PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 86 — 2009
Wizualizacja wyników badań marketingowych – podejścia, metody i zastosowania

Joanna Trzęsiok, Michał Trzęsiok
Akademia Ekonomiczna w Katowicach

PROPOZYCJA METODY WIZUALIZACJI WYNIKÓW KLASYFIKACJI WSPOMAGAJĄCEJ PROFILOWANIE KLAS

1. Wstęp

Badania empiryczne wskazują, że wśród metod klasyfikacji największą dokładnością charakteryzują się te należące do eksploracyjnej analizy danych. Metody, takie jak: zagregowane drzewa klasyfikacyjne Breimana, metoda wektorów nośnych czy sieci neuronowe, pozwalają na zbudowanie modelu charakteryzującego się relatywnie niskimi błędami klasyfikacji [Trzęsiok 2006]. Często jednak modele te działają na zasadzie „czarnej skrzynki" i trudno jest interpretować otrzymane rezultaty. Ponadto przed wykorzystaniem uzyskanych wyników klasyfikacji niejednokrotnie wymagane jest przeprowadzenie merytorycznej ich weryfikacji. Właśnie słaba interpretowalność otrzymywanych modeli jest często największą wadą wymienionych nieparametrycznych metod dyskryminacji.

Obecnie w badaniach duży nacisk kładzie się nie tylko na poszukiwanie nowych, coraz dokładniejszych metod, ale również na zwiększenie możliwości pozyskiwania dodatkowej wiedzy o badanym zjawisku z modeli otrzymanych znanymi już metodami. Wpisując się w ten nurt, w artykule zaproponowano metodę wizualizacji wyników klasyfikacji wspomagającą profilowanie klas. Jest ona modyfikacją znanej metody doboru zmiennych do modelu przez budowanie rankingu z wykorzystaniem procedury eliminacji zmiennych redundantnych. W wyniku działania tej metody otrzymujemy informacje o tym, które ze zmiennych objaśniających mają największy wpływ na uzyskaną klasyfikację obiektów, a które zmienne można uznać za nieistotne. Ta dodatkowa wiedza jest szczególnie ważna dla decydentów i znacznie wspomaga proces podejmowania decyzji.

Zgodnie z klasyfikacją przedstawioną w pracy [Guyon i in. 2006] wyróżnić można trzy podejścia do problemu doboru zmiennych do modelu:

- filtrowanie zmiennych (*filters*), które obejmuje techniki doboru zmiennych niezależnie od zastosowanej metody klasyfikacji. Filtrowanie odbywa się na etapie przygotowania danych;
- symulacyjne przeszukiwanie podzbiorów zmiennych (*wrappers*) polegające na wielokrotnym wykorzystaniu metody klasyfikacji do oceny jakości modeli budowanych na różnych zestawach zmiennych. Do metod tych zaliczamy m.in. strategię wspinaczki, selekcję czy eliminację zmiennych;
- metody zagnieżdżone (*embedded methods*), czyli takie, w których kryterium doboru zmiennych jest integralną częścią algorytmu metody.

Metoda wizualizacji wyników klasyfikacji przedstawiona w artykule bazuje na procedurze składającej się z dwóch etapów.

1. W pierwszym etapie dla każdej klasy osobno wyznaczany jest ranking zmiennych ze względu na ich moc dyskryminującą – zdolność do odróżniania obiektów danej klasy od wszystkich innych obserwacji. Przy ocenie siły wpływu poszczególnych zmiennych objaśniających na wynik klasyfikacji wykorzystuje się jedną z metod symulacyjnego przeszukiwania podzbiorów zmiennych – metodę eliminacji.

2. W drugim etapie otrzymane informacje (ranking istotności zmiennych) są w prosty sposób kodowane i przedstawiane w sposób graficzny, umożliwiając badaczowi przeprowadzenie profilowania klas.

Zaprezentowane podejście w prosty i intuicyjny sposób łączy wyniki klasyfikacji obiektów z klasyfikacją zmiennych objaśniających.

2. Metoda doboru zmiennych do modelu, tworzenie rankingu zmiennych dla każdej klasy

W tej części pracy przedstawiono procedurę opartą na metodzie doboru zmiennych do modelu, pozwalającą na określenie siły wpływu zmiennych objaśniających na wynik klasyfikacji. Do budowy rankingu zmiennych wykorzystano procedurę ich eliminacji dla każdej klasy osobno, tj. budowano modele dyskryminacyjne, w których obserwacje z jednej ustalonej klasy oddzielano od wszystkich pozostałych obiektów traktowanych jako jedna klasa (*one against all*).

W metodzie eliminacji punktem wyjścia jest pełen zestaw zmiennych, z którego iteracyjnie usuwane zostają zmienne objaśniające – po jednej w każdym kroku iteracji aż do momentu, gdy zbiór zmiennych jest pusty. Usuwana jest zawsze ta zmienna, która w najmniejszym stopniu zmienia wartość przyjętego wcześniej kryterium.

W literaturze (zob. [Guyon i in. 2006; Rakotomamonjy 2003]) najczęściej jako kryterium wykorzystuje się minimalny błąd klasyfikacji modelu uwzględniający wszystkie możliwe obserwacje, również nowe, których przynależność do klas nie jest znana. Kryterium to ma jednak jedynie charakter teoretyczny, gdyż jego wartość nie jest znana. W praktyce estymuje się więc błąd klasyfikacji za pomocą metody sprawdzania krzyżowego. Tak obliczony błąd oznaczać będziemy przez $CVerr$. Alternatywnym podejściem jest przyjęcie jako kryterium miary zgodności klasyfika-

cji np. indeksu Randa. W podejściu tym klasyfikację obserwacji ze zbioru uczącego, otrzymaną na podstawie modelu zbudowanego na pełnym zestawie zmiennych, traktować będziemy jako wzorzec. W kolejnych krokach algorytmu budowane będą modele ze zredukowaną liczbą zmiennych objaśniających, które zostaną porównane ze wzorcem. Za każdym razem ze zbioru tymczasowo wyłączamy jedną zmienną (kolejno), ale ostatecznie, w danym etapie, wyeliminowana zostaje ta, która ma najmniejszy wpływ na wyniki klasyfikacji. Jest to zmienna odpowiadająca modelowi, którego zgodność klasyfikacji ze wzorcem będzie największa. Maksymalna wartość indeksu Randa wskazuje więc właśnie tę zmienną, której usunięcie w stopniu najmniejszym z możliwych zmienia klasyfikację.

Procedurę eliminacji zmiennych przeprowadzamy po kolei dla każdej klasy i w ten sposób otrzymujemy informację, które zmienne mają kluczowe znaczenie dla kształtu poszczególnych klas.

Przedstawiony algorytm doboru zmiennych do modelu przedstawić można w następujących krokach:

1. Zbuduj model dyskryminacyjny f_0, wykorzystując kompletny zbiór zmiennych objaśniających $V_0 = \{X_1, X_2, \ldots, X_m\}$, dobierając optymalne wartości parametrów metody;

2. Dla każdej klasy $k = 1, \ldots, K$, traktując obiekty nienależące do klasy k jako jedną klasę, wykonaj kroki:

Dla $j = 1, \ldots, m-1$ wykonaj polecenia:

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

b. Dla wszystkich zbudowanych w poprzednim kroku modeli porównaj zgodność otrzymanych wyników klasyfikacji ze wzorcem f_0, obliczając w tym celu indeks Randa.

c. Ostatecznie w kroku j usuń tę zmienną, dla której wartość indeksu Randa jest największa (jej usunięcie w najmniejszym stopniu zmienia wynik klasyfikacji). Zredukowany zbiór zmiennych oznacz przez V_j.

d. Przyjmij jako model f_j ten, który zbudowany był na zredukowanym zbiorze zmiennych oznaczonym przez V_j. Dla modelu f_j oblicz błąd klasyfikacji $CVerr_j$ metodą sprawdzania krzyżowego z podziałem zbioru danych na pięć części.

Procedurę powtarzamy tak długo, aż w zbiorze zmiennych pozostanie tylko jeden element. Jest to właśnie ta zmienna, która ma największy wpływ na kształt danej klasy. Rezultatem powyższego algorytmu są rankingi zmiennych dla każdej klasy z osobna.

Oprócz uzyskania rankingu dla każdej klasy, możliwe jest również zidentyfikowanie zbioru zmiennych nieistotnych. Wiadomo, że najmniej istotna zmienna została usunięta w pierwszym kroku budowy rankingu. Zachodzi jednak pytanie, ile z tych usuniętych zmiennych można uznać za redundantne.

Mając dany, dla każdej klasy, ciąg $\{CVerr_j\}_{j=1,\ldots,m-1}$ wartości błędu klasyfikacji (obliczonego metodą sprawdzania krzyżowego), wybieramy model odpowiadający naj-

mniejszej wartości $CVerr_j$. Zmienne, które nie są uwzględnione w tym modelu, można uznać za nieistotne. Jednak pomiar błędu klasyfikacji metodą sprawdzania krzyżowego jest obarczony błędem estymacji (tzw. błędem standardowym pomiaru):

$$SE_j = \frac{s_j}{\sqrt{T}},$$

gdzie T jest liczbą części, na które był dzielony zbiór uczący przy stosowaniu metody sprawdzania krzyżowego, s_j zaś to odchylenie standardowe błędu klasyfikacji obliczanego dla różnych części walidacyjnych wyróżnionych ze zbioru uczącego. W swojej pracy Hastie, Tibshirani i Friedman [2001] zaproponowali uwzględnienie faktu, że błąd $CVerr_j$ obarczony jest standardowym błędem pomiaru przez wybór właśnie tego modelu, dla którego błąd jest nie większy niż:

$$\min(CVerr_j) + SE_j.$$

Wszystkie zmienne nieuwzględnione w tym modelu uważamy za redundantne w opisie danej klasy.

Metoda eliminacji oparta jest na strategii wspinaczki i prowadzi do uzyskania rozwiązania optymalnego jedynie w sensie lokalnym. Unikamy jednak przeszukiwania wszystkich możliwych podzbiorów zmiennych, co znacznie skraca czas wykonywania algorytmu. Zaletą tego podejścia jest również stosunkowo niska złożoność obliczeniowa.

3. Przykład ilustrujący działanie procedury

Działanie przedstawionej procedury zilustrowano na przykładzie zbioru danych *Glass*. Jest to zbiór standardowo wykorzystywany do badania własności metod wielowymiarowej analizy statystycznej. Jest on przykładem zastosowania zagadnień dyskryminacji w kryminalistyce, przedstawia bowiem klasyfikację odłamków szkła znalezionych w miejscu popełnienia przestępstwa. Fragment szkła może zostać wykorzystany jako dowód w sprawie pod warunkiem jego poprawnej klasyfikacji.

Zbiór *Glass* zawiera 214 obserwacji – fragmentów szkła, charakteryzowanych przez 9 zmiennych objaśniających, którymi są współczynnik załamania światła (Z.1) oraz procentowa zawartość tlenków różnych metali (Z.2-Z.9). Obiekty w tym zbiorze reprezentują 6 klas.

Wszystkie obliczenia oraz wykresy wykonano w języku programu **R** z wykorzystaniem autorskich procedur.

3.1. Etap I – tworzenie rankingu zmiennych dla każdej klasy z osobna

W pierwszym etapie przeprowadzona została procedura eliminacji zmiennych i stworzono rankingi zmiennych objaśniających dla każdej klasy z osobna. Przedstawiony w poprzednim punkcie algorytm tworzenia rankingu zmiennych jest proce-

durą uniwersalną, która może być wykorzystana dla dowolnej metody dyskryminacji. W pracy zastosowano go dla czterech wybranych metod analizy danych [Walesiak, Gatnar 2009; Hastie, Tibshirani, Friedman 2001]:

- metody wektorów nośnych SVM (S*upport Vector Machines*),
- zagregowanych drzew klasyfikacyjnych Breimana (*random forest*),
- sieci neuronowych (*neural network*),
- metody *k* najbliższych sąsiadów (*k Nearest Neighbours*).

Trzy pierwsze metody możemy zaliczyć do grupy nieparametrycznych metod dyskryminacji, działających na zasadzie „czarnej skrzynki". Stworzenie rankingu możliwe jest jednak również dla klasycznej metody *k* najbliższych sąsiadów.

Wyniki działania zaproponowanej procedury w postaci graficznej zostały przedstawione dla wszystkich metod dyskryminacji w dalszej części pracy na rys. 1-4. Wyniki pośrednie w postaci tabelarycznej przedstawiono w tab. 1 wyłącznie dla jednej z klas w przypadku zastosowania metody wektorów nośnych (w celu zilustrowania pierwszego etapu procedury). W wyniku działania procedury otrzymano wiele analogicznych tabel – dla każdej metody dyskryminacji tyle tabel, ile jest klas w zbiorze danych.

Tabela 1. Wynik działania pierwszego etapu procedury – ocena mocy dyskryminacyjnej zmiennych w modelu SVM dla klasy pierwszej

Numer iteracji	Usunięta zmienna	Indeks Randa	Błąd klasyfikacji *CVerr*	Błąd standardowy
1	Z.5	0,936	0,229	0,034
2	Z.8	0,919	0,257	0,032
3	Z.3	0,910	**0,197**	0,030
4	Z.2	0,869	***0,224***	0,009
5	Z.1	0,853	0,243	0,036
6	Z.9	0,837	0,234	0,016
7	Z.6	0,744	0,239	0,029
8	Z.7	0,568	0,360	0,043
9	Z.4			

Źródło: opracowanie własne.

Z tabeli 1 można odczytać, że w pierwszej iteracji zidentyfikowano zmienną Z.5, której usunięcie w najmniejszym stopniu zmieniało wyniki klasyfikacji (które w 93,6% były zgodne z klasyfikacją otrzymaną na komplecie zmiennych). Oznacza to, że zmienna Z.5 ma najmniejszą moc dyskryminacyjną dla klasy pierwszej. W drugiej iteracji z pozostałego zestawu zmiennych wyeliminowano zmienną Z.8. Jako ostatnia w zbiorze zmiennych opisujących klasę pierwszą pozostała najistotniejsza zmienna – Z.4. Chcąc więc uzyskać ranking zmiennych ze względu na ich moc dyskryminacyjną, w opisie klasy pierwszej należy przeczytać drugą kolumnę tab. 1 od końca.

Jak to zostało przedstawione w punkcie 2 artykułu, dodatkowo w celu wskazania w rankingu miejsca podziału zmiennych na istotne i redundantne każdorazowo po usunięciu kolejnej zmiennej obliczano błąd klasyfikacji metodą sprawdzania krzyżowego (*CVerr*) wraz z odpowiadającym mu błędem standardowym pomiaru. Najmniejszy błąd *CVerr* (zaznaczony w tab. 1 pogrubioną czcionką) uzyskano w trzeciej iteracji. Uwzględnienie (dodanie) błędu standardowego daje wartość $\min(CVerr_j) + SE_j$. Błąd klasyfikacji modelu zbudowanego w czwartej iteracji nie przekracza tej wartości, dlatego ten model zostaje uznany za najlepiej opisujący wybraną klasę. Zatem nieuwzględnione w nim zmienne Z.5, Z.8, Z.3 oraz Z.2 należy uznać za nieistotne.

W tabeli 2 przedstawiono rankingi zmiennych (odwrócona kolejność w stosunku do drugiej kolumny tab. 1) dla każdej klasy zbioru *Glass* osobno w przypadku zastosowania metody wektorów nośnych. Pogrubioną czcionką zaznaczono zmienne istotne.

Tabela 2. Ranking zmiennych w modelu SVM dla każdej klasy zbioru danych *Glass* osobno

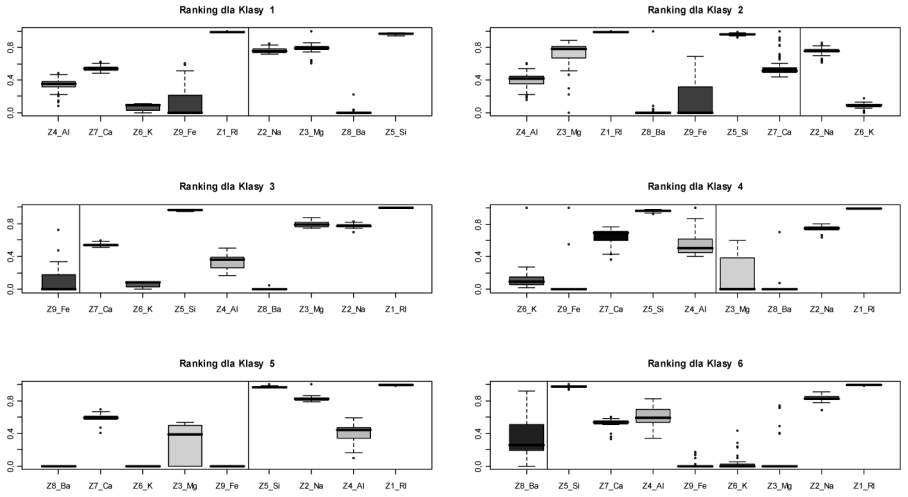
Pozycja w rankingu	Klasa 1 Szyba okienna typu *float*	Klasa 2 Szyba okienna zwyczajna	Klasa 3 Szyba samochodowa typu *float*	Klasa 4 Opakowanie szklane	Klasa 5 Zastawa stołowa	Klasa 6 Reflektor
1	**Z.4**	**Z.4**	**Z.9**	**Z.4**	**Z.7**	**Z.8**
2	**Z.7**	**Z.3**	Z.7	**Z.7**	**Z.5**	Z.7
3	**Z.6**	**Z.1**	Z.6	**Z.9**	**Z.4**	Z.5
4	**Z.9**	Z.5	Z.5	Z.5	**Z.2**	Z.4
5	**Z.1**	Z.9	Z.4	Z.8	**Z.3**	Z.9
6	Z.2	Z.8	Z.8	Z.6	**Z.9**	Z.6
7	Z.3	Z.7	Z.3	Z.3	**Z.6**	Z.3
8	Z.8	Z.2	Z.2	Z.2	**Z.8**	Z.2
9	Z.5	Z.6	Z.1	Z.1	Z.1	Z.1

Źródło: opracowanie własne.

Wyniki w tab. 2 w znacznym stopniu wspomagają profilowanie każdej z klas. W przedstawionym przykładzie widać m.in., że do opisu klasy 3 wystarczy jedna zmienna – Z.9. Oznacza to, że do identyfikacji szyby samochodowej wystarczy posłużyć się jedną cechą, która jest dla tej klasy obiektów bardzo charakterystyczna. Z tabeli 2 nie można jednak odczytać zakresu wybranych istotnych zmiennych w ramach różnych klas w celu porównania. Możliwość taką uzyskuje się po przedstawieniu znormalizowanych wartości zmiennych objaśniających na wykresach pudełkowych dla każdej klasy osobno.

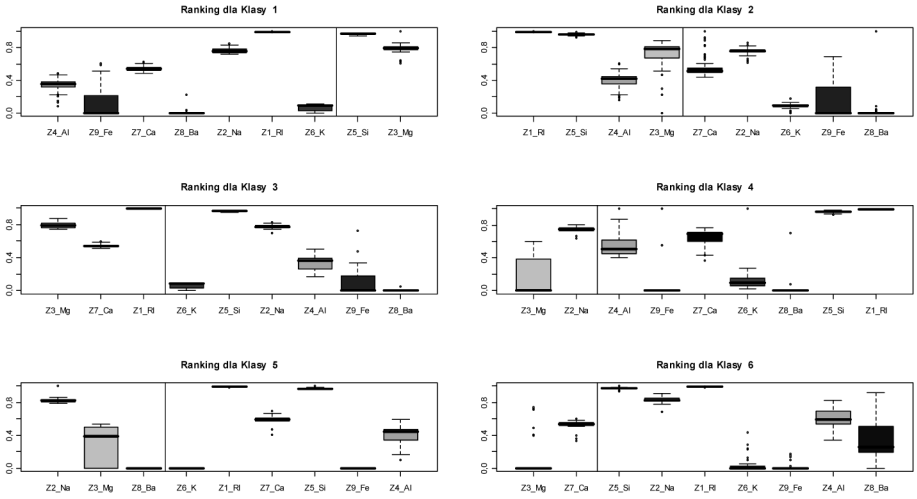
3.2. Etap II – graficzna prezentacja wyników

W drugim etapie rankingi zmiennych objaśniających, otrzymane dla każdej metody klasyfikacji, zostały w prosty sposób zakodowane i przedstawione w sposób graficzny. Unormowanie zmiennych (podzielenie przez największą pod względem



Rys. 1. Graficzna prezentacja wyników dla metody SVM

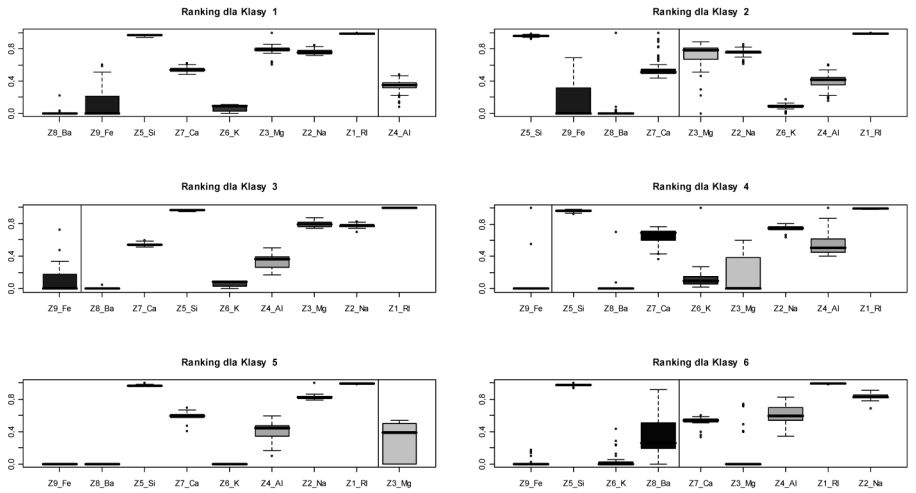
Źródło: opracowanie własne.



Rys. 2. Graficzna prezentacja wyników dla metody Random Forest

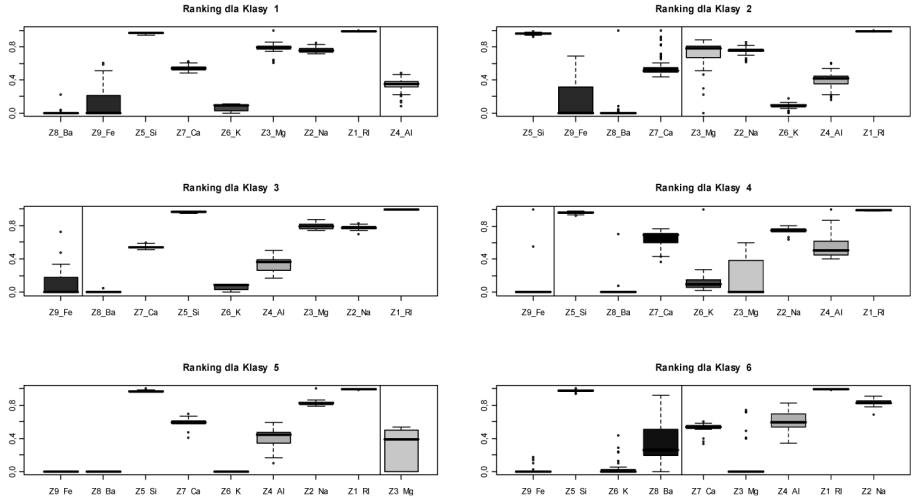
Źródło: opracowanie własne.

wartości bezwzględnej wartość zmiennej w obrębie danej klasy) umożliwia przedstawienie na jednym wykresie wykresów pudełkowych dla różnych zmiennych, a także pozwala na sprawne porównywanie, w jakim obszarze zakresu zmienności znajduje się wybrana zmienna w różnych klasach. Na rysunkach 1-4 w sposób graficzny przedstawiono wyniki działania zaproponowanej procedury. Wykresy pudeł-



Rys. 3. Graficzna prezentacja wyników dla sieci neuronowych

Źródło: opracowanie własne.



Rys. 4. Graficzna prezentacja wyników dla metody *k* najbliższych sąsiadów

Źródło: opracowanie własne.

kowe dają obraz zakresu zmienności analizowanych zmiennych objaśniających ustawionych w kolejności uzyskanej w rankingu (od lewej zmienne najistotniejsze), dla każdej klasy z osobna. Pionowa linia rozdziela zmienne istotne dla opisu danej klasy od redundantnych.

Każdy z rysunków jest syntetycznym sposobem prezentacji profilu pojedynczej klasy. Poszczególne zmienne mogą mieć przyporządkowane na stałe kolory wykresów pudełkowych dla lepszej i szybszej orientacji.

Rozważając np. rys. 1, stwierdzamy, że (dla metody SVM) – oprócz informacji, które opisane już zostały przy omawianiu tab. 1 i 2 – z rysunku tego możemy odczytać, że zmienna Z.3 reprezentująca zawartość tlenków magnezu jest istotną zmienną w opisie klas 2 i 5 (w dyskryminacji pozostałych klas nie jest istotna). Natomiast położenie pudełka wskazuje na to, że klasa 2 charakteryzuje się wysokimi wartościami tej zmiennej, klasa 5 zaś raczej niskimi. Jest to jedynie bardzo mały fragment opisu, który tworzony jest na etapie profilowania klas. Przedstawiony jednak został wyłącznie w celu zilustrowania możliwości interpretacyjnych oferowanych przez procedurę.

Porównanie opisów klas dla różnych metod dyskryminacji (porównanie rysunków 1, 2, 3 i 4 ze sobą) wyraźnie pokazuje, że mechanizmy identyfikowania i rozdzielania klas w tych metodach bardzo się różnią (inne cechy zostaną zidentyfikowane jako charakterystyczne i przez to istotne dla danej klasy w metodzie *k–NN*, a inne w metodzie *Random Forest* lub pozostałych).

4. Podsumowanie

W artykule przedstawiono propozycję metody wizualizacji wyników klasyfikacji wspomagającej profilowanie klas. Procedura ta pozwala na uzyskanie dodatkowej wiedzy o modelach dyskryminacyjnych otrzymanych metodami eksploracyjnej analizy danych, które często działają na zasadzie „czarnej skrzynki". Modele zbudowane za pomocą metod nieparametrycznych dają wyniki klasyfikacji charakteryzujące się wysoką dokładnością, jednak bez użycia dodatkowych narzędzi otrzymane wyniki trudno jest interpretować.

Zaproponowana metoda wizualizacji wyników klasyfikacji pozwala ocenić moc dyskryminującą zmiennych objaśniających. W wyniku jej działania otrzymujemy łatwy w interpretacji ranking zmiennych dla każdej klasy z osobna, przez co może być ona skutecznie wykorzystywana do profilowania klas. Ponadto, oprócz możliwości tworzenia rankingu zmiennych, metoda pozwala na niearbitralny podział zmiennych na istotne i redundantne oraz na graficzną prezentację wyników.

Literatura

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

Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning*, Springer Verlag, N.Y 2001.

Rakotomamonjy A., *Variable Selection Using SVM–based Criteria*, „Journal of Machine Learning Research" 2003 nr 3, s. 1357-1370.

Trzęsiok M., *Metoda wektorów nośnych na tle innych metod wielowymiarowej analizy danych*, [w:] Taksonomia 13, *Klasyfikacja i analiza danych – teoria i zastosowania*, K. Jajuga, M. Walesiak (red.), AE, Wrocław 2006, s. 536-542.

Walesiak M., Gatnar E. (red.), *Statystyczna analiza danych z wykorzystaniem programu R*, PWN, Warszawa 2009.

THE PROPOSAL OF VISUALIZATION OF CLASSIFICATION RESULTS SUPPORTING CLASS DESCRIPTION

Summary

After building the classification model, at the stage of the class description we try to extract knowledge from the model. We search for the description of classification rules, the natural language. The paper presents the simple algorithm for building the ranking of predictor variables based on their descriptive power (for every class separately) and uses boxplots to enable interpretation and give some insight.

The procedure is universal and can be applied to classic or data mining methods. SVMs, Random Forest, Neural Network and *k*-Nearest Neighbours were used for illustration with **R** software.