PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 427

Taksonomia 27
Klasyfikacja i analiza danych – teoria i zastosowania

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2016

Redaktor Wydawnictwa: Agnieszka Flasińska

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 stronach internetowych
www.pracenaukowe.ue.wroc.pl
www.wydawnictwo.ue.wroc.pl

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

Wersja pierwotna: publikacja drukowana

Zamówienia na opublikowane prace należy składać na adres:
Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
ul. Komandorska 118/120, 53-345 Wrocław
tel./fax 71 36 80 602; e-mail:econbook@ue.wroc.pl
www.ksiegarnia.ue.wroc.pl

Druk i oprawa: TOTEM

Spis treści

Wstęp

W dniach 14–16 września 2015 r. w Hotelu Novotel Gdańsk Marina w Gdańsku odbyła się XXIV Konferencja Naukowa Sekcji Klasyfikacji i Analizy Danych PTS (XXIX Konferencja Taksonomiczna) „Klasyfikacja i analiza danych – teoria i zastosowania", zorganizowana przez Sekcję Klasyfikacji i Analizy Danych Polskiego Towarzystwa Statystycznego oraz Katedrę Statystyki Wydziału Zarządzania Uniwersytetu Gdańskiego.

W trakcie dwóch sesji plenarnych oraz 13 sesji równoległych wygłoszono 58 referatów poświęconych aspektom teoretycznym i aplikacyjnym zagadnienia klasyfikacji i analizy danych. Odbyła się również sesja plakatowa, na której zaprezentowano 14 plakatów.

Teksty 24 recenzowanych artykułów naukowych stanowią zawartość prezentowanej publikacji z serii Taksonomia nr 27. Teksty 25 recenzowanych artykułów naukowych znajdują się w Taksonomii nr 26.

Krzysztof Jajuga, Marek Walesiak

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 427 • 2016

Taksonomia 27 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

LOKALNA OCENA MOCY DYSKRYMINACYJNEJ ZMIENNYCH

LOCAL EVALUATION OF A DISCRIMINATION POWER OF THE VARIABLES

DOI: 10.15611/pn.2016.427.15

Streszczenie: Metoda *k* najbliższych sąsiadów pomimo swej prostoty daje w niektórych zastosowaniach zaskakująco dobre rezultaty. Jednym z jej atutów jest w naturalny sposób rozwiązany problem dyskryminacji wielu klas. Poważną wadą natomiast jest wrażliwość na zmienne nieistotne. Zmienne takie powodują tu gwałtowny spadek dokładności klasyfikacji. K. Kira i L.A. Rendell oraz I. Kononenko zaproponowali kryterium doboru zmiennych dedykowane metodzie *k* najbliższych sąsiadów. Praktycznym problemem stosowania ich algorytmu jest wybór parametrów: liczby iteracji, liczby najbliższych sąsiadów, wartości progowej ważności zmiennych. W artykule podjęto dyskusję nad tym problemem. Algorytm poddany będzie empirycznej weryfikacji, gdzie wykorzystane będą zbiory rzeczywiste z dołączonymi zmiennymi bez mocy dyskryminacyjnej.

Słowa kluczowe: metoda *k* najbliższych sąsiadów, dobór zmiennych, lokalna ocena ważności zmiennych.

Summary: The drawback of the kNN method is its sensitivities on irrelevant variables. K. Kira and L.A. Rendell and I. Kononenko have proposed a filter which evaluates variable importance using local information. A practical problem of the use of their algorithm is the choice of parameters (the number of iterations, the number of nearest neighbors and the threshold). In this paper we empirically verify the algorithm using real data and artificially generated variables without discrimination power.

Keywords: *k* nearest neighbors, filters, local evaluation of variable importance.

1. Wstęp

Jednym z najprostszych ideowo klasyfikatorów jest metoda *k* najbliższych sąsiadów (w skrócie kNN od *k nearest neighbors*). W niektórych zastosowaniach daje zaskakująco dobre rezultaty, np. w rozpoznawaniu obrazów [King, Feng, Suther-

land 1995]. Jednym z atutów tej metody jest w naturalny sposób rozwiązany problem dyskryminacji wielu klas. Poważną wadą natomiast, wrażliwość na zmienne nieistotne, to jest takie, które nie mają wpływu na zmienną objaśnianą. Zmienne takie mogą powodować gwałtowny spadek dokładności klasyfikacji. Z tego powodu ważnym etapem stosowania metody kNN jest dobór zmiennych. Wydaje się, że odpowiednim dla tego klasyfikatora kryterium jakościującym zmienne są wagi wyliczane przez algorytm Relief, zaproponowany przez K. Kirę i L.A. Rendella [1992], a następnie modyfikowany przez I. Kononenkę [1994]. Autorzy inspirowani metodą kNN opracowali sposób oceny ważności zmiennych na podstawie informacji lokalnej, jaka tkwi w odległościach między obiektami. Algorytm reprezentuje podejście wielowymiarowe. Jest w stanie rozpoznać pewne interakcje między zmiennymi (zob. [Kubus 2015a]), które nie są rozpoznawalne przez kryteria ani jednowymiarowe, ani wielowymiarowe, takie jak pojemność informacji Hellwiga [1969] czy korelacja grupowa [Hall 2000].

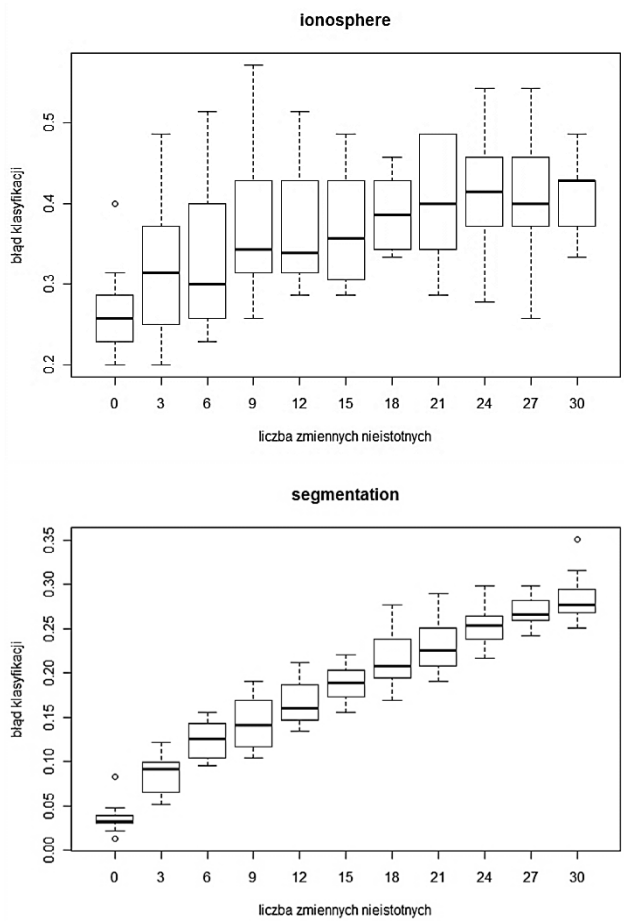
Warunkiem skuteczności algorytmu Relief jest wybór właściwych wartości jego parametrów: liczby iteracji, liczby najbliższych sąsiadów oraz progowej wagi, dzielącej zmienne na ważne i nieważne. Celem artykułu jest empiryczna weryfikacja tego algorytmu. Dyskutowany będzie problem optymalnego wyboru jego parametrów. W badaniach wykorzystane będą zbiory rzeczywiste oraz sztucznie generowane zmienne bez mocy dyskryminacyjnej.

2. Metoda kNN a zmienne nieistotne

Metoda k najbliższych sąsiadów jest metodą nieparametryczną, która dokonuje klasyfikacji na podstawie odległości między obiektami. Została zaproponowana już w 1951 r. przez E. Fixa i J. Hodgesa [1951], a jej ideę odzwierciedla przysłowie „Z jakim przestajesz, takim się stajesz". Etap uczenia (estymacji parametrów modelu) jest pominięty, a klasyfikacji dokonuje się na podstawie klas najbliższych obiektów. Tak więc dla obserwacji $\boldsymbol{x}$ poddawanej klasyfikacji wyznacza się k najmniej od niej oddalonych obiektów ze zbioru uczącego. W najprostszym przypadku obiekt $\boldsymbol{x}$ przypisany jest do klasy najczęściej występującej wśród k najbliższych sąsiadów (tzw. głosowanie majoryzacyjne). Inny sposób głosowania polega na wprowadzeniu wag, które są funkcjami odległości obiektu rozpoznawanego $\boldsymbol{x}$ od najbliższych sąsiadów. Funkcja ta powinna być malejąca, przyjmować wartości nieujemne oraz osiągać maksimum dla odległości równej 0. Wagi są w odpowiednich klasach sumowane, a obserwacja $\boldsymbol{x}$ przypisana do tej z maksymalną sumą. Modyfikacja ta nazywana jest metodą ważonych k najbliższych sąsiadów [Hechenbichler, Schliep 2004].

Ważnymi aspektami w stosowaniu metody kNN są: wybór parametru k, wybór funkcji odległości oraz formuły normalizacyjnej, wybór sposobu ważenia odległości. Osobny problem stanowią duży wymiar przestrzeni cech oraz zmienne nie-

istotne. Zilustruje to następujący przykład. Do wybranych zbiorów z tab. 2 dołączano w kolejnych dziesięciu iteracjach zmienne o jednakowych rozkładach w klasach, zatem nie mające mocy dyskryminacyjnej. Na rysunku 1 pokazano, jak rośnie błąd klasyfikacji z liczbą wprowadzanych zmiennych nieistotnych. W zbiorze *ionosphere* są dwie klasy, a w *segmentation* siedem. Błąd klasyfikacji rozumiany jest w artykule jako frakcja niepoprawnie sklasyfikowanych obiektów.



Rys. 1. Błędy klasyfikacji estymowane 10-częściowym sprawdzaniem krzyżowym

Źródło: opracowanie własne.

W przypadku analizy zbiorów, gdzie dane były gromadzone bez dostatecznej wiedzy merytorycznej o wpływie predyktorów na badane zjawisko, selekcja zmiennych jest szczególnie zalecana. Pozwala uzyskać dokładniejsze klasyfikatory oraz wydobywa wiedzę na temat ważności czynników wpływających na zmienną

objaśnianą. Wśród trzech podejść do selekcji zmiennych [Guyon i in. 2006] dużą popularność mają metody doboru zmiennych na podstawie kryterium oceny, której dokonuje się przed konstrukcją klasyfikatora (*filters*).

3. Algorytm Relief i jego modyfikacje

Najczęściej stosowane kryteria doboru zmiennych mają charakter jednowymiarowy (statystyki testowe, miary bazujące na entropii), to znaczy oceniają indywidualny wpływ predyktora na zmienną objaśnianą. Są one łatwe w implementacji i szybkie w działaniu. Z drugiej strony, mogą nie wykryć łącznego wpływu predyktorów na zmienną objaśnianą oraz nie eliminują zmiennych powielających informacje (zmiennych redundantnych). Kryteria wielowymiarowe oceniają zdolność dyskryminacyjną podzbioru zmiennych. Są one przeważnie skonstruowane tak, by maksymalizować siłę związku predyktorów ze zmienną objaśnianą i jednocześnie minimalizować zależność między predyktorami [Hellwig 1969; Hall 2000]. Wspólną cechą tych kryteriów jest ich charakter globalny, to znaczy ich wartości wyliczane są z użyciem obiektów z całej dziedziny badanego zjawiska.

K. Kira i L.A. Rendell [1992] opracowali algorytm Relief, który nadaje wagi zmiennym na podstawie informacji lokalnej. Jego inspiracją była metoda *k* najbliższych sąsiadów. Algorytm (tab. 1) w sposób iteracyjny modyfikuje wagi zmiennych w oparciu o odległości między losowo wybranym obiektem a jego najbliższym sąsiadem z tej samej klasy (*nearest hit*) oraz najbliższym sąsiadem z klasy przeciwnej (*nearest miss*). Algorytm można stosować dla zmiennych binarnych lub ilościowych, przy czym te drugie są normalizowane tak, by przyjmowały wartości z przedziału [0; 1]. Stosowanym przez autorów czynnikiem normalizacyjnym jest zakres zmiennej. Formuła aktualizacyjna (krok 4 algorytmu) jest tak skonstruowana, że waga *j*-tej zmiennej rośnie, gdy jej wartości dużo różnią się w przypadku obiektów (najbliższych sąsiadów) z różnych klas i odwrotnie, mało się różnią w przypadku obiektów (najbliższych sąsiadów) z tych samych klas.

Tabela 1. Algorytm Relief

1.	Ustal wartości początkowe: wektor wag zmiennych $\boldsymbol{W} = (W_1, \ldots, W_p) = \mathbf{0}$, liczbę iteracji M.
2.	Wylosuj obiekt $\boldsymbol{x_i} = (x_1, \ldots, x_p)$ ze zbioru uczącego.
3.	Znajdź dla niego najbliższy obiekt tej samej klasy $\boldsymbol{x_{(h)}}$ oraz najbliższy obiekt klasy przeciwnej $\boldsymbol{x_{(m)}}$.
4.	Dla każdej zmiennej $X_1, \ldots, X_p$ aktualizuj jej wagę wg formuły: $W_j \leftarrow W_j - (x_j - x_{(h)j})^2 + (x_j - x_{(m)j})^2$.
5.	Kroki 2–4 powtarzaj M razy.
6.	Ostatecznie: $\boldsymbol{W} \leftarrow \boldsymbol{W}/M$.

Źródło: opracowanie własne na podstawie [Kira, Rendell 1992].

Podobnie jak w innych metodach doboru zmiennych algorytm Relief ustala ranking zmiennych. Należy zatem użyć wartości progowej w celu podjęcia decyzji, które zmienne usunąć. Zastosowane normalizacje zapewniają, że wagi są z przedziału [–1; 1]. Zmienne, którym przyporządkowano wagi mniejsze od zera, uważane są za nieistotne. Można też przyjąć bardziej radykalną wartość progową τ. K. Kira i L.A. Rendell [1992], traktując wagi jako zmienne losowe, wykazują na podstawie nierówności Czebyszewa, że przyjmując:

$$\tau = 1 / \sqrt{\alpha M} , \tag{1}$$

prawdopodobieństwo popełnienia błędu pierwszego rodzaju jest mniejsze od przyjętego poziomu istotności α. W przypadku, gdy liczba iteracji M jest równa liczbie obserwacji w zbiorze uczącym, mamy wariant deterministyczny algorytmu.

Algorytm Relief doczekał się kilku modyfikacji i rozszerzeń. I. Kononenko [1994] proponuje brać k najbliższych sąsiadów tej samej klasy i k najbliższych sąsiadów klasy przeciwnej. W formule aktualizacyjnej (krok 4 algorytmu) różnice wartości j-tej zmiennej ważone prawdopodobieństwami *a priori* są wówczas uśredniane dla wszystkich najbliższych sąsiadów. Takie podejście zapewnia większą odporność na szum zawarty w danych. We wspomnianym artykule rozwiązano też problem brakujących danych oraz problem dyskryminacji wielu klas. W ten sposób zmodyfikowana przez I. Kononenkę [1994] wersja algorytmu nazywana jest w literaturze ReliefF.

Kluczowe znaczenie dla skutecznego stosowania algorytmu ReliefF ma ustalenie jego parametrów. Można tego dokonać w sposób całkowicie zautomatyzowany, stosując sprawdzanie krzyżowe (*tuning*). Takie podejście ma jednak swoje słabe strony. Dla dużych zbiorów danych i dużej liczby parametrów znacząco wzrasta czas obliczeń. Z kolei w przypadku małych zbiorów wyniki mogą być mało stabilne. Algorytm ReliefF wymaga ustalenia trzech parametrów: liczby najbliższych sąsiadów, liczby iteracji oraz wartości progowej ważności zmiennych. Cenne mogą być zatem wszelkie wyniki empiryczne, które dawałyby sugestie co do ich wyboru.

4. Badania empiryczne

W przeprowadzonym badaniu przedyskutowano problemy ustalenia parametrów algorytmu ReliefF. Następnie porównano jakość klasyfikacji z selekcją oraz bez selekcji zmiennych. Jako klasyfikatora użyto metodę ważonych k najbliższych sąsiadów. Metoda ta może być stosowana z różnymi funkcjami odległości, funkcjami ważącymi odległości, czy też liczbami najbliższych sąsiadów. Parametry te mogą być ustalane automatycznie przez procedurę sprawdzania krzyżowego. Gdyby jednak dołączyć listę parametrów algorytmu ReliefF, przestrzeń przeszukiwań ich optymalnego zestawu znacząco zwiększyłaby wymiar. Wobec tego w badaniu zdecydowano się przyjąć (chyba najczęściej stosowaną) odległość euklidesową

oraz ważącą odległości funkcję Epanechnikova, która znalazła skuteczne zastosowanie na przykład w zadaniu *scoringu* marketingowego [Kubus 2015b]. Ma ona postać:

$$K(d) = \frac{3}{4}\left(1 - d^2\right), \tag{2}$$

gdzie d jest odległością między obiektem rozpoznawanym a obiektem ze zbioru najbliższych sąsiadów. Do ustalenia liczby sąsiadów wykorzystano wynik teoretyczny, jaki uzyskali G.G. Enas i S.C. Choi [1986], to znaczy przyjęto $k \approx N^{2/8}$, gdzie N jest liczbą obiektów w zbiorze uczącym. W badaniu wykorzystano zbiory powszechnie używane do porównań metod statystycznego uczenia [Frank, Asuncion 2010] (tab. 2).

Tabela 2. Zbiory danych wykorzystane w badaniu

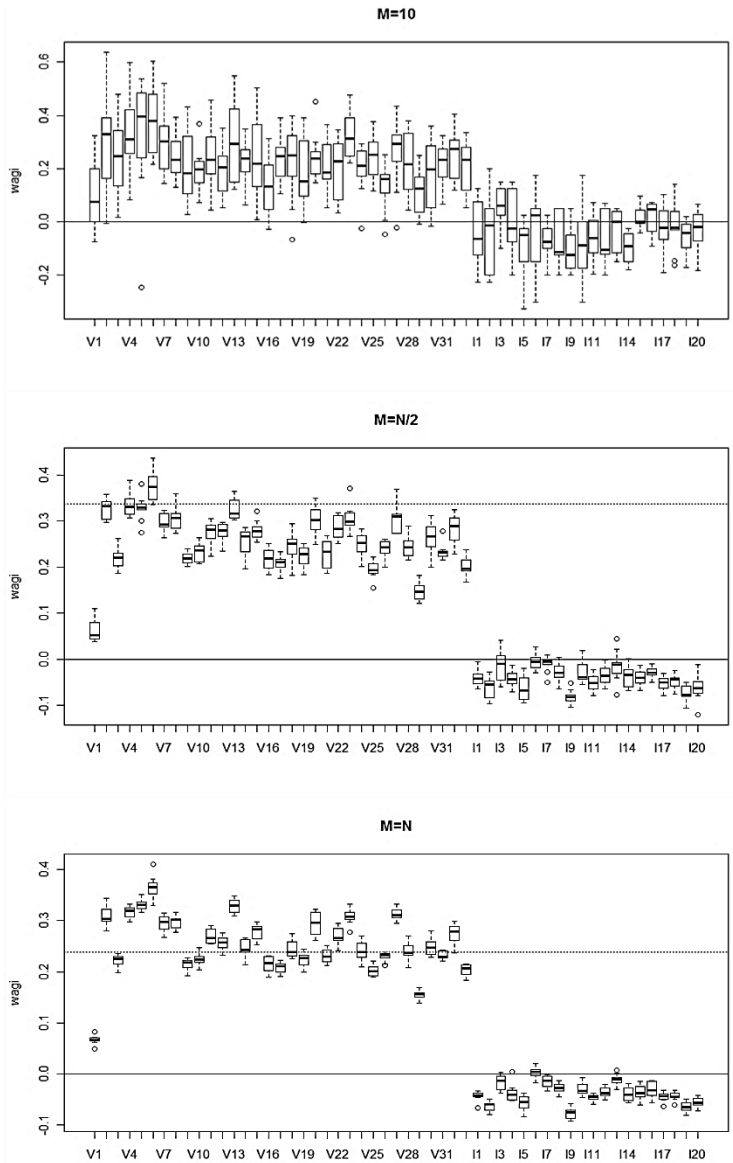
Zbiory	Liczba obiektów	Liczba zmiennych	Liczba klas
biodeg	1055	41	2
cardiotocographic	2126	21	3
ecoli	336	7	8
glass	214	9	6
ionosphere	351	33	2
segmentation	2310	19	7
sonar	208	60	2

Źródło: repozytorium Uniwersytetu Kalifornijskiego, http://archive.ics.uci.edu/ml/.

Do ustalenia liczby sąsiadów w algorytmie ReliefF zaadaptowano wspomnianą sugestię G.G. Enasa i S.C. Choi [1986] dotyczącą liczby sąsiadów w klasyfikatorze kNN. Kwestia liczby iteracji łatwo może być rozstrzygnięta przez przyjęcie wariantu deterministycznego. Jednak dla dużych N oznacza to wysoki koszt obliczeń. Z kolei zbyt mała liczba iteracji może wiązać się z utratą stabilności wyników lub z pominięciem ważnej informacji tkwiącej w danych. Rysunek 2 przedstawia wagi zmiennych oszacowane przez ReliefF (w 10-częściowym sprawdzaniu krzyżowym) dla różnych liczb iteracji. Do ilustracji wykorzystano zbiór *ionosphere*, do którego dołączono 20 zmiennych o jednakowych rozkładach warunkowych (był to rozkład normalny lub zero-jedynkowy). Zmienne te, bez mocy dyskryminacyjnej, oznaczono literą *I*.

Ustalenie zbyt małej liczby iteracji (tutaj 10)[1] powoduje, że wagi mają większe wariancje i nie wszystkie zmienne nieistotne są rozpoznawane (to znaczy ich średnie wagi są większe od zera). Z kolei wagi zmiennych w przypadku N iteracji oraz $N/2$ iteracji są podobne, a zmienne nieistotne stosunkowo dobrze rozpoznawane.

[1] Liczba iteracji równa 10 jest rekomendowana w pracy I. Kononenki [1994].



Rys. 2. Wagi predyktorów oszacowane algorytmem ReliefF dla zbioru *ionosphere* z wprowadzonymi zmiennymi nieistotnymi (*I*). Linie poziome reprezentują wybrane wartości progowe

Źródło: obliczenia własne.

Linie poziome na rys. 2 reprezentują wybrane wartości progowe τ dla wag. Wagi mniejsze wskazują na to, że zmienne są nieistotne. Próg równy zero (linia ciągła) dosyć dobrze oddziela sztucznie wprowadzone zmienne od oryginalnych,

choć nie zapewnia ich idealnego rozpoznawania. Dla $M = 10$ algorytm wprowadzał w sprawdzaniu krzyżowym średnio 6 zmiennych nieistotnych, a dla $M = N/2$ i wariantu deterministycznego – jedną. Wydaje się, że korzystniejsze byłoby ustalenie nieco większej wartości progowej τ. Z kolei próg wyliczony według formuły (1), zaznaczony na rys. 2 linią przerywaną, jest zdecydowanie zbyt radykalny. W przypadku, gdy M=10 uzyskano τ = 1,41, co skutkowało odrzuceniem wszystkich zmiennych. Podobne wyniki uzyskano w zbiorach *biodeg* oraz *sonar* z identycznie wprowadzanymi zmiennymi nieistotnymi. W *sonar* algorytm wprowadzał średnio od 3 do 6 zmiennych nieistotnych dla rozważanych liczb iteracji M.

W związku z trudnościami w ustaleniu wartości progowej τ, w artykule proponuje się następującą procedurę. Metodą sprawdzania krzyżowego szacowany będzie błąd klasyfikacji dla klasyfikatorów z jedną, najlepszą w rankingu zmienną, z dwoma zmiennymi itd. Przyjmiemy, że błąd minimalny wskazuje optymalny podzbiór zmiennych. Do celów porównawczych przeprowadzono badanie, wykorzystując zbiory z pracy [Kubus 2015c], to jest *biodeg*, *ionosphere* oraz *sonar* z dołączonymi zmiennymi nieistotnymi. Wyniki prezentuje tab. 3. Symbole (10b) oraz (10n) przy nazwach zbiorów oznaczają, że wprowadzano do nich odpowiednio 10 zmiennych nieistotnych binarnych lub z rozkładu normalnego. Proponowana procedura jedynie w zbiorze *sonar* nie rozpoznała bezbłędnie wszystkich zmiennych nieistotnych. Średnia liczba takich wprowadzeń była jednak niewielka, mniejsza od jeden.

Tabela 3. Średnie liczby zmiennych nieistotnych wprowadzanych do klasyfikatorów kNN przy 50-krotnym podziale zbiorów na część uczącą i testową

Zbiór ze zmiennymi nieistotnymi	*biodeg* (10b)	*biodeg* (10n)	*ionosphere* (10b)	*ionosphere* (10n)	*sonar* (10b)	*sonar* (10n)
Średnia liczba zmiennych nieistotnych	0	0	0	0	0,96	0,20

Źródło: obliczenia własne.

Tabela 4. Średnie błędy klasyfikacji z błędami standardowymi (w %) estymowane 50 razy na zbiorach testowych (1/3 oryginalnego zbioru danych) oraz wartości p jednostronnego testu sumy rang

Zbiory	kNN – bez selekcji zmiennych	ReliefF i kNN	Wartości p
biodeg	14,2 (0,2)	15,1 (0,3)	0,99355
cardiotocographic	9,0 (0,1)	8,4 (0,2)	0,00075
ecoli	17,4 (0,4)	17,9 (0,4)	0,71677
glass	31,9 (0,8)	26,2 (0,7)	0,00001
ionosphere	13,4 (0,5)	12,1 (0,4)	0,05528
segmentation	4,2 (0,1)	3,8 (0,1)	0,00627
sonar	15,2 (0,6)	15,9 (0,5)	0,74402

Źródło: obliczenia własne.

Uzyskane dotychczas wyniki zweryfikowano na danych rzeczywistych. Zastosowano dwukrotnie klasyfikator kNN z funkcją Epanechnikova (2) ważącą odległości. W pierwszym przypadku bez selekcji zmiennych, w drugim stosując algorytm ReliefF. Liczbę sąsiadów – zarówno w klasyfikatorze, jak i w algorytmie selekcji zmiennych – ustalano według sugestii G.G. Enasa i S.C. Choi [1986]. Liczbę iteracji w ReliefF przyjęto $N/2$, natomiast wartość progową ustalano za pomocą zaproponowanej wcześniej procedury. Rezultaty podsumowuje tab. 4. Jedynie w zbiorze *biodeg* algorytm ReliefF nie poprawił dokładności klasyfikacji. W przypadku zbiorów *cardiotocographic*, *glass*, *segmentation* poprawa była znacząca. Zauważmy, że są to zbiory z liczbą klas większą od 2.

5. Podsumowanie

Algorytm Relief w odróżnieniu od większości popularnych kryteriów doboru zmiennych dokonuje oceny ważności zmiennych na podstawie informacji lokalnej. Podobnie jak w wielu metodach *data mining*, praktycznym problemem jego stosowania jest ustalenie optymalnych wartości jego parametrów. W artykule sugeruje się, by liczbę sąsiadów przyjąć taką jaką G.G. Enas i S.C. Choi [1986] przyjęli dla klasyfikatora kNN. Zaproponowano też procedurę wyboru liczby zmiennych rozwiązującą problem ustalenia wartości progowej dla wag. Pokazano empirycznie, że sugerowana w literaturze liczba iteracji 10 [Kononenko 1994] nie gwarantuje poprawnego rozpoznawania zmiennych nieistotnych oraz że wariant deterministyczny można przybliżyć wykonując połowę iteracji. Zaproponowane ustawienia dały obiecujące rezultaty na zbiorach danych rzeczywistych.

Literatura

Enas G.G., Choi S.C., 1986, *Choice of the smoothing parameter and efficiency of k-nearest neighbor classification*, Computer and Mathematics with Applications, no. 12A(2), s. 235–244.

Fix E., Hodges J., 1951, *Discriminatory analysis – nonparametric discrimination: Consistency properties*, Technical Report 21-49-004,4, US Air Force, School of Aviation Medicine, Randolph Field, TX.

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

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

Hall M., 2000, *Correlation-based feature selection for discrete and numeric class machine learning*, [w:] P. Langley (red.), *Proceedings of the 17th International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, s. 359–366.

Hechenbichler K., Schliep K.P., 2004, *Weighted k-Nearest-Neighbor Techniques and Ordinal Classification*, Discussion Paper 399, SFB 386, Ludwig-Maximilians Universität, München.

Hellwig Z., 1969, *Problem optymalnego wyboru predykant*, Przegląd Statystyczny, nr 3/4, s. 221–237.

King R.D., Feng C., Sutherland A., 1995, *StatLog: Comparison of classification algorithms on large real-world problems*, Applied Artificial Intelligence, vol. 9, no. 3, s. 289–333.

Kira K., Rendell L.A., 1992, *The feature selection problem: Traditional methods and a new algorithm,* [w:] AAAI'92 Proceedings of the Tenth National Conference on Artificial Intelligence, MIT Press, Cambridge, MA, s. 129–134.

Kononenko I., 1994, *Estimating attributes: Analysis and extensions of RELIEF*, [w:] F. Bergadano, L.D. Raedt (red.), *Machine Learning ECML-94*, Springer, Berlin–Heidelberg, s. 171–182.

Kubus M., 2015a, *Feature selection and the chessboard problem*, Acta Universitatis Lodziensis, Folia Oeconomica, Statistical Analysis in Theory and Practice, nr 1 (311), s. 17–25 (DOI 11089/21).

Kubus M., 2015b, *Identyfikacja potencjalnych nabywców polis ubezpieczeniowych w warunkach mocno niezbilansowanej próby uczącej*, Ekonometria, nr 2(48), s. 89–99 (DOI: 10.15611/ekt.2015.2.08).

Kubus M., 2015c, *Rekurencyjna eliminacja cech w metodach dyskryminacji*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 384, Taksonomia 24: *Klasyfikacja i analiza danych – teoria i zastosowania*, s. 154–162 (DOI: 10.15611/pn.2015.384.16).