_j|, \tag{8}$$

natomiast Zou i Hastie [2005] komponent kary będący ich kombinacją (*elastic net*):

$$P_\alpha(b) = \sum_{j=1}^{p} \left(\alpha b_j^2 + (1-\alpha)|b_j| \right). \tag{9}$$

Selekcja zmiennych (przez zerowanie niektórych współczynników) możliwa jest w przypadku, gdy komponent kary ma postać (8) lub (9). Parametr lambda decyduje o rozmiarze kary i w efekcie steruje złożonością modelu. Jego ustalenie jest główną trudnością stosowania modeli z regularyzacją. Zwykle w tym celu stosuje się ocenę błędu klasyfikacji przez sprawdzanie krzyżowe lub kryteria informacyjne. Studium porównawcze tych kryteriów dla przypadku regresji liniowej można znaleźć w pracy Kubusa [2011]. Zadanie estymacji parametrów strukturalnych (5) ma rozwiązanie w postaci zamkniętej jedynie w przypadku regresji grzbietowej. LASSO wymaga rozwiązania zadania programowania kwadratowego z liniowymi ograniczeniami, ale zwykle stosuje się metody przybliżone. Obecnie najbardziej popularny jest algorytm LARS Efrona i in. [2004], który cechuje mała złożoność obliczeniowa. Wykorzystuje się go też w implementacjach *elastic net*, gdyż można udowodnić,

że zadanie estymacji w tym przypadku da się przeformułować na zadanie LASSO. W regularyzowanej regresji logistycznej zadanie estymacji wymaga zastosowania metod numerycznych. W literaturze pojawiło się wiele propozycji rozwiązania tego problemu. Przegląd najważniejszych wyników i studium porównawcze można znaleźć np. w artykułach [Lee i in. 2006; Yuan i in. 2012]. Obecnie rekomendowany jako najszybszy jest algorytm *coordinate descent* Friedmana i in. [2010], który został zaimplementowany w pakiecie `glmnet` programu R.

4. Eksperyment

W artykule proponuje się wykorzystanie LMP z regularyzacją w postaci *elastic net* do wstępnej selekcji zmiennych. Na zredukowanym w ten sposób podzbiorze zmiennych objaśniających budowany będzie klasyczny model regresji logistycznej. Zaproponowane podejście oznaczane będzie w tabelach wynikowych symbolem EN+RL. Należy jeszcze podkreślić, że do selekcji zmiennych stosowano 10-częściowe sprawdzanie krzyżowe. Eliminowano zmienne, dla których mediana współczynników szacowanych w tej procedurze była równa 0. Do celów porównawczych wykorzystano: regresję logistyczną (RL) bez selekcji zmiennych, regresję logistyczną z regularyzacją LASSO (RL+L1) oraz LMP z regularyzacją *elastic net* (EN), który był stosowany zarówno do selekcji zmiennych, jak i budowy modelu.

Tabela 1. Wykorzystane zbiory danych

Zbiór	Liczba obserwacji	Liczba zmiennych	Liczba klas
Pima	768	8	2
ionosphere	351	33	2
sonar	208	60	2

Źródło: *UCI Repository of Machine Learning Databases*.

W przeprowadzonym eksperymencie wykorzystano trzy zbiory z repozytorium Uniwersytetu Kalifornijskiego [Frank, Asuncion 2010], których krótka charakterystyka znajduje się w tab. 1. Do oryginalnych zbiorów wprowadzano sztucznie generowane zmienne nieistotne na trzy sposoby:

1) 10 zmiennych z rozkładu zero-jedynkowego z jednakowymi frakcjami zer i jedynek,

2) 10 zmiennych z rozkładu $N(0; 1)$,

3) 10 zmiennych z rozkładu zero-jedynkowego z jednakowymi frakcjami zer i jedynek oraz 10 zmiennych z rozkładu $N(0; 1)$.

W ten sposób do badań uzyskano 9 zbiorów, które w tab. 2-4 oznaczone są według klucza ZBIÓR_NR, gdzie NR oznacza sposób generowania zmiennych nieistotnych.

Tabela 2. Mediany (w nawiasach) oraz średnie liczby wprowadzanych do modeli zmiennych nieistotnych (z błędami standardowymi) w procedurze sprawdzania krzyżowego

Zbiór	RL + L1	EN + RL
Pima 1	(0) 3,0 +/– 1,5	0
Pima 2	(0) 1,0 +/– 1,0	0
Pima 3	(0) 0,0 +/– 0,0	0
ionosphere 1	(8) 7,7 +/– 0,6	0
ionosphere 2	(5,5) 5,4 +/– 0,4	0
ionosphere 3	(13) 12,7 +/– 0,7	0
sonar 1	(6,5) 5,8 +/– 0,4	0
sonar 2	(3) 3,3 +/– 0,8	0
sonar 3	(1) 3,6 +/– 1,4	1

Źródło: obliczenia własne dla zbiorów z tab. 1 po wprowadzeniu zmiennych nieistotnych.

Tabela 3. Błędy klasyfikacji (w %) estymowane 10-częściowym sprawdzaniem krzyżowym (z błędami standardowymi)

Zbiór	RL	RL + L1	EN	EN + RL
Pima 1	23,9 +/– 2,7	24,2 +/– 1,8	24,3 +/– 1,4	25,1 +/– 1,6
Pima 2	25,9 +/– 1,7	23,7 +/– 1,5	23,4 +/– 2,2	24,8 +/– 2,1
Pima 3	25,0 +/– 2,1	23,7 +/– 1,3	27,9 +/– 1,5	24,6 +/– 1,5
ionosphere 1	15,1 +/– 2,3	11,7 +/– 2,4	13,7 +/– 2,7	12,0 +/– 1,9
ionosphere 2	13,4 +/– 1,7	14,0 +/– 1,1	14,5 +/– 1,3	11,3 +/– 2,1
ionosphere 3	12,2 +/– 1,9	12,0 +/– 3,0	14,2 +/– 1,9	11,4 +/– 1,7
sonar 1	29,8 +/– 1,8	25,9 +/– 3,4	25,9 +/– 1,8	22,5 +/– 2,0
sonar 2	29,3 +/– 1,6	25,4 +/– 1,8	26,0 +/– 1,7	25,0 +/– 2,3
sonar 3	31,7 +/– 2,5	26,0 +/– 4,2	25,5 +/– 5,3	21,1 +/– 2,6

Źródło: obliczenia własne dla zbiorów z tab. 1 po wprowadzeniu zmiennych nieistotnych.

W tab. 2 zestawiono liczby zmiennych nieistotnych wprowadzane do modeli w procedurze sprawdzania krzyżowego. Zaproponowane w artykule podejście okazało się niemal bezbłędne, podczas gdy regularyzowana regresja logistyczna wprowadzała nieraz dość znaczną liczbę zmiennych nieistotnych. Następnie uzyskane wyniki zweryfikowano oceną jakości modeli. W tym celu szacowano błędy klasyfikacji za pomocą 10-częściowego sprawdzania krzyżowego (zob. tab. 3). Zaproponowana metoda dawała na ogół nieco mniejsze błędy dla zbiorów z dość dużą liczbą zmiennych objaśniających (*ionosphere i sonar*), jednak różnice nie były sta-

tystycznie istotne, co zbadano testem Kruskala-Wallisa. Jeszcze jednym argumentem przemawiającym za proponowanym podejściem jest porównanie czasów pracy algorytmów (zob. tab. 4).

Tabela 4. Czas (w sekundach) pracy procedury sprawdzania krzyżowego (procesor 2,1 GHz oraz 4,0 GB RAM)

Zbiór	RL + L1	EN + RL
Pima 1	5,8	5,3
Pima 2	8,9	5,5
Pima 3	8,7	9,3
ionosphere 1	49,7	11,7
ionosphere 2	55,1	11,6
ionosphere 3	46,6	15,6
sonar 1	27,7	21,7
sonar 2	31,3	21,3
sonar 3	24,3	26,7

Źródło: obliczenia własne dla zbiorów z tab. 1 po wprowadzeniu zmiennych nieistotnych.

5. Podsumowanie

W artykule zaproponowano wykorzystanie liniowego modelu prawdopodobieństwa z regularyzacją jako metody doboru zmiennych do modelu regresji logistycznej. W etapie selekcji zmiennych dodatkowo zastosowano procedurę sprawdzania krzyżowego dla efektywniejszej eliminacji zmiennych nieistotnych. Przeprowadzone symulacje potwierdziły atrakcyjność takiego podejścia. Zaproponowana metoda identyfikuje zmienne nieistotne o wiele skuteczniej od powszechnie stosowanej regularyzowanej regresji logistycznej i jest na ogół szybsza, co może mieć znaczenie w analizie zbiorów z dużą liczbą zmiennych. Uzyskane modele charakteryzowały się też często nieco mniejszymi błędami klasyfikacji, lecz różnice nie były statystycznie istotne.

Literatura

Efron B., Hastie T., Johnstone I., Tibshirani R. (2004), *Least angle regression*, „Annals of Statistics" 32(2), s. 407-499.

Frank A., Asuncion A. (2010), *UCI Machine Learning Repository*, Irvine, CA: University of California, School of Information and Computer Science [http://archive.ics.uci.edu/ml].

Friedman J., Hastie T., Tibshirani R. (2010), *Regularization paths for generalized linear models via coordinate descent*, „Journal of Statistical Software", 33(1), s. 1-22.

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

Hoerl A.E., Kennard R. (1970), *Ridge regression: biased estimation for nonorthogonal problems*, „Technometrics" 12, s. 55-67.

Kubus M. (2011), *On model selection in some regularized linear regression methods*, XXX Konferencja Wielowymiarowa Analiza Statystyczna, Łódź (w druku).

Lee S., Lee H., Abbeel P., Ng A.Y. (2006), *Efficient L_1 regularized logistic regression*, In 21th National Conference on Artificial Intelligence (AAAI), s. 401-407.

Tibshirani R. (1996), *Regression shrinkage and selection via the lasso,* „J.Royal. Statist. Soc. B." 58, s. 267-288.

Yuan G., Ho C., Lin C. (2012), *An improved GLMNET for L1-regularized logistic regression*, „Journal of Machine Learning Research" 13, s.1999-2030.

Zou H., Hastie T. (2005), *Regularization and variable selection via the elastic net,* „Journal of the Royal Statistical Society" Series B. 67(2): s. 301-320.

REGULARIZED LINEAR PROBABILITY MODEL AS A FILTER

Summary: The application of regularized linear probability model as a filter which precedes the logistic regression is proposed in this paper. Additionally the cross-validation is applied in the feature selection stage. Such an approach guaranties more efficient elimination of the irrelevant variables than commonly used regularized logistic regression and classification errors of compared methods do not differ significantly. The datasets from UCI Repository were used in empirical study and noisy variables were generated from Bernoulli or normal distributions.

Keywords: feature selection, regularization, linear probability model.