EKONOMETRIA ECONOMETRICS 3(53) • 2016

ISSN 1507-3866
e-ISSN 2449-9994

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

PROPOZYCJA AGREGOWANEGO KLASYFIKATORA kNN Z SELEKCJĄ ZMIENNYCH

THE PROPOSITION OF THE KNN ENSEMBLE WITH FEATURE SELECTION

DOI: 10.15611/ekt.2016.3.03
JEL Classification: C01, C14, C52

Streszczenie: Modele zagregowanych drzew klasyfikacyjnych zyskały uznanie ze względu na poprawę stabilności i często redukcję obciążenia. Adaptacja tego podejścia do metody *k* najbliższych sąsiadów (kNN) napotyka jednak na pewne trudności: względnie duża stabilność tych klasyfikatorów oraz wzrost błędu klasyfikacji, gdy w zbiorze uczącym są zmienne bez mocy dyskryminacyjnej. W artykule proponuje się agregowany klasyfikator kNN z selekcją zmiennych. Jego dokładność klasyfikacji zweryfikowana jest na zbiorach rzeczywistych z dołączonymi zmiennymi nieistotnymi.

Słowa kluczowe: metoda *k* najbliższych sąsiadów, podejście wielomodelowe, selekcja zmiennych, algorytm ReliefF.

Summary: Aggregated classification trees have gained recognition due to improved stability, and frequently reduced bias. However, the adaptation of this approach to the *k* nearest neighbors method (kNN), faces some difficulties: the relatively high stability of these classifiers, and an increase of misclassifications when the variables without discrimination power are present in the training set. In this paper we propose aggregated kNN classifier with feature selection. Its classification accuracy has been verified on the real data with added irrelevant variables.

Keywords: *k* nearest neighbors, ensemble, feature selection, ReliefF algorithm.

1. Wstęp

Wraz z szybko rozwijającą się technologią informacyjną i gromadzeniem dużych baz danych znaczenia nabierają techniki eksploracyjnej analizy danych (*data mining*). Dane traktowane są jako zasób, a wydobyta z nich informacja wspomaga decyzje menadżerskie. W ostatnim dwudziestoleciu dużą popularność zyskały modele zagregowanych drzew klasyfikacyjnych, które znajdują zastosowanie między innymi w badaniach lojalności czy satysfakcji klientów. Podejście wielomodelowe ma wiele

zalet. Można udowodnić, że błąd modelu zagregowanego jest mniejszy od średniego błędu modeli bazowych [Breiman 1996; Fumera, Roli 2001]. Model zagregowany jest bardziej stabilny, co ma teoretyczne uzasadnienie w dekompozycji błędu klasyfikacji lub predykcji [Breiman 1996; Freund, Schapire 1997; Gatnar 2008]. W przypadku drzew wzmacnianych (*boosting*) występuje też redukcja obciążenia [Freund, Schapire 1997]. Modele zagregowane łatwo adaptować do przetwarzania danych pojawiających się sekwencyjnie (w odstępach czasu) lub do przetwarzania bardzo dużych zbiorów danych, które w tym przypadku można dzielić na mniejsze części, dla których buduje się modele, a następnie agreguje wyniki. Zadanie to może być realizowane na wielu komputerach. Trzy często przytaczane w literaturze argumenty teoretyczne przemawiające za podejściem wielomodelowym podał też T.G. Dietterich [2000]. Łączenie modeli w jeden model zagregowany rozwijano głównie dla drzew klasyfikacyjnych. Można jednak znaleźć w literaturze próby łączenia innych modeli, np. sieci neuronowych [Opitz, Maclin 1999] czy logicznych reguł klasyfikacji [Friedman, Popescu 2005].

W artykule podjęto próbę adaptacji podejścia wielomodelowego do metody *k* najbliższych sąsiadów (*k nearest neighbours* – kNN). L. Breiman [1996], badając przyczyny sukcesu podejścia wielomodelowego, zwraca uwagę na ważność zróżnicowania modeli bazowych. Postulat ten łatwiej spełnić modelom niestabilnym, kiedy niewielkie zmiany w zbiorze uczącym wpływają na postać modelu. Klasyfikatory *k* najbliższych sąsiadów jako stabilne nie są rekomendowane do agregowania. Mimo to w literaturze przedmiotu można znaleźć próby adaptacji podejścia wielomodelowego do metody kNN [Bay 1999; Domeniconi, Yan 2004; Zhou, Yu 2005; Gul i in. 2014]. Poważnym problemem stosowania metody *k* najbliższych sąsiadów są zmienne bez mocy dyskryminacyjnej, nazywane dalej zmiennymi nieistotnymi (*irrelevant variables*). Błąd klasyfikacji dosyć szybko rośnie wraz z liczbą takich zmiennych w zbiorze danych [Kubus 2016].

Głównym celem artykułu jest zweryfikowanie własnej propozycji klasyfikatora kNN, wykorzystującego selekcję zmiennych oraz podejście wielomodelowe. Zastosowany będzie algorytm doboru zmiennych ReliefF [Kononenko 1994], specjalnie dedykowany metodzie *k* najbliższych sąsiadów. Badania empiryczne przeprowadzone będą na powszechnie wykorzystywanych do celów porównawczych zbiorach danych z repozytorium Uniwersytetu Kalifornijskiego [Frank, Asuncion 2010]. Do zbiorów rzeczywistych dołączane będą również zmienne nieistotne generowane tak, by miały jednakowe rozkłady w klasach, co oznacza brak mocy dyskryminacyjnej.

2. Metoda kNN a podejście wielomodelowe

Metoda *k* najbliższych sąsiadów (kNN), mimo swej prostoty, daje w niektórych zastosowaniach zaskakująco dobre rezultaty, na przykład w rozpoznawaniu obrazów [King, Feng, Sutherland 1995]. Jej niewątpliwym atutem to w naturalny sposób rozwiązany problem dyskryminacji wielu klas. Metoda wykorzystuje jedynie odległo-

ści między obiektami, rozumiane jako ich podobieństwo ze względu na określone cechy. Dla obserwacji $\boldsymbol{x}$ poddawanej klasyfikacji wyznacza się k najmniej od niej oddalonych obiektów (najbliższych sąsiadów) ze zbioru uczącego. W najprostszym przypadku obiekt $\boldsymbol{x}$ przypisany jest do klasy najczęściej występującej wśród tych k obiektów (tzw. głosowanie majoryzacyjne). Inny sposób głosowania polega na wprowadzeniu wag, które są funkcjami odległości obiektu rozpoznawanego $\boldsymbol{x}$ od najbliższych sąsiadów. Wagi są w odpowiednich klasach sumowane, a obserwacja $\boldsymbol{x}$ przypisana do tej z maksymalną sumą [Hechenbichler, Schliep 2004]. Stosując metodę kNN, należy wybrać regułę normalizacji zmiennych oraz sposób obliczania odległości. Kluczowe znaczenie ma jednak wybór parametru k, tzn. liczby najbliższych sąsiadów. Można go ustalić adaptacyjnie, w sposób całkowicie zautomatyzowany, stosując sprawdzanie krzyżowe. Z kolei G.G. Enas i S.C. Choi [1986], na podstawie przeprowadzonych symulacji, proponują, by przyjąć $k \approx N^{2/8}$ lub $k \approx N^{3/8}$, gdzie N jest liczbą obiektów w zbiorze uczącym. Poważną wadą metody kNN jest wzrost błędu klasyfikacji w sytuacji, gdy w zbiorze uczącym są zmienne niemające wpływu na zmienną objaśnianą [Kubus 2016]. Gdy analityk nie dysponuje dostateczną wiedzą na temat badanego zjawiska i celem analizy jest wydobycie wiedzy z danych (*knowledge discovery*), wykrywanie takich zmiennych decyduje o skuteczności metody kNN.

Wykorzystanie podejścia wielomodelowego w połączeniu z metodą kNN napotyka pewne problemy. Jak już wspomniano, agregowane klasyfikatory powinny być zróżnicowane oraz dostatecznie dokładne, tzn. przynajmniej nieco dokładniejsze od klasyfikacji na podstawie prawdopodobieństw *a priori* [Breiman 1996]. Pierwszy postulat realizowany jest przeważnie przez dobór różnych prób uczących do klasyfikatorów bazowych lub rzutowanie obiektów na różne podprzestrzenie. W przypadku modeli mało stabilnych, małe zmiany w zbiorze uczącym wpływają na postać klasyfikatora, co zapewnia odpowiednie zróżnicowanie. Klasyfikatory kNN wykorzystujące jedynie odległości między obiektami są względnie stabilne. Z kolei rzutowanie obiektów na różne podprzestrzenie koresponduje w metodzie k najbliższych sąsiadów z problemem zmiennych nieistotnych. Zwrócił na to uwagę już S.D. Bay [1999], który proponował najprostsze podejście, polegające na losowym wyborze podprzestrzeni. C. Domeniconi i B. Yan [2004] opracowali więc system wag dla zmiennych tak, by były losowane z różnym prawdopodobieństwem do kolejnych klasyfikatorów bazowych. Inny sposób osiągnięcia zróżnicowania klasyfikatorów bazowych zastosowali Z.H. Zhou i Y. Yu [2005], wprowadzając losowość do miary odległości. Z kolei A. Gul i in. [2014] stosują kontrolę jakości pojedynczych klasyfikatorów, nie wprowadzając do modelu tych, które nie osiągnęły ustalonej, progowej wartości funkcji oceny. Zabieg ten przypomina nieco upraszczanie wstępne drzew klasyfikacyjnych (*pre-pruning*), a w efekcie jest również sposobem radzenia sobie z problemem zmiennych nieistotnych. Gdy takie zmienne zostaną wylosowane do klasyfikatora i spowodują znaczny spadek jego dokładności, klasyfikator nie zostanie wprowadzony do modelu agregowanego.

Tabela 1. Proponowany algorytm agregacji klasyfikatorów kNN

1	Ustal wartości początkowe: – liczbę klasyfikatorów bazowych M, – maksymalną liczbę zmiennych q w pojedynczym klasyfikatorze.
2	Wylosuj ze zbioru uczącego próbę bootstrapową U_b .
3	Na próbie U_b zastosuj selekcję zmiennych, by otrzymać podzbiór zmiennych S o największej mocy dyskryminacyjnej.
4	Jeśli $card(S) > q$, to wylosuj q zmiennych z podzbioru S.
5	Na zbiorze U_b oraz dla zredukowanego zestawu zmiennych zastosuj klasyfikator kNN. Zapamiętaj wektor oszacowanych prawdopodobieństw *a posteriori* (lub klasyfikacji) dla obiektów ze zbioru rozpoznawanego.
6	Kroki 2-5 powtarzaj M razy.
7	Dokonaj agregacji M wektorów prawdopodobieństw *a posteriori* (lub klasyfikacji).

Źródło: opracowanie własne.

W odpowiedzi na problem wpływu zmiennych nieistotnych na dokładność klasyfikacji w metodzie kNN w artykule proponuje się następujący algorytm (tab. 1). Kluczowe znaczenie ma stosowany w kroku 3 algorytm selekcji zmiennych. Tu zaimplementowano dobór zmiennych algorytmem ReliefF [Kononenko 1994], który nadaje zmiennym wagi według lokalnie ocenianej mocy dyskryminacyjnej. Jest to metoda selekcji zmiennych, która − podobnie jak rozważane klasyfikatory − wykorzystuje jedynie odległości między obiektami. Pozostałe parametry proponowanego w tab. 1 algorytmu to: liczba klasyfikatorów bazowych, maksymalna liczba zmiennych w pojedynczym klasyfikatorze, liczba najbliższych sąsiadów oraz sposób agregacji wyników klasyfikacji. W przypadku posłużenia się prawdopodobieństwami *a posteriori* można wyznaczyć ich średnią arytmetyczną lub medianę. Zauważmy, że jeśli liczba zmiennych o dużej mocy dyskryminacyjnej jest większa od parametru q, to algorytm ma charakter stochastyczny.

3. Ocena lokalnej mocy dyskryminacyjnej zmiennych

Obecnie w literaturze przedmiotu metody selekcji zmiennych dzielone są na trzy grupy (zob. np. [Guyon i in. 2006]). Historycznie pierwsze były metody doboru zmiennych stosowane przed algorytmem uczącym (*filters*). Moc dyskryminacyjna zmiennych oceniana jest w nich na podstawie kryterium (np. miary bazujące na entropii), które nie ma bezpośredniego związku z budowanym modelem. W drugiej grupie metod selekcji zmiennych wykorzystuje się podejście polegające na ocenie jakości modeli budowanych na różnych podzbiorach zmiennych objaśniających (*wrappers*). Do trzeciej grupy należą metody, w których selekcja zmiennych jest integralną częścią algorytmu uczącego. Tak jest na przykład w drzewach klasyfikacyjnych, w których do modelu wprowadzane są zmienne (warunki w węzłach),

które lokalnie optymalizują kryterium jednorodności. Celem selekcji zmiennych jest uzyskanie takiego podzbioru zmiennych objaśniających *S*, by model cechował się jak najlepszą zdolnością przewidywania dla nowych obiektów (spoza zbioru uczącego). Zmienne spoza podzbioru *S* nazywane są nieistotnymi (*irrelevant variables*), choć ich zdefiniowanie przysparza pewnych trudności [Guyon i in. 2006]. Po pierwsze, zmienne, które indywidualnie nie mają mocy dyskryminacyjnej, mogą ją mieć w kontekście z innymi zmiennymi. Ilustruje to benchmarkowy problem szachownicy [Kubus 2015]. Po drugie, rozwiązanie zadania selekcji zmiennych może nie być jedyne. Ponieważ selekcja zmiennych jest zadaniem optymalizacji kombinatorycznej, decydującą rolę odgrywa tu technika przeszukiwania przestrzeni podzbiorów zmiennych.

Metody doboru zmiennych (*filters*) wchodzą w zakres etapu przygotowywania danych do modelowania. Cieszą się one dużą popularnością ze względu na swą prostotę i szybkość działania. Dotyczy to najczęściej stosowanych kryteriów jednowymiarowych, które oceniają moc dyskryminacyjną pojedynczych zmiennych. Słabością tego podejścia, wynikającą z uproszczonego schematu przeszukiwania przestrzeni podzbiorów zmiennych, jest brak możliwości oceny łącznego wpływu podzbioru zmiennych na zmienną objaśnianą. Ponadto kryteria jednowymiarowe nadają w przybliżeniu jednakową ocenę zmiennym redundantnym, to jest powielającym informacje. W związku z tym w literaturze zaproponowano miary dokonujące oceny podzbiorów zmiennych, na przykład pojemność informacji Z. Hellwiga [1969]. W swej konstrukcji uśredniają one wyniki miar korelacji między dwoma zmiennymi. Inne podejście zaproponowali K. Kira i L.A. Rendell [1992] w algorytmie Relief. Nadaje on wagi zmiennym na podstawie oceny lokalnej mocy dyskryminacyjnej.

Inspiracją algorytmu Relief była metoda *k* najbliższych sąsiadów i tej metodzie jest on przede wszystkim dedykowany. Wagi zmiennych (na początku równe zeru) są w kolejnych krokach modyfikowane na podstawie odległości między losowo wybranym obiektem a jego najbliższym sąsiadem z tej samej klasy (*nearest hit*) oraz najbliższym sąsiadem z klasy przeciwnej (*nearest miss*). Waga 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. Taka konstrukcja sprawia, że kryterium oceny mocy dyskryminacyjnej zmiennych opracowane w algorytmie Relief ma charakter wielowymiarowy. I. Kononenko [1994] zaproponował wybór *k* najbliższych sąsiadów tej samej klasy i *k* najbliższych sąsiadów klasy przeciwnej. Ponadto podał sposób rozwiązania problemu dyskryminacji wielu klas. Jego wersja algorytmu nazywana jest w literaturze ReliefF. Na wyjściu uzyskuje się ranking ważności zmiennych ze względu na moc dyskryminacyjną. Podobnie jak w wielu metodach eksploracyjnej analizy danych skuteczność stosowania algorytmu ReliefF zależy od prawidłowego ustalenia jego parametrów. Są nimi: liczba iteracji, liczba najbliższych sąsiadów oraz wartość progowa dla ważności zmiennych. Studium empiryczne tego zagadnienia

można znaleźć w pracy M. Kubusa [2016], w której dobre rezultaty uzyskano, przyjmując liczbę iteracji równą połowie liczebności zbioru uczącego, natomiast do ustalenia liczby najbliższych sąsiadów wykorzystano wspomnianą sugestię G.G. Enasa i S.C. Choi [1986] dotyczącą klasyfikatorów kNN. Z kolei wartość progowa wyznaczana była według następującej procedury. Dla klasyfikatorów z jedną, najlepszą w rankingu zmienną, z dwoma zmiennymi itd. szacowano metodą sprawdzania krzyżowego błąd klasyfikacji oraz przyjęto, że błąd minimalny wskazuje optymalny podzbiór zmiennych.

4. Badania empiryczne

Badania empiryczne przeprowadzono na powszechnie wykorzystywanych do celów porównawczych zbiorach danych z repozytorium Uniwersytetu Kalifornijskiego [Frank, Asuncion 2010]. Były to zbiory: *cardiotocographic* (2126, 21, 3), *ecoli* (336, 7, 8), *glass* (214, 9, 6), *ionosphere* (351, 33, 2), *segmentation* (2310, 19, 7), *sonar* (208, 60, 2). Liczby w nawiasach oznaczają kolejno: liczbę obiektów, liczbę zmiennych objaśniających oraz liczbę klas. Do zbiorów rzeczywistych dołączono również zmienne bez mocy dyskryminacyjnej z rozkładu $N(0;1)$, generowane tak, by miały jednakowe rozkłady w klasach. Liczba dołączanych zmiennych nieistotnych w każdym zbiorze stanowiła w przybliżeniu 10% oryginalnej liczby zmiennych objaśniających. Celem badania było zweryfikowanie propozycji klasyfikatora kNN wykorzystującego podejście wielomodelowe oraz selekcję zmiennych (tab. 1). W badaniu przyjęto: liczbę iteracji $M = 100$ oraz liczbę zmiennych $q \approx \sqrt{p}$, gdzie p jest liczbą zmiennych objaśniających. Takie ustawienia przyjmował w swych badaniach L. Breiman [2001] w metodzie lasów losowych, z którą proponowany tu klasyfikator będzie w dalszej części porównany. Liczbę sąsiadów w klasyfikatorach bazowych przyjęto jako równą 1. W tym przypadku klasyfikatory kNN mają najbardziej skomplikowane granice między klasami, co oznacza, że mogą być nieco mniej stabilne. Agregacji dokonano metodą głosowania większościowego. Algorytm ReliefF wywołano z ustawieniami rekomendowanymi w pracy M. Kubusa [2016], opisanymi w poprzednim punkcie. Błąd klasyfikacji estymowany był 50 razy z wykorzystaniem zbioru testowego, który stanowił za każdym razem 1/3 liczebności oryginalnego zbioru danych.

Wyniki zestawiono w tab. 2. Porównano w niej zaproponowany klasyfikator (A-R-1NN) z klasycznym kNN dla $k = 1$ oraz z klasyfikatorem agregowanym bez selekcji zmiennych (A-1NN). Niemal zawsze klasyfikator agregowany wykazywał przewagę nad klasycznym. Jedynym wyjątkiem był zbiór *segmentation* bez dołączonych zmiennych nieistotnych. Wprowadzenie selekcji zmiennych do klasyfikatora agregowanego prowadziło do mniejszych błędów klasyfikacji. Analizę istotności różnicy w błędach dla dwóch rozważanych podejść wielomodelowych przeprowadzono jednostronnym testem sumy rang. W przypadku zbiorów z dołączonymi zmiennymi nieistotnymi uzyskano wartość $p = 0{,}06995$ dla zbioru *ionosphere* oraz

p = 0,12429 dla zbioru *sonar*. Pozostałe wartości p były mniejsze od 0,001. Podkreślmy, że w przypadku zbiorów *glass* oraz *ecoli* wprowadzono sztucznie zaledwie jedną zmienną, a w zbiorach *cardiotocographic* oraz *segmentation* − dwie. Można się spodziewać, że gdyby proporcja liczby zmiennych nieistotnych do liczby wszystkich zmiennych była większa, to różnice w błędach klasyfikacji byłyby jeszcze większe. Wówczas prawdopodobieństwo wylosowania zmiennych nieistotnych do klasyfikatora bazowego byłoby większe. W przypadku zbiorów oryginalnych istotną poprawę dokładności (poziom istotności 0,05) klasyfikatora agregowanego przez selekcję zmiennych uzyskano trzy razy. Dla zbiorów *cardiotocographic* oraz *glass* wartości p były równe odpowiednio 0,00067 oraz 0,03900, a dla zbioru *segmentation* uzyskano $p < 0{,}000001$.

Tabela 2. Średnie błędy klasyfikacji z błędami standardowymi (w %) estymowane 50 razy na zbiorach testowych. Do oryginalnych zbiorów dołączono 10% zmiennych nieistotnych z $N(0;1)$

Zbiory z dołączonymi zmiennymi nieistotnymi	1-NN	A-1NN	A-R-1NN
cardiotocographic	11,9 +/– 0,2	10 +/– 0,2	8,4 +/– 0,1
ecoli	20,2 +/– 0,5	18,9 +/– 0,5	16,5 +/– 0,5
glass	34,2 +/– 0,7	26,3 +/– 0,5	22,6 +/– 0,7
ionosphere	16,1 +/– 0,4	6,7 +/– 0,3	6,2 +/– 0,3
segmentation	8,7 +/– 0,1	4,2 +/– 0,1	3 +/– 0,1
sonar	15,2 +/– 0,6	14,1 +/– 0,6	13,1 +/– 0,6
Zbiory oryginalne	1-NN	A-1NN	A-R-1NN
cardiotocographic	9,6 +/– 0,2	9,1 +/– 0,2	8,3 +/– 0,1
ecoli	19,6 +/– 0,5	17,7 +/– 0,5	17,3 +/– 0,5
glass	29,9 +/– 0,6	24 +/– 0,6	22,1 +/– 0,7
ionosphere	13,2 +/– 0,3	6,3 +/– 0,3	6,2 +/– 0,3
segmentation	3,9 +/– 0,1	4,2 +/– 0,1	3,4 +/– 0,1
sonar	14,4 +/– 0,6	13,2 +/– 0,6	13,4 +/– 0,5

Źródło: obliczenia własne.

Opracowany algorytm A-R-1NN ma pewne cechy wspólne z metodą lasów losowych [Breiman 2001] − przede wszystkim: bootstrapowe próby uczące oraz losowanie zmiennych do klasyfikatorów bazowych. Zróżnicowanie modeli bazowych osiąga się przez maksymalne dopasowanie do danych ze zbioru uczącego. Mianowicie w drzewach nie stosuje się przycinania, a w metodzie k najbliższych sąsiadów rozważa tylko jeden obiekt najbliższy (k = 1). Obie metody stosują selekcję zmiennych. Algorytm A-R-1NN (tab. 1) wykorzystuje ReliefF [Kononenko 1994], nato-

miast drzewa mają wpisany mechanizm doboru zmiennych w schemat ich konstrukcji (*embedded method*). Również wybór parametrów *M* oraz *q* wzorowano na pracy L. Breimana [2001]. Nasuwa się zatem pytanie, czy zaproponowana metoda może konkurować z lasami losowymi? Tabela 3 przedstawia porównanie błędów klasyfikacji. Sposób ich estymacji był identyczny jak w poprzednim badaniu porównawczym. Na podstawie przeprowadzonego porównania nie można wykazać wyższości którejś z tych metod. Co ciekawe, dla dwóch zbiorów: *cardiotocographic, segmentation*, lasy losowe są wyraźnie dokładniejsze w klasyfikacji, a dla dwóch innych − *glass, sonar* − widać zdecydowaną przewagę klasyfikatora A-R-1NN. Wydaje się, że przyczyną takiego wyniku może być mała liczebność zbiorów *glass* oraz *sonar*. Zbadanie zbioru *ecoli* metodą lasów losowych napotkało trudności implementacyjne w programie R. Przyczyną była bardzo mała liczebność jednej z klas.

Tabela 3. Porównanie błędów klasyfikacji (w %) estymowanych 50 razy na zbiorach testowych dla lasów losowych oraz proponowanego agregowanego klasyfikatora kNN z selekcją zmiennych

Zbiory z dołączonymi zmiennymi nieistotnymi	Lasy losowe	A-R-1NN
cardiotocographic	6,1 +/– 0,1	8,4 +/– 0,1
glass	24,9 +/– 0,7	22,6 +/– 0,7
ionosphere	6,2 +/– 0,3	6,2 +/– 0,3
segmentation	2,6 +/– 0,1	3 +/– 0,1
sonar	18,8 +/– 0,6	13,1 +/– 0,6
Zbiory oryginalne	Lasy losowe	A-R-1NN
cardiotocographic	5,9 +/– 0,1	8,3 +/– 0,1
glass	23,2 +/– 0,7	22,1 +/– 0,7
ionosphere	6,2 +/– 0,3	6,2 +/– 0,3
segmentation	2,5 +/– 0,1	3,4 +/– 0,1
sonar	18,2 +/– 0,7	13,4 +/– 0,5

Źródło: obliczenia własne.

5. Podsumowanie

W artykule zaproponowano agregowany klasyfikator kNN z selekcją zmiennych. Choć klasyfikatory kNN są względnie stabilne, można je różnicować, stosując różne podprzestrzenie. Problem stanowią jednak zmienne bez mocy dyskryminacyjnej, powodujące w metodzie *k* najbliższych sąsiadów wzrost błędu klasyfikacji. Kluczowe staje się zatem zadanie selekcji zmiennych, kiedy wydaje się, że właściwym

kryterium oceny zmiennych są wagi nadawane przez algorytm ReliefF [Kononenko 1994]. Badanie empiryczne, w którym do oryginalnych zbiorów dołączano zmienne bez mocy dyskryminacyjnej generowane z rozkładu normalnego, dało obiecujące wyniki. Zaproponowany klasyfikator prowadził do mniejszych błędów klasyfikacji niż: klasyczna metoda 1-NN oraz agregowany klasyfikator 1-NN bez selekcji zmiennych. Atrakcyjność zaproponowanego klasyfikatora zweryfikowano też z cieszącą się dużą popularnością i uznaniem metodą lasów losowych [Breiman 2001]. W przypadku mniej licznych zbiorów danych (nieco więcej niż 200 obiektów) zaproponowany klasyfikator dawał znacznie mniejsze błędy klasyfikacji. Dla dużych zbiorów danych (ponad 2000 obiektów) przewagę wykazywały jednak lasy losowe.

Warto jeszcze nadmienić, że zaproponowany w tab. 1 algorytm łatwo poddać modyfikacjom. Można go stosować z różnymi metodami selekcji zmiennych oraz w zadaniu regresji.

Literatura

Bay S.D., 1999, *Nearest neighbour classification from multiple feature subsets*, Intelligent Data Analysis, 3(3), s. 191-209.

Breiman L., 1996, *Bagging predictors*, Machine Learning, 24(2), s. 123-140.

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

Dietterich T.G., 2000, *Ensemble methods in machine learning*, [w:] *Multiple Classifier Systems. First International Workshop*, vol. 1857, Springer-Verlag.

Domeniconi C., Yan B., 2004, *Nearest neighbour ensemble*, IEEE Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004), vol. 1.

Enas G.G., Choi S.C., 1986, *Choice of the smoothing parameter and efficiency of k-nearest neighbor classification*, Computer and Mathematics with Applications, 12A(2), s. 235-244.

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/.

Freund Y., Schapire R.E., 1997, *A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting*, Journal of Computer and System Sciences, no. 55, s. 119-139.

Friedman J.H., Popescu B.E., 2005, *Predictive learning via rule ensembles*, Technical Report, Department of Statistics, Stanford University.

Fumera G., Roli F., 2001, *Error rejection in linearly combined multiple classifiers*, Proceedings of International Workshop on Multiple Classifier Systems, Springer, Cambridge, UK.

Gatnar E., 2008, *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa.

Gul A., Perperoglou A., Khan Z., Mahmoud O., Miftahuddin M., Adler W., Lausen B., 2014, *Ensemble of a subset of kNN classifiers*, Advances in Data Analysis and Classification, s. 1-14.

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

Hechenbichler K., Schliep K.P., 2004, *Weighted k-Nearest-Neighbour 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*, Proceedings AAAI-92, MIT Press.

Kononenko I., 1994, *Estimating attributes: Analysis and extensions of RELIEF*, Proceedings European Conference on Machine Learning.

Kubus M., 2015, *Feature selection and the chessboard problem*, Acta Universitatis Lodziensis, Folia Oeconomica, Statistical Analysis in Theory and Practice, no. 1 (311), s. 17-25.

Kubus M., 2016, *Lokalna ocena mocy dyskryminacyjnej zmiennych*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 427, Taksonomia 27, *Klasyfikacja i analiza danych – teoria i zastosowania*, Wrocław.

Opitz D., Maclin R., 1999, *Popular ensemble methods: An empirical study*, Journal of Artificial Intelligence Research, 11, s. 169-198.

Zhou Z.H., Yu Y., 2005, *Adapt bagging to nearest neighbour classifiers*, Journal of Computer Science and Technology, vol. 20(1), s. 48-54.