PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 7 (1207) 2008
TAKSONOMIA 15
Klasyfikacja i analiza danych – teoria i zastosowania

Michał Trzęsiok
Akademia Ekonomiczna w Katowicach

WYBÓR WARTOŚCI PARAMETRÓW PRZEZ WALIDACJĘ WYNIKÓW KLASYFIKACJI TAKSONOMICZNEJ METODY WEKTORÓW NOŚNYCH

1. Wstęp

Twórcą metody wektorów nośnych (SVM – *support vector machines*) jest Vladimir Vapnik, który pierwotnie zaproponował jej algorytm przeznaczony do rozwiązywania zadań dyskryminacji [Vapnik 1998]. W 2001 r. Ben-Hur, Horn, Siegelmann i Vapnik zaproponowali nową metodę taksonomiczną (SVC – *support vector clustering*) – wykorzystującą podejście analogiczne jak w dyskryminacyjnej metodzie wektorów nośnych. Metoda SVC jest metodą bardzo elastyczną – pozwala na generowanie skupień o nieliniowych, bardzo nieregularnych kształtach, lecz nie wymaga nakładania *a priori* założeń dotyczących liczby i kształtu skupień. Te dwie najważniejsze cechy wyniku grupowania metodą taksonomiczną, czyli liczba i kształty skupień, zależą przede wszystkim od dwóch parametrów metody SVC. Bardzo ważną kwestią jest więc odpowiedni dobór wartości parametrów metody SVC.

W niniejszej pracy przedstawiono propozycję wykorzystania jednej z metod walidacji wyników klasyfikacji – analizy replikacji – jako kryterium wyboru wartości kluczowych parametrów taksonomicznej metody wektorów nośnych.

2. Taksonomiczna metoda wektorów nośnych – krótki opis algorytmu

2.1. Etap pierwszy – wyznaczanie optymalnej hiperkuli zawierającej obrazy obserwacji ze zbioru danych

Szczegółowy opis algorytmu taksonomicznej metody wektorów nośnych można znaleźć w [Ben-Hur i in. 2001]. Pierwszy etap metody SVC jest bardzo podobny do dyskryminacyjnej metody SVM, która jest opisana w polskiej literaturze

m.in. w [Trzesiok 2004]. Z tego powodu w tym rozdziale przedstawiono jedynie główną ideę metody oraz najważniejsze elementy formalnego, analitycznego zapisu zadania optymalizacyjnego, przez który realizowana jest metoda. W punkcie 2.2 omówiony został bardziej szczegółowo drugi etap metody SVC, związany z identyfikowaniem przynależności obiektów do skupień.

Niech dany będzie zbiór uczący $D=\{\mathbf{x}^1, \ldots, \mathbf{x}^N\}$, gdzie $\mathbf{x}^i \in \mathbf{R}^d$, dla $i=1, \ldots, N$. Taksonomiczna metoda wektorów nośnych, podobnie jak jej dyskryminacyjny odpowiednik, w pierwszym kroku transformuje dane z przestrzeni pierwotnej w przestrzeń o znacznie większym wymiarze za pomocą nieliniowego przekształcenia $\varphi: \mathbf{R}^d \to \mathbf{Z}$. W nowej przestrzeni cech wyznaczana jest hiperkula o najmniejszym możliwym promieniu, zawierająca obrazy wszystkich obserwacji ze zbioru D. Zadanie wyznaczenia takiej kuli oznacza poszukiwanie jej środka oznaczonego przez $\mathbf{a}$ oraz promienia R, co można zapisać [Schölkopf, Smola 2002] w sposób analityczny jako zadanie optymalizacyjne o postaci:

$$\min_{R\in\mathbf{R},\mathbf{a}\in\mathbf{Z},\xi_i\geq 0} R^2 + \frac{1}{\nu N}\sum_{i=1}^{N}\xi_i, \tag{1}$$

z warunkami ograniczającymi:

$$\bigwedge_{i\in\{1,\ldots,N\}} \left\|\varphi(\mathbf{x}^i)-\mathbf{a}\right\|^2 \leq R^2+\xi_i, \quad \xi_i \geq 0, \tag{2}$$

gdzie: $\nu \in [0,1]$ – parametr mechanizmu regularyzacji, za pomocą którego użytkownik określa górną granicę frakcji obiektów ze zbioru D, które mogą zostać sklasyfikowane jako obserwacje nietypowe (ich obraz w przestrzeni cech może znajdować się poza wyznaczaną hiperkulą – dla tych obserwacji zachodzi: $\xi_i > 0$).

Rozwiązanie powyższego zadania optymalizacyjnego można wyznaczyć metodą mnożników Lagrange'a. Po przekształceniu funkcji Lagrange'a do postaci dualnej można zadanie (1)-(2) zapisać następująco:

$$\begin{aligned} &\min_{\alpha} \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j \mathrm{K}\left(\mathbf{x}^i,\mathbf{x}^j\right) - \sum_{i=1}^{N}\alpha_i \mathrm{K}\left(\mathbf{x}^i,\mathbf{x}^i\right) \\ &\text{dla } 0\leq\alpha_i\leq\frac{1}{\nu N}, \sum_{i=1}^{N}\alpha_i = 1, \end{aligned} \tag{3}$$

gdzie: α_i – współczynniki Lagrange'a,

$\mathrm{K}(\mathbf{u},\mathbf{v}) = \varphi(\mathbf{u})\cdot\varphi(\mathbf{v})$ – funkcja jądrowa, definiująca iloczyn skalarny w nowej przestrzeni cech.

Rozwiązanie zadania optymalizacyjnego (3) można przedstawić w postaci:

$$a = \sum_{i=1}^{N} \alpha_i \varphi(\mathbf{x}^i),$$
$$R^2 = \mathrm{K}(\mathbf{x}^s, \mathbf{x}^s) + \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j \mathrm{K}(\mathbf{x}^i, \mathbf{x}^j) - 2\sum_{i=1}^{N} \alpha_i \mathrm{K}(\mathbf{x}^i, \mathbf{x}^s), \tag{4}$$

gdzie: $\mathbf{x}^s$ – dowolny *wektor nośny*, czyli obserwacja, której w rozwiązaniu zadania (3) odpowiada niezerowy mnożnik Lagrange'a $(\alpha_s > 0)$.

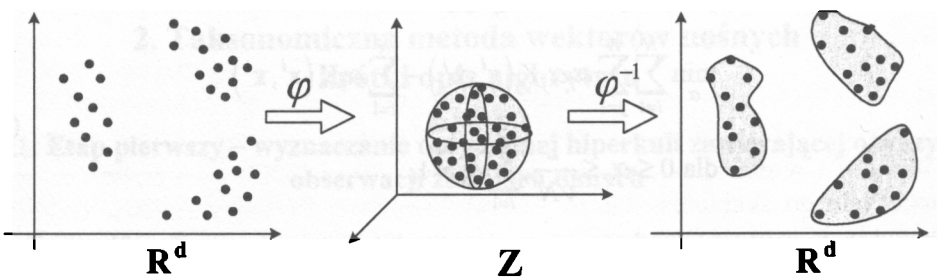
Wykorzystanie funkcji jądrowych jest charakterystyczną cechą metody wektorów nośnych. Pozwala ono na przeszukiwanie bardzo bogatej przestrzeni funkcji klasyfikujących, bez konieczności definiowania wprost nieliniowej transformacji φ.

Wyznaczony środek i promień kuli pozwalają na zdefiniowanie funkcji *f*. Zwraca ona wartość 1 dla obserwacji, których obraz w nowej przestrzeni cech znajduje się wewnątrz wyznaczonej hiperkuli, a wartość –1 w przeciwnym wypadku. Funkcję *f* można zapisać w następującej postaci:

$$f(\mathbf{x}) = \mathrm{sgn}\left(R^2 - \left(\mathrm{K}(\mathbf{x},\mathbf{x}) + \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j \mathrm{K}(\mathbf{x}^i,\mathbf{x}^j) - 2\sum_{i=1}^{N} \alpha_i \mathrm{K}(\mathbf{x}^i,\mathbf{x}) \right) \right). \tag{5}$$

Przeciwobraz zbioru punktów tworzących hipersferę w $\mathbf{Z}$ dla przekształcenia $\varphi: \mathbf{R}^d \rightarrow \mathbf{Z}$ tworzy w przestrzeni pierwotnej $\mathbf{R}^d$ rozłączne kontury – obwiednie danych. Te kontury mogą być zinterpretowane jako brzegi skupień. Punktom hiperkuli z przestrzeni $\mathbf{Z}$ odpowiadają jedynie punkty ograniczone konturami w przestrzeni pierwotnej.

Na rys. 1 przedstawiono ideę taksonomicznej metody wektorów nośnych. W skrócie można ją opisać jako procedurę, w której dane przekształcane są za pomocą nieliniowej transformacji w przestrzeń o większym wymiarze, gdzie wyznaczana jest optymalna hiperkula zawierająca ich obrazy. Hipersfera, poprzez transformację odwrotną, wyznacza w przestrzeni danych rozłączne kontury, które mogą być interpretowane jako brzegi skupień.

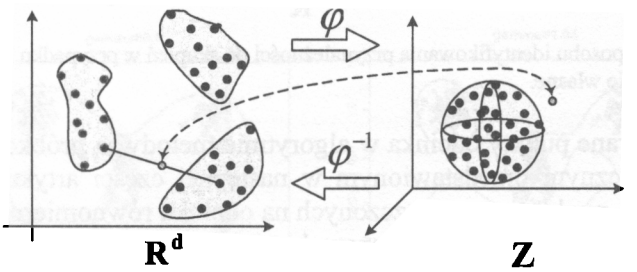


Rys. 1. Ilustracja idei taksonomicznej metody wektorów nośnych
Źródło: opracowanie własne.

Przedstawiony pierwszy etap metody wektorów nośnych daje opis konturów ograniczających zidentyfikowane skupienia w przestrzeni pierwotnej, lecz nie rozróżnia poszczególnych skupień między sobą. W etapie drugim wskazane zostanie, w jaki sposób algorytm może rozpoznawać, w którym skupieniu znajduje się dany obiekt.

2.2. Etap drugi – wskazanie przynależności do poszczególnych skupień

Można zauważyć, że gdy dane są dwa punkty należące do dwóch różnych skupień, to odcinek łączący te punkty zawiera też punkty „pośrednie", nienależące do żadnego skupienia (zob. rys. 2), czyli punkty, których obrazy w przestrzeni **Z** leżą poza wyznaczoną hiperkulą (funkcja f dana wzorem (5) zwraca dla tych punktów wartość –1).



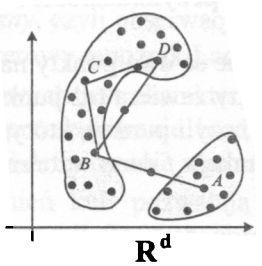
Rys. 2. Ilustracja sposobu identyfikowania przynależności do skupień (dana jest funkcja opisująca globalnie kontury wszystkich skupień jednocześnie)
Źródło: opracowanie własne.

Niech $\overline{\mathbf{x}^i\mathbf{x}^k}$ oznacza odcinek łączący punkty $\mathbf{x}^i, \mathbf{x}^k$. Zdefiniujmy *macierz połączeń* $A=[a_{ik}]_{N\times N}$, która będzie identyfikować, czy dana para punktów $\mathbf{x}^i, \mathbf{x}^k$ należy do tego samego skupienia:

$$a_{ik}=\begin{cases}0, & \text{gdy} \quad \bigvee\limits_{\mathbf{x}\in\overline{\mathbf{x}^i\mathbf{x}^k}} f(\mathbf{x})=-1,\\ 1, & \text{gdy} \quad \bigwedge\limits_{\mathbf{x}\in\overline{\mathbf{x}^i\mathbf{x}^k}} f(\mathbf{x})=1.\end{cases} \tag{6}$$

Wartość 1 w macierzy oznacza, że na pewno odpowiednie dwa punkty należą do tego samego skupienia, ale ze względu na możliwość wyznaczenia skupień niewypukłych, wartość 0 niekoniecznie oznacza, że punkty należą do różnych skupień. Przykład ilustrujący sposób identyfikowania przynależności do skupień w przypadku klas niewypukłych przedstawiono na rys. 3, na którym punkty B i D należą do tego samego skupienia, choć w macierzy połączeń odpowiadająca tej

parze wartość jest równa 0. Natomiast parze B, C oraz parze C, D odpowiada w macierzy połączeń wartość 1. Konieczne zatem jest przekształcenie macierzy połączeń tak, żeby reprezentowała relację przechodnią, a dokładniej – relację równoważności, tj. relację zwrotną, symetryczną i przechodnią. Dopiero po takim uzupełnieniu można z niej odczytać strukturę skupień. Innymi słowy, skupienia są wyznaczone przez połączone fragmenty grafu, generowanego przez macierz A.



Rys. 3. Ilustracja sposobu identyfikowania przynależności do skupień w przypadku klas niewypukłych
Źródło: opracowanie własne.

Klasyfikowane punkty odcinka w algorytmie metody są próbkowane. W przykładzie numerycznym przedstawionym w następnej części artykułu sprawdzano każdorazowo 11 punktów rozmieszczonych na odcinku równomiernie.

Do rozstrzygnięcia pozostaje jeszcze kwestia przynależności do skupień tych obiektów, które podczas budowy modelu SVC, w związku z ustaloną wartością parametru ν, zostały sklasyfikowane jako obserwacje oddalone. Obserwacje te, jako że ich obrazy nie należą do hiperkuli, nie zostaną przydzielone do żadnego skupienia. Jedną z możliwości jest potraktowanie każdej z obserwacji oddalonych jako jednoelementowego skupienia. Druga możliwość to wskazanie przynależności takiego obiektu do najbliższego skupienia (ta opcja została zastosowana w badaniu, którego wyniki przedstawiono w punkcie 4).

W znanych dostępnych pakietach statystycznych (SPSS, Statistica, SAS, a także w R) metoda wektorów nośnych jest oprogramowana jedynie w zakresie realizowania zadań dyskryminacji, regresji oraz wyznaczania uogólnionego kwantyla rozkładu. W przypadku tych zadań, w odróżnieniu od opisanej powyżej metody taksonomicznej, w nowej przestrzeni cech wyznaczana jest optymalna hiperpłaszczyzna, a nie hipersfera. Można jednak wykazać [Schölkopf, Smola 2002], że wybór funkcji jądrowej Gaussa:

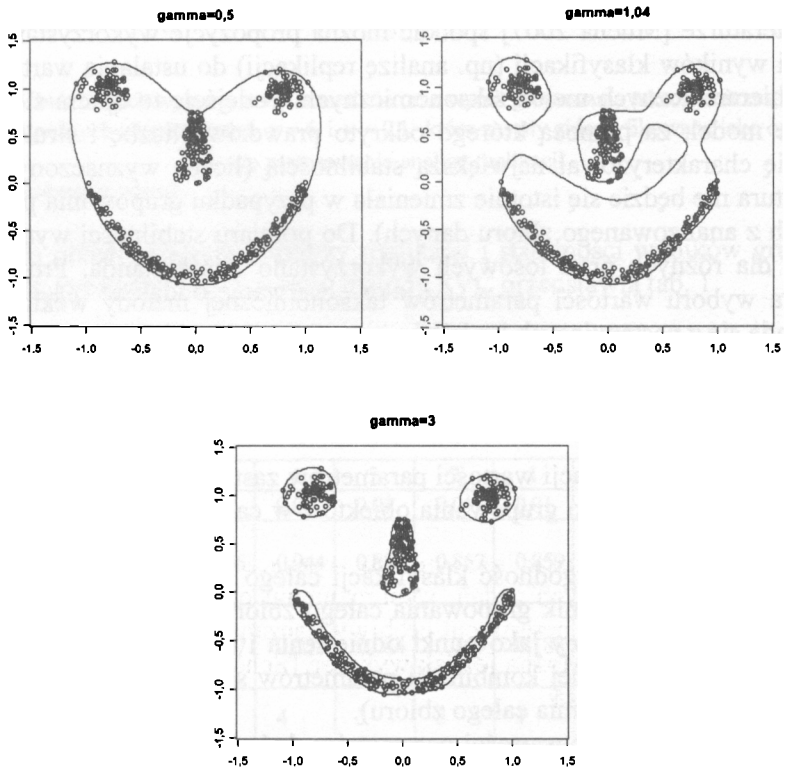
$$K(\mathbf{u},\mathbf{v}) = \exp(-\gamma\|\mathbf{u}-\mathbf{v}\|^2) \tag{7}$$

sprawia, iż problem znajdowania *hipersfery* zawierającej obraz zbioru danych jest równoważny problemowi wyznaczania *hiperpłaszczyzny* oddzielającej obraz danych od początku układu współrzędnych. Ta uwaga pozwoliła na wykorzystanie funkcji `svm(..., type="one-classification")` z biblioteki `e1071` programu R

do zrealizowania pierwszego etapu metody SVC. Drugi etap (identyfikacja przynależności obiektów do skupień z daną funkcją f definiującą ich kontury) został zrealizowany za pomocą autorskich funkcji napisanych w języku programu R.

3. Wrażliwości liczby otrzymywanych skupień oraz ich kształtu na wartości parametrów metody SVC

Oprócz parametru regularyzacji ν kluczowe znaczenie dla wyniku grupowania taksonomiczną metodą wektorów nośnych ma wybór parametru γ funkcji jądrowej Gaussa (7). Do zilustrowania znaczenia tego wyboru posłużono się zbiorem danych *Smiley*, wygenerowanym komputerowo za pomocą procedury zawartej w bibliotece `mlbench` pakietu statystycznego R (zob. rys. 4). Jest to jeden z ogólnodostępnych zbiorów zaprojektowanych specjalnie do sprawdzania własności metod wielowymiarowej analizy statystycznej.



Rys. 4. Wyniki grupowania na zbiorze danych *Smiley* taksonomiczną metodą wektorów nośnych dla różnych wartości parametru γ funkcji jądrowej Gaussa ($\gamma = 0{,}5$, $\gamma = 1{,}04$, $\gamma = 3$)

Źródło: opracowanie własne.

Z przedstawionego przykładu wynika, że metoda SVC jest bardzo elastyczna i pozwala na wyznaczenie skupień o bardzo różnorodnych kształtach (nieliniowych, niewypukłych). Bardzo cenną własnością metody SVC jest to, że jej stosowanie nie wymaga przyjmowania żadnych założeń dotyczących liczby skupień i ich kształtu, choć wiadomo, że charakterystyki te silnie zależą od wartości parametru γ funkcji jądrowej Gaussa (im większa jego wartość, tym większa liczba zidentyfikowanych skupień). Wynik grupowania zależy także od wartości parametru regularyzacji $\nu \in [0,1]$ metody SVC (parametr ten umożliwia rozpoznanie prawdziwej struktury klas w przypadku występowania obiektów nietypowych lub nierozłącznych skupień). Jak dotąd nie zaproponowano jednak efektywnej metody wyznaczania wartości tych parametrów.

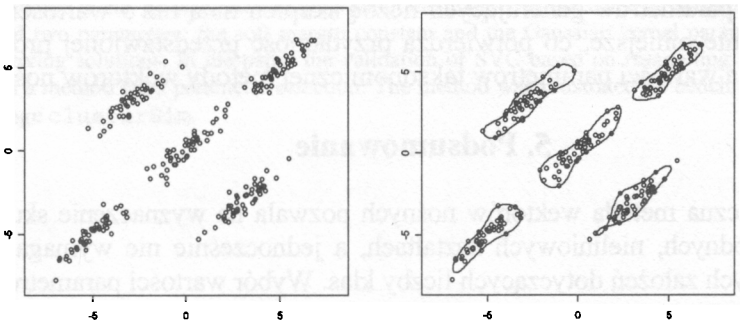
4. Propozycja ustalania wartości parametrów taksonomicznej metody wektorów nośnych przez analizę replikacji

W literaturze [Mucha 2007] spotkać można propozycje wykorzystania metod walidacji wyników klasyfikacji (np. analizę replikacji) do ustalania wartości parametrów hierarchicznych metod taksonomicznych. Podejście to opiera się na założeniu, że model, za pomocą którego odkryto prawdziwą liczbę i strukturę klas, będzie się charakteryzował największą stabilnością (liczba wyznaczonych klas i ich struktura nie będzie się istotnie zmieniała w przypadku grupowania prób wylosowanych z analizowanego zbioru danych). Do pomiaru stabilności wyników klasyfikacji dla różnych prób losowych wykorzystano miarę Randa. Proponowana procedura wyboru wartości parametrów taksonomicznej metody wektorów nośnych składa się z następujących kroków:

1. Wylosować z analizowanego zbioru danych B prób (tu $B = 20$).
2. Ustalić przeszukiwaną przestrzeń parametrów (różne kombinacje wartości γ i ν).
3. Dla każdej kombinacji wartości parametrów zastosować taksonomiczną metodę wektorów nośnych do grupowania obiektów w całym zbiorze i w wylosowanych próbach.
4. Porównać parami zgodność klasyfikacji całego zbioru i próbek, obliczając wartość miary Randa (wynik grupowania całego zbioru dla danej kombinacji parametrów jest tu traktowany jako punkt odniesienia i dlatego wyniki grupowania prób losowych dla tej samej kombinacji parametrów są zawsze porównywane jedynie z wynikiem grupowania całego zbioru).
5. Wybrać ten wariant wartości parametrów, który daje modele o najbardziej stabilnych wynikach klasyfikacji.

Do zilustrowania przedstawionej procedury wykorzystano zbiór danych zaprojektowany do porównywania własności metod statystycznej analizy wielowymiarowej i wygenerowany za pomocą funkcji `cluster.Gen(..., model=6)` za-

wartej w pakiecie `clusterSim` programu R. Ze względu na możliwość zilustrowania wyników grupowania wybrano zbiór z przestrzeni $\mathbf{R}^2$, w którym obserwacje skupione są w pięciu klasach liniowo nieseparowalnych (zob. rys. 5). W analizowanym zbiorze każda klasa zawierała 50 obserwacji.



Rys. 5. Zbiór danych wykorzystany w analizie oraz wynik grupowania taksonomiczną metodą wektorów nośnych z parametrami $\gamma = 6$ i $\nu = 0{,}1$, które zostały zidentyfikowane jako optymalne po zastosowaniu analizy replikacji

Źródło: opracowanie własne.

Wyniki przeprowadzonej analizy replikacji i stabilności wyników grupowania dla różnych kombinacji parametrów metody SVC przedstawia tab. 1.

Tabela 1. Uśrednione po wszystkich próbach losowych wartości miary Randa oraz współczynnik zmienności miary Randa dla różnych kombinacji wartości parametrów taksonomicznej metody wektorów nośnych (dodatkowo w każdym przypadku zamieszczono informację o liczbie skupień zidentyfikowanych w całym zbiorze danych)

Wartość γ	6	5	4	3	2	1	**6**	5	4	3
Wartość ν	0,01	0,01	0,01	0,01	0,01	0,01	**0,1**	0,1	0,1	0,1
Uśredniona wartość miary Randa	0,979	0,976	0,944	0,832	0,857	0,859	**0,980**	0,938	0,949	0,841
Współczynnik zmienności dla miary Randa	0,034	0,037	0,058	0,084	0,149	0,188	**0,035**	0,041	0,057	0,097
Zidentyfikowana liczba skupień	5	5	4	3	3	1	**5**	4	4	3

Źródło: opracowanie własne.

Przeprowadzona procedura wyznaczania wartości parametrów taksonomicznej metody wektorów nośnych oparta na analizie replikacji wskazała, że stabilność

wyników grupowania jest największa w modelu SVC z parametrami $\gamma = 6$ i $\nu = 0{,}1$. Należy zauważyć, że w tym przypadku wskazany model charakteryzuje się największą wartością uśrednionej miary Randa, a także współczynnik zmienności dla miary Randa jest najmniejszy, co dodatkowo wzmacnia wniosek o stabilności wyników grupowania dla tej kombinacji parametrów. Ponadto w tym przypadku metoda SVC poprawnie zidentyfikowała liczbę skupień równą 5 (zob. rys. 5). W kombinacji parametrów generujących liczbę skupień inną niż 5 wartości miary Randa są istotnie mniejsze, co potwierdza przydatność przedstawionej procedury do wyznaczania wartości parametrów taksonomicznej metody wektorów nośnych.

5. Podsumowanie

Taksonomiczna metoda wektorów nośnych pozwala na wyznaczenie skupień o bardzo różnorodnych, nieliniowych kształtach, a jednocześnie nie wymaga przyjmowania żadnych założeń dotyczących liczby klas. Wybór wartości parametrów γ i ν metody SVC ma kluczowe znaczenie dla kształtu i liczby wyznaczanych skupień. Wyniki analizy replikacji mogą stanowić przesłanki do wyboru wartości parametrów metody SVC. Porównując stabilność wyników grupowania prób losowych i całego zbioru danych dla różnych kombinacji parametrów metody, można wybrać ten układ parametrów, dla którego wartość miary Randa jest największa. Podstawową wadą tego podejścia jest to, że czas obliczeń potrzebnych do zrealizowania metody SVC bardzo szybko rośnie wraz ze wzrostem wymiaru kostki danych. Zastosowanie analizy replikacji (procedur próbkowania) dodatkowo znacznie zwiększa ten czas. Ta własność metody SVC istotnie ogranicza jej zakres zastosowań.

Literatura

Ben-Hur A., Horn D., Siegelman H. T., Vapnik V. (2001), *Support Vector Clustering*, „Journal of Machine Learning Research", 2, s. 125-137.

Mucha H.-J. (2007), *On Validation of Hierarchical Clustering*, [w:] *Advances in Data Analysis*, red. R. Decker, H.-J. Lenz, Springer Verlag, Berlin, s. 115-122.

Schölkopf B., Smola A.J. (2002), *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge.

Trzęsiok M. (2004), *Analiza wybranych własności metody dyskryminacji wykorzystującej wektory nośne*, [w:] *Postępy ekonometrii*, red. A.S. Barczak, AE, Katowice.

Vapnik V. (1998), *Statistical Learning Theory*, John Wiley & Sons, New York.

ON PARAMETER SELECTION AND VALIDATION OF SUPPORT VECTOR CLUSTERING

Summary

Support Vector Clustering (SVC) is a new method for unsupervised classification. The advantage of this method is that no prior assumptions about the number or the shape of clusters are required. The choice of two parameters: the soft margin constant and the Gaussian kernel parameter, is crucial to the clustering solutions. In the paper the validation of SVC based on resampling techniques is proposed as a method for a parameter selection. The method was illustrated on benchmark data set from `R` package `clusterSim`.