PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 384

Taksonomia 24
Klasyfikacja i analiza danych – teoria i zastosowania

Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2015

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

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

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

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
e-ISSN 2392-0041 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Zamówienia na opublikowane prace należy składać na adres:
Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
tel./fax 71 36 80 602; e-mail:econbook@ue.wroc.pl
www.ksiegarnia.ue.wroc.pl

Druk i oprawa: TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 384 • 2015
Taksonomia 24 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania e-ISSN 2392-0041

Mariusz Kubus
Politechnika Opolska
e-mail: m.kubus@po.opole.pl

REKURENCYJNA ELIMINACJA CECH W METODACH DYSKRYMINACJI

Streszczenie: Jednym z podejść do selekcji zmiennych jest ocena ich podzbiorów za pomocą kryterium jakości modelu (*wrapper*). Jego zaleta to ścisły związek estymacji parametrów modelu z doborem zmiennych. Niestety, nawet zachłanne przeszukiwanie w selekcji krokowej jest kosztowne obliczeniowo w przypadku dużego wymiaru i złożonych obliczeniowo algorytmów uczących. Guyon i in. [2002] zaproponowali dla metody SVM algorytm rekurencyjnej eliminacji cech. Jego ideę można przenieść na inne metody statystycznego uczenia. W artykule zbadana będzie efektywność rekurencyjnej eliminacji cech stosowanej z różnymi metodami dyskryminacji. Pod uwagę wzięta będzie zdolność do identyfikacji zmiennych nieistotnych, dokładność klasyfikacji oraz czas pracy algorytmu. Badania przeprowadzone będą na zbiorach, do których sztucznie wprowadzano zmienne nieistotne.

Słowa kluczowe: selekcja zmiennych, rekurencyjna eliminacja cech, strategie przeszukiwania.

DOI: 10.15611/pn.2015.384.16

1. Wstęp

Jednym z kluczowych problemów w modelowaniu klasyfikatorów jest występowanie w zbiorach danych zmiennych nieistotnych, a więc takich, które nie mają wpływu na badane zjawisko reprezentowane przez zmienną objaśnianą. W wielu aplikacjach badacz nie ma dostatecznej wiedzy o tym, jakie czynniki wpływają na zmienną objaśnianą. Analizie poddawane są duże zbiory, by w sposób automatyczny i zalgorytmizowany wydobyć wiedzę z danych. Wydawać by się mogło, że im więcej informacji (zgromadzonych danych), tym lepiej. Jednak nie zawsze tak jest. Zmienne nieistotne mogą wywoływać zjawisko nadmiernego dopasowania modelu do danych ze zbioru uczącego, co w efekcie obniża zdolność generalizacji modelu, tzn. dokładność klasyfikacji nowych obiektów. Prawidłowość taka jest uzasadniana teoretycznie kompromisem obciążeniowo-wariancyjnym (zob. np. [Hastie i in. 2009]). Wraz ze złożonością modelu, która może być charakteryzowana liczbą zmiennych, maleje błąd resubstytucji, lecz rośnie błąd estymowany na zbiorze

danych, który nie brał udziału w etapie uczenia. Dodatkowe problemy pojawiają się w zadaniach wysokowymiarowych. Wraz ze wzrostem wymiaru przestrzeni cech wzrastają w tempie wykładniczym wymogi co do liczebności próby, by zachować dokładność estymacji – jest to tzw. przekleństwo wielowymiarowości (curse of dimensionality). Studium tego problemu przedstawili np. Blumer i in. [1987], Hastie i in. [2009, s.22-27] czy Langley i Sage [1997].

W celu redukcji zbędnej informacji powszechnie stosowane są metody selekcji zmiennych. Dzieli się je na trzy grupy [Blum i Langley 1997]: metody doboru zmiennych (*filters*), selekcję zmiennych przez selekcję modeli (*wrappers*) oraz metody selekcji, które są integralną częścią algorytmów uczących (*embedded methods*). Mocną stroną drugiego podejścia jest ścisły związek oceny zmiennych z oceną jakości modelu. Niestety, jest ono złożone obliczeniowo. Odpowiedzią na ten problem była propozycja rekurencyjnej eliminacji cech (*Recursive Feature Elimination – RFE*) zaproponowana przez Guyon i in. [2002] w powiązaniu z metodą SVM.

W artykule podjęte zostaną badania nad algorytmem RFE stosowanym z różnymi metodami dyskryminacji. Zwrócimy uwagę na zdolność identyfikacji zmiennych nieistotnych, dokładność klasyfikacji uzyskanych modeli oraz czas pracy algorytmów.

2. Selekcja zmiennych jako zadanie optymalizacji kombinatorycznej

Problem selekcji zmiennych można sformułować jako zadanie optymalizacji kombinatorycznej (zob. np. [Tsamardinos i Aliferis 2003]). Załóżmy, że dysponujemy próbą uczącą U:

$$U = \{(\boldsymbol{x_1}, y_1),...,(\boldsymbol{x_N}, y_N) : \boldsymbol{x_i} \in \boldsymbol{X} = (X_1,...,X_p), y_i \in Y, i \in \{1,...,N\}\} . \quad (1)$$

Przyjmijmy pewną rodzinę modeli F oraz kryterium jakości Q podzbioru zmiennych $S \subseteq \boldsymbol{X}$. W podejściu polegającym na selekcji modeli (*wrappers*) kryterium Q jest zarazem oceną jakości modelu $f_S \in F$ budowanego na podzbiorze S. Uniwersalnym kryterium jest błąd generalizacji, a więc w przypadku dyskryminacji błąd klasyfikacji dla nowych obiektów, spoza zbioru uczącego. Zwykle jest on estymowany metodami repróbkowania (np. przez sprawdzanie krzyżowe) lub z wykorzystaniem zbioru testowego. Wartość kryterium Q zależy od zbioru uczącego U, rodziny modeli F oraz podzbioru zmiennych S. Przy założeniu, że U się nie zmienia i że ograniczamy się do wybranej klasy modeli F, wartość kryterium zależy jedynie od S. Zadanie selekcji zmiennych można więc postawić jako problem znalezienia takiego podzbioru zmiennych objaśniających $S \subseteq \boldsymbol{X}$, dla którego funkcja $Q(S)$ przyjmuje wartość optymalną. Ponieważ zbiór wszystkich możliwych podzbiorów zmiennych jest skończony, mamy do czynienia z problem optymalizacji kombinatorycznej. Rozwiązanie polegające na obliczeniu wartości kryterium Q dla wszyst-

kich możliwych kombinacji zmiennych jest w większych wymiarach niepraktyczne. Po pierwsze, ze względu na złożoność obliczeniową problemu. Po drugie, ze względu na problem nadmiernego przeszukiwania (*oversearching*), które może prowadzić do nadmiernego dopasowania, a tym samym obniżenia zdolności generalizacji [Quinlan, Cameron-Jones 1995; Jensen, Cohen 2000]. Zatem jeśli przeszukiwanie wyczerpujące wszystkie kombinacje zmiennych jest niepraktyczne lub nawet niewskazane, to należy uznać, że kluczowym komponentem algorytmów selekcji zmiennych jest sposób przeszukiwania przestrzeni wszystkich podzbiorów zmiennych objaśniających.

Techniki przeszukiwania dzieli się na heurystyczne i stochastyczne. Ich zadaniem jest wskazanie, które podzbiory są warte brania pod uwagę, a więc niejako wytyczają ścieżkę w przestrzeni wszystkich kombinacji zmiennych. Techniki stochastyczne są bardziej kosztowne obliczeniowo i rzadziej stosowane w dużych wymiarach. Wśród heurystycznych najpopularniejsza jest strategia wspinaczki (*hill-climbing*), inaczej nazywana strategią zachłanną (*greedy search*). W zadaniu selekcji zmiennych strategię tę wykorzystuje selekcja krokowa nazywana też selekcją sekwencyjną, a w modelach liniowych regresją krokową. Inne strategie przeszukiwania są nieco bardziej kosztowne obliczeniowo. W przeszukiwaniu wiązką (*beam search*) prowadzi się kilka strategii wspinaczki równolegle. Z kolei w przeszukiwaniu typu najpierw najlepszy (*best-first*) stosowane są liczne powroty do poprzednich stanów.

Alternatywnym podejściem jest wykorzystanie rankingów zmiennych, co znacznie zawęża przestrzeń przeszukiwań. Najprostsze podejście polega na rangowaniu zmiennych, a następnie budowaniu modeli zagnieżdżonych, ich ocenie oraz wyborze optymalnego [Kubus 2013]. Ranking zmiennych można też przeprowadzać wielokrotnie w procedurze iteracyjnej (co z założenia ma stabilizować model) i takie podejście przyjęto w algorytmie rekurencyjnej eliminacji cech [Guyon i in. 2002].

3. Krokowa oraz rekurencyjna eliminacja cech

Ponieważ tytułowy algorytm rekurencyjnej eliminacji cech – RFE przypomina nieco wspomnianą już procedurę eliminacji krokowej, postaramy się tu wyeksponować różnicę, która skutkuje znacznym zmniejszeniem złożoności obliczeniowej na korzyść algorytmu RFE. W obu przypadkach punktem początkowym analizy jest pełny zestaw zmiennych objaśniających. W selekcji krokowej, w kolejnej iteracji buduje się i ocenia modele dla podzbiorów powstałych przez usunięcie jednej zmiennej, zatem p modeli na podzbiorach $p-1$ elementowych. Najlepszy model identyfikuje najlepszy podzbiór w tej iteracji. Dla tego podzbioru ponownie rozważa się wszystkie możliwe usunięcia pojedynczej zmiennej i procedura kontynuowana jest do wyczerpania zmiennych lub spełnienia kryterium stopu. W efekcie powstaje ciąg optymalnych w każdej iteracji podzbiorów, z których ostatecznie

wybierany jest najlepszy i przyjmowany jako rozwiązanie. Pomimo że eliminacja krokowa wykorzystuje strategię wspinaczki, złożoność obliczeniowa jest wciąż duża dla dużych p oraz kosztownych obliczeniowo algorytmów uczących. Takim jest niewątpliwie metoda SVM, która wymaga wewnętrznej walidacji swych parametrów.

W celu zmniejszenia kosztu obliczeń Guyon i in. [2002] zaproponowali algorytm RFE, w którym dokonuje się wielokrotnego rangowania zmiennych. Po zbudowaniu i ocenie modelu dla pełnego zestawu zmiennych budowany jest ranking zmiennych. W tym celu wykorzystywane są tylko informacje, jakie niesie model. Najgorsza ze zmiennych jest usuwana i budowany jest model dla kolejnego podzbioru ze zmniejszoną o jeden liczbą zmiennych. Takie podejście zapewnia, że w każdej iteracji wystarczy zbudować tylko jeden model, co znacznie redukuje złożoność obliczeniową. Podobnie jak w krokowej eliminacji procedura kontynuowana jest do wyczerpania zmiennych lub spełnienia kryterium stopu oraz efektem jest ciąg optymalnych w każdej iteracji podzbiorów, z których wybierany jest najlepszy. Rolę rankingu w liniowym SVM odgrywają oszacowane wagi. W przypadku SVM z funkcją jądrową kryterium rangujące zmienne jest związane z maksymalizacją marginesu i ma postać:

$$Q = \left| W^2(\boldsymbol{\alpha}) - W^2_{(-j)}(\boldsymbol{\alpha}) \right| \tag{2}$$

oraz

$$W^2_{(-j)}(\boldsymbol{\alpha}) = \sum \alpha_k \alpha_l y_k y_l K(\boldsymbol{x}_k^{-j}, \boldsymbol{x}_l^{-j}), \tag{3}$$

gdzie $\boldsymbol{x}_k^{-j}$ oznacza k-ty obiekt z usuniętą realizacją j-tej zmiennej. Wartości $W^2(\boldsymbol{\alpha})$ oblicza się analogicznie jak we wzorze (3), ale bez usuwania realizacji zmiennych.

Oryginalnie zaproponowany przez Guyon i in. [2002] algorytm RFE można stosować z wieloma metodami dyskryminacji (czy też regresji). Jest to możliwe, o ile na podstawie zbudowanego modelu jesteśmy w stanie zbudować ranking zmiennych. Pseudokod algorytmu RFE w ujęciu ogólnym został przedstawiony w tab. 1. Liczba zmiennych usuwanych w 4 kroku algorytmu może być ustalona (w małych zbiorach po prostu $k = 1$), określona regułą arytmetyczną (np. połowa zmiennych) lub regułą statystyczną (np. test istotności). Najprostszym kryterium rangującym zmienne na podstawie informacji dostarczonej przez model są wartości bezwzględne współczynników modelu liniowego. Problematyczne jednak mogą być duże wariancje oszacowanych współczynników. Spotkamy się z taką sytuacją np. w przypadku wystąpienia współliniowości. Zhu i Hastie [2004] powiązali algorytm RFE z logistyczną regresją grzbietową. Jak wiadomo, tak estymowane współczynniki mają mniejsze wariancje. W niniejszym artykule proponujemy, by wykorzystać regularyzowaną regresję logistyczną z komponentem kary w postaci elastycznej sieci [Zou, Hastie 2005]. Motywacją takiego wyboru były nie tylko

Tabela 1. Algorytm RFE – ujęcie ogólne

1.	Ustal wartości początkowe: aktualny podzbiór zmiennych $S = \boldsymbol{X}$, minimalną liczbę zmiennych $p_0 < p$, ciąg optymalnych w każdej iteracji podzbiorów $\boldsymbol{B} = (\mathbf{0},\ldots,\mathbf{0})$.
2.	Wykorzystując aktualny podzbiór S, zbuduj model f_S oraz ranking ważności zmiennych.
3.	Oceń model f_S i zapamiętaj w ciągu $\boldsymbol{B}$.
4.	Ze zbioru S usuń k najgorszych w rankingu zmiennych.
5.	Kroki 2-4 powtarzaj do momentu, gdy w S pozostanie p_0 zmiennych.
6.	Wybierz najlepszy spośród zapamiętanych w ciągu $\boldsymbol{B}$ modeli zagnieżdżonych.

Źródło: opracowanie własne na podstawie [Guyon i in. 2002].

mniejsze wariancje oszacowanych współczynników oraz szybkość algorytmu uczącego, ale też wewnętrzny mechanizm selekcji zmiennych, który skutkuje statystycznie uzasadnioną eliminacją większej liczby zmiennych w niektórych iteracjach. Na uwagę zasługuje fakt, że popularna wśród praktyków procedura usuwania w kolejnych iteracjach zmiennych, którym odpowiadają nieistotne statystycznie współczynniki modelu liniowego, wpisuje się w schemat algorytmu RFE. Jej zalety to możliwość usuwania więcej niż jednej zmiennej w każdej iteracji oraz naturalne kryterium stopu (istotność wszystkich współczynników). Granitto i in. [2006] stosowali rekurencyjną eliminację cech z metodą lasów losowych [Breiman 2001]. Ranking generowany przez las losowy powstaje przez agregację wartości miary jakości podziału dla każdej zmiennej we wszystkich modelach składowych.

4. Eksperyment

W badaniu efektywności algorytmu RFE brano pod uwagę trzy charakterystyki. Po pierwsze, skuteczność identyfikacji zmiennych nieistotnych. W tym celu do danych rzeczywistych sztucznie wprowadzano zmienne nieistotne, a więc nieróżniące się rozkładem w klasach. Ponieważ algorytm selekcji zmiennych może zbyt radykalnie usuwać zmienne – eliminując nie tylko zmienne nieistotne, ale też niektóre ważne dla dokładnej klasyfikacji – drugim składnikiem oceny jest błąd klasyfikacji uzyskanych modeli. Trzecią charakterystyką, odgrywającą nieraz w praktyce ważną rolę, jest czas pracy algorytmu. W badaniu wykorzystano zbiory danych z repozytorium Uniwersytetu Kalifornijskiego (http://archive.ics.uci.edu/ml/): *biodeg* (1055 obiektów, 41 zmiennych), *ionosphere* (351 obiektów, 33 zmienne) oraz *sonar* (208 obiektów, 60 zmiennych). Każdy z nich jest zadaniem klasyfikacji binarnej.

Do oryginalnych zbiorów dołączano zmienne nieistotne na dwa sposoby. W pierwszym przypadku generowano 10 zmiennych z rozkładu zero-jedynkowego z jednakową frakcją zer i jedynek. W drugim przypadku 10 zmiennych nieistotnych generowano z rozkładu normalnego. Realizacje pierwszych trzech losowano niezależnie z $N(0;1)$. Czwarta zmienna była kombinacją liniową pierwszych

trzech z dodanym szumem gaussowskim. Kolejne sześć zmiennych wygenerowano tak, by były parami skorelowane. Uzyskane w ten sposób zbiory oznaczono w tab. 2-4, dodając do nazwy symbole (10b) w pierwszym przypadku lub (10n) w drugim. Każdy zbiór był 50 razy dzielony losowo na próbę uczącą i próbę testową (1/3 oryginalnego zbioru uczącego), na której estymowano błąd klasyfikacji.

W badaniu porównawczym zastosowano algorytm RFE z różnymi metodami dyskryminacji. Poniżej przedstawiamy ich opis i symbole stosowane w tab. 2-4.

- Regresja logistyczna (RL): ranking ważności zmiennych wg wartości modułu statystyki *z*; w każdej iteracji usuwana jedna zmienna; wybór końcowego modelu na podstawie kryterium BIC.
- Regularyzowana regresja logistyczna z komponentem kary w postaci elastycznej sieci (RRL): ranking według wartości bezwzględnej oszacowanych współczynników. Usuwanie zmiennych odpowiadających zerowym współczynnikom lub jednej, najgorszej w rankingu. Wybór końcowego modelu na podstawie kryterium BIC.
- Liniowy model wektorów nośnych (SVM-l): ranking według modułu oszacowanych wag. Usuwanie połowy zmiennych w iteracji, jeśli ich liczba jest większa od 10, lub jednej zmiennej w przeciwnym razie. Parametr kosztu ustalany spośród cost=2^(-2:2) za pomocą 10-częściowego sprawdzania krzyżowego. Wybór końcowego modelu na podstawie błędu klasyfikacji szacowanego 10-częściowym sprawdzaniem krzyżowym z regułą jednego błędu standardowego.
- Metoda wektorów nośnych z jądrem radialnym (SVM-r): jak wyżej, ale ranking według formuły (2-3).
- Lasy losowe (RF): ranking generowany przez las losowy. W każdej iteracji usuwana jedna zmienna. Liczba drzew = 200. Wybór końcowego modelu na podstawie błędu klasyfikacji szacowanego 10-częściowym sprawdzaniem krzyżowym z regułą jednego błędu standardowego.

Dla porównania przedstawione będą też wyniki regresji logistycznej z eliminacją krokową (KE-RL). Tu także do wyboru modelu końcowego stosowano kryterium informacyjne BIC. Ponadto zastosowano kryterium stopu, którym był brak poprawy wartości BIC.

Tabela 2 przedstawia średnie liczby wprowadzanych do modeli zmiennych nieistotnych. Najskuteczniejsze okazały się tu dwie metody powiązane z algorytmem RFE: RF oraz RRL. Weryfikacja tych wyników za pomocą błędu klasyfikacji uzyskanych modeli nie potwierdza jednak wyższości tych metod (tab. 3). Wyraźnie najmniejsze błędy klasyfikacji uzyskano w metodach SVM-r oraz RF. Wprawdzie w zbiorze *biodeg* mniejszy błąd uzyskano metodą SVM-l, lecz różnica nie była statystycznie istotna. Modele regresji logistycznej okazały się nieadekwatne do badanych zbiorów. Jednak zwróćmy uwagę, że przy porównywalnych błędach klasyfikacji metoda RRL lepiej identyfikowała zmienne nieistotne.

Tabela 2. Średnie liczby wprowadzanych do modeli zmiennych nieistotnych w 50 eksperymentach

Zbiory	Rekurencyjna eliminacja cech – RFE					Selekcja krokowa
	RL	RRL	SVM-l	SVM-r	RF	KE-RL
biodeg (10b)	0,12	0,2	0	1,2	0	0,12
biodeg (10n)	0,06	0,08	1,06	0,9	0,62	0,06
ionosphere (10b)	4,38	0,32	1,06	1,76	0	4,36
ionosphere (10n)	4,04	0,04	0,28	4,22	0	3,94
sonar (10b)	1,74	0,02	1,38	0,64	0	1,58
sonar (10n)	1,98	0	0,54	0,5	0,08	2,02

Źródło: obliczenia własne.

Tabela 3. Średnie błędy klasyfikacji z błędami standardowymi (w %) estymowane na zbiorach testowych w 50 eksperymentach

Zbiory	Rekurencyjna eliminacja cech – RFE					Selekcja krokowa
	RL	RRL	SVM-l	SVM-r	RF	KE-RL
biodeg (10b)	16,2 (0,2)	17,5 (0,3)	14,9 (0,3)	15,5 (0,3)	15,2 (0,3)	16,6 (0,3)
biodeg (10n)	16,1 (0,3)	17,5 (0,2)	14,6 (0,3)	15,1 (0,3)	15,3 (0,2)	16,1 (0,3)
ionosphere (10b)	14,9 (0,4)	13,6 (0,4)	14,6 (0,3)	8,1 (0,3)	7,9 (0,3)	15,6 (0,5)
ionosphere (10n)	16,2 (0,5)	14,4 (0,4)	14,4 (0,4)	8,9 (0,3)	7,7 (0,4)	15,5 (0,4)
sonar (10b)	28,2 (0,8)	27,5 (0,8)	27,0 (0,7)	21,9 (0,7)	23,0 (0,9)	29,6 (0,7)
sonar (10n)	29,7 (0,8)	27,8 (0,8)	28,4 (0,7)	21,9 (0,8)	24,5 (0,8)	29,4 (0,7)

Źródło: obliczenia własne.

W tabeli 4 przedstawiono średnie czasy pracy algorytmów, tzn. czas potrzebny na selekcję zmiennych, budowę modelu oraz jego ocenę na zbiorze testowym. Niestety, dokładnie klasyfikujące modele SVM-r wymagają zdecydowanie dłuższego czasu obliczeń. Różnice stałyby się jeszcze wyraźniejsze, gdyby chcieć oszacować błąd przez 10-częściowe sprawdzanie krzyżowe. Zaznaczmy też, że badane zbiory nie miały dużej liczby zmiennych w porównaniu do niektórych zastosowań,

Tabela 4. Średnie czasy pracy algorytmów (procesor 2,1 GHz oraz 4,0 GB RAM)

Zbiory	Rekurencyjna eliminacja cech – RFE					Selekcja krokowa
	RL	RRL	SVM-l	SVM-r	RF	KE-RL
biodeg (10b)	3 s	23 s	4 min 56 s	26 min 29 s	3 min 46 s	1 min 32 s
biodeg (10n)	3 s	22 s	5 min	19 min 28 s	3 min 43 s	1 min 34 s
ionosphere (10b)	2 s	6 s	1 min 18 s	2 min 15 s	50 s	27 s
ionosphere (10n)	2 s	6 s	1 min 20 s	2 min 39 s	50 s	25 s
sonar (10b)	3 s	4 s	1 min 21 s	2 min 20 s	1 min 17 s	1 min 23 s
sonar (10n)	3 s	4 s	1 min 21 s	2 min 19 s	1 min 18 s	1 min 22 s

Źródło: obliczenia własne.

np. selekcja genów czy klasyfikacja tekstów, na co praktycy powinni zwrócić uwagę. W zadaniach wysokowymiarowych algorytm RFE-SVM-r mógłby stać się mało praktyczny. Z tej perspektywy konkurencyjne wydają się lasy losowe.

5. Podsumowanie

Algorytm RFE [Guyon i in. 2002] opracowano dla zmniejszenia złożoności obliczeniowej w porównaniu do popularnej eliminacji krokowej. Zaproponowany oryginalnie dla metody SVM może być stosowany z innymi metodami dyskryminacji. Ma to duże znaczenie, gdyż klasyfikatory RFE-SVM, choć cechują się wysoką zdolnością generalizacji, są wymagające co do czasu obliczeń i nie zawsze potrafią bezbłędnie identyfikować zmienne nieistotne. Przeprowadzone badanie empiryczne wskazuje na konkurencyjność kombinacji lasów losowych z algorytmem RFE. W badaniu okazała się ona szybsza, lepiej identyfikowała zmienne nieistotne i prowadziła do porównywalnych błędów klasyfikacji. W przypadku gdy model regresji logistycznej jest adekwatny do zadania dyskryminacji, godny uwagi jest algorytm RFE-RRL. Można też go potraktować jako metodę doboru zmiennych przed estymacją modelu (*filter*).

Literatura

Blum A.L., Langley P., 1997, *Selection of relevant features and examples in machine learning*, Artificial Intelligence, v. 97 n. 1-2, s. 245-271.

Blumer A., Ehrenfeucht A., Haussler D., Warmuth M.K., 1987, *Occam's razor*, Information Processing Letters, 24, s. 377-380.

Breiman L., 2001, *Random forests*, Machine Learning, 45, s. 5-32.

Granitto P.M., Furlanello C., Biasioli F., Gasperi F., 2006, *Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products,* Chemometrics and Intelligent Laboratory Systems, vol. 83, 2, s. 83-90.

Guyon I., Weston J., Barnhill S., Vapnik V., 2002, *Gene selection for cancer classification using support vector machines,* Machine Learning, 46: 389-422.

Hastie T., Tibshirani R., Friedman J., 2009, *The Elements of Statistical Learning: Data Mining, Inferance, and Prediction,* 2nd edition, Springer, New York.

Jensen D.D., Cohen P.R., 2000, *Multiple comparisons in induction algorithms*, Machine Learning, 38(3): s. 309-338.

Kubus M., 2013, *Some Remarks on Feature Ranking Based Wrappers*, [w:] Cz. Domański (red.) *Methods and Applications of Multivariate Statistical Analysis*, Acta Universitatis Lodziensis, Folia Oeconomica, 286, s. 147-154.

Langley P., Sage S., 1997, *Scaling to Domains with Many Irrelevant Features*, [in:] R. Greiner (ed.), Computational Learning Theory and Natural Learning Systems (vol. 4). Cambridge, MA: MIT Press.

Quinlan J.R., Cameron-Jones R.M., 1995, *Oversearching and layered search in empirical learning.* [in] Mellish C. (Ed.), Proceedings of the 14th International Joint Conference on Artificial Intelligence, Morgan Kaufman, s. 1019-1024.

Tsamardinos I., Aliferis C.F., 2003, *Towards principled feature selection: relevancy, filters and wrappers.* In Proceedings of the Workshop on Artificial Intelligence and Statistics.

Zhu J., Hastie T., 2004, *Classification of gene microarrays by penalized logistic regression*, Biostatistics, 5, 3, s. 427-443.

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.

RECURSIVE FEATURE ELIMINATION IN DISCRIMINATION METHODS

Summary: Recursive feature elimination (RFE) with SVM was proposed in [Guyon at al. 2002] to reduce computational cost in comparison to stepwise selection. The algorithm can be applied with a wider range of discrimination methods. This article shows a general scheme of RFE algorithm. The empirical comparison of RFE combined with a few discrimination methods is also given.

Keywords: feature selection, recursive feature elimination, heuristic search.