PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 468 • 2017

Taksonomia 28
Klasyfikacja i analiza danych – teoria i zastosowania

ISSN 1899-3192
e-ISSN 2392-0041

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

PROBLEM ZMIENNYCH ZAKŁÓCAJĄCYCH W AGREGOWANYCH KLASYFIKATORACH KNN

A PROBLEM OF THE NOISY VARIABLES IN THE AGGREGATED KNN CLASSIFIERS

DOI: 10.15611/pn.2017.468.12
JEL Classification: C01, C14, C52

Streszczenie: Podejście wielomodelowe w dyskryminacji i regresji zyskało duże uznanie ze względu na poprawę stabilności modeli oraz ich dokładności przewidywań. Agregowanie klasyfikatorów *k* najbliższych sąsiadów (kNN) napotyka jednak poważne problemy. Metoda kNN, wykorzystująca w klasyfikacji wyłącznie odległości między obiektami, jest względnie stabilna, przez co zróżnicowanie klasyfikatorów bazowych można osiągnąć, jedynie wybierając różne podprzestrzenie. Tu z kolei napotykamy problem zmiennych zakłócających (*noisy variables*), to jest takich, które nie mają wpływu na zmienną objaśnianą, a które w metodzie kNN znacznie obniżają dokładność klasyfikacji. W artykule dokonano przeglądu zaproponowanych w literaturze metod agregowania klasyfikatorów kNN oraz zweryfikowano je z własną propozycją algorytmu. W badaniach wykorzystano zbiory danych rzeczywistych z dołączonymi zmiennymi zakłócającymi.

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

Summary: Ensemble learning in discrimination and regression has gained a great appreciation due to the improved stability of the model, and often improved accuracy of the predictions. Aggregating of *k* nearest neighbors classifiers (kNN), however, faces serious problems. The kNN method, which uses only the distances between objects, is relatively stable, so the diversity of base classifiers can only be achieved by choosing different subspaces. Here, in turn, we encounter the problem of noisy variables, those that do not affect the response variable, and which result in decreasing the accuracy of the kNN classifier. This article reviews the methods of the aggregated kNN classifiers, which were proposed in the literature. We also verify our own proposition of the algorithm. The real data with noisy variables added are used in the empirical study.

Keywords: ensemble learning, *k* nearest neighbours method, feature selection.

1. Wstęp

Podejście wielomodelowe w dyskryminacji i regresji już od dwudziestu lat cieszy się dużym uznaniem i popularnością. W Polsce ukazała się monografia E. Gatnara [2008] podsumowująca stan wiedzy w tej dziedzinie. Od tego czasu nadal podejmowano wiele prac badawczych w tym obszarze. D. Rozmus [2008] wykorzystuje agregację modeli w taksonomii. A. Dudek [2009] adaptuje to podejście do zmiennych symbolicznych. T. Górecki i M. Krzyśko [2015] łączą klasyfikatory metodami regresyjnymi. Podejście wielomodelowe rozwijane było głównie dla drzew klasyfikacyjnych [Breiman 1996; Freund, Schapire 1997; Breiman 2001], ale można znaleźć w literaturze przedmiotu propozycję łączenia innych klasyfikatorów: sieci neuronowych [Opitz, Maclin 1999], logicznych reguł klasyfikacji [Friedman, Popescu 2005], modeli SVM [Trzęsiok 2006] czy klasyfikatorów *k* najbliższych sąsiadów.

Modele zagregowane można skutecznie stosować, o ile spełnione są pewne warunki. Po pierwsze, modele bazowe muszą cechować się dostateczną dokładnością. Powinna ona być przynajmniej nieco większa od klasyfikacji na podstawie prawdopodobieństw *a priori* (w regresji od średniej wartości empirycznej). Po drugie, modele bazowe powinny być zróżnicowane (tzn. nie mogą identycznie klasyfikować tych samych obserwacji ze zbioru uczącego). K. Tumer i J. Ghosh [1996], a następnie L. Breiman [2001] pokazali, że im bardziej modele bazowe różnią się od siebie (inaczej klasyfikują te same obserwacje ze zbioru uczącego), tym dokładniejsze są wyniki predykcji modelu zagregowanego. Zróżnicowaniu sprzyja niestabilność, dlatego też L. Breiman [1996] rekomendował do podejścia wielomodelowego drzewa klasyfikacyjne lub sieci neuronowe, natomiast odradzał agregowania klasyfikatorów kNN, które są względnie stabilne, gdyż do klasyfikacji używają jedynie odległości między obiektami. Mimo to w literaturze pojawiały się takie propozycje [Ho 1998; Bay 1999; Domeniconi, Yan 2004; Zhou, Yu 2005b; Gul i in. 2014; Kubus 2016b]. Trzy główne strategie mające na celu uzyskanie zróżnicowania klasyfikatorów bazowych polegają na: losowaniu obiektów do prób uczących, rzutowaniu obiektów w podprzestrzenie oraz na zmianie parametrów w modelach bazowych. Pierwsza technika nie sprawdza się w przypadku względnie stabilnej metody kNN. Uwaga badaczy skupiła się więc na wyborze podprzestrzeni. Tu jednak napotykamy problem zmiennych, które nie mają mocy dyskryminacyjnej, nazywanych zmiennymi zakłócającymi (*noisy variables*). Zmienne takie powodują obniżenie dokładności klasyfikacji w metodzie kNN (zob. np. [Kubus 2016a]).

W artykule pokażemy, że zaproponowane w literaturze metody agregowania klasyfikatorów kNN na różne sposoby starają się rozwiązać problem zmiennych zakłócających. Przedstawimy możliwości modyfikacji własnej propozycji agregowanego klasyfikatora kNN zaproponowanego w pracy [Kubus 2016b]. W badaniach empirycznych wykorzystane będą zbiory rzeczywiste oraz sztucznie generowane zmienne bez mocy dyskryminacyjnej. Spróbujemy wykazać, że selekcja zmiennych

jest niezbędnym komponentem algorytmu agregowania klasyfikatorów kNN, decydującym o jego dokładności.

2. Agregowane klasyfikatory kNN

Zastosowanie metody *k* najbliższych sąsiadów w podejściu wielomodelowym napotyka dwa zasadnicze problemy. Po pierwsze, klasyfikatory kNN są względnie stabilne, tzn. niewielkie zmiany w próbie uczącej nie wpływają znacząco na wyniki klasyfikacji [Breiman 1996]. Można sobie wyobrazić, że usunięcie kilku obserwacji ze zbioru uczącego wpływa jedynie lokalnie na niewielką zmianę granic między klasami. W metodzie drzew klasyfikacyjnych usunięcie kilku obserwacji może spowodować nawet wybór innych zmiennych w węzłach, a więc znacząco zmieni się postać modelu. Po drugie, metoda kNN jest wrażliwa na zmienne zakłócające (zob. np. [Kubus 2016a]). Rzutowanie obiektów zbioru uczącego na różne podprzestrzenie jest właściwie jedynym sposobem uzyskania odpowiedniego zróżnicowania klasyfikatorów bazowych. Jeśli jednak w zbiorze danych są zmienne bez mocy dyskryminacyjnej, dokładność klasyfikacji modeli bazowych może gwałtownie spadać, co w efekcie prowadzi do mało dokładnego modelu agregowanego.

Pierwsze propozycje agregowania klasyfikatorów kNN pochodzą od T.K. Ho [1998][1] oraz S.D. Baya [1999]. Polegają one na losowym wyborze *q* zmiennych objaśniających do klasyfikatorów bazowych. S.D. Bay [1999] w swym algorytmie MFS (*Multiple Feature Subsets*) agregował 100 klasyfikatorów jednego najbliższego sąsiada (1NN), przyjmując metrykę euklidesową, głosowanie większościowe i wybierając liczbę *q* za pomocą sprawdzania krzyżowego. Autor zauważył też, że w przypadku, gdy w zbiorze są zmienne bez mocy dyskryminacyjnej, klasyfikator taki może dawać sporo błędnych klasyfikacji. Kolejne metody zaproponowane w literaturze na różne sposoby odpowiadają na problem zmiennych zakłócających. Poniżej zamieszczono ich krótką prezentację.

2.1. Algorytm ENNWDS

C. Domeniconi oraz B. Yan [2004] opracowali algorytm ENNWDS (*Ensemble of Nearest Neighbors in Weight-Driven Subspaces*), w którym nowatorskim pomysłem jest losowanie zmiennych do klasyfikatorów bazowych z różnymi prawdopodobieństwami. Najpierw więc zmiennym przypisywane są wagi, które sumują się do jedynki. W tym celu autorzy stosują algorytm ADAMENN (*Adaptive Metric Nearest Neighbors*) [Domeniconi, Peng, Gunopulos 2002]. Wykorzystywana jest w nim ważona odległość chi-kwadrat między prawdopodobieństwami *a posteriori*. Porównywane są prawdopodobieństwa wystąpienia *j*-tej klasy po zaobserwowaniu obiektu

[1] T.K. Ho zaproponowała także analogiczną metodę RSM (*Random Subspaces Method*) dla drzew klasyfikacyjnych (zob. też [Gatnar 2008]).

w oryginalnej przestrzeni zmiennych, z prawdopodobieństwem dla obiektu zrzutowanego na oś reprezentującą wybraną zmienną. Należy podkreślić, że oceny ważności zmiennych przypisywane są lokalnie, gdyż odległości chi-kwadrat są uśredniane dla najbliższych sąsiadów rozpoznawanego obiektu. Parametrami metody ENNWDS są: liczba klasyfikatorów bazowych, liczba losowanych zmiennych oraz liczba najbliższych sąsiadów. Autorzy ustalili liczbę klasyfikatorów bazowych na 200 oraz metodą jednoczęściowego sprawdzania krzyżowego (*leave one out cross validation*) dobierali kombinację parametrów k oraz q (liczba losowanych zmiennych do klasyfikatorów bazowych). Stosowali ważoną metrykę euklidesową oraz badali trzy schematy agregowania: głosowanie większościowe, uśrednianie prawdopodobieństw *a posteriori* oraz metodę Borda.

2.2. Algorytm FASBIR

Kolejna prezentowana metoda agregowania klasyfikatorów kNN zawiera etap selekcji zmiennych. Jej implementacją jest algorytm FASBIR (*Filtered Attribute Subspace based Bagging with Injected Randomness*) zaproponowany przez Z.H. Zhou i Y. Yu [2005b]. W celu uzyskania jak największego zróżnicowania klasyfikatorów bazowych autorzy zastosowali następujące techniki: losowy wybór zmiennych po wcześniejszej selekcji, losowanie prób bootstrapowych oraz losowy wybór potęgi w metryce Minkowskiego (tab. 1). Oryginalnie algorytm ten badany był dla: 100 klasyfikatorów bazowych, zbioru potęg $P = \{1, 2, 3\}$, wartości progowej t równej 33% średniej wartości przyrostu informacji dla wszystkich zmiennych oraz dla liczby q równej połowie liczby zmiennych po etapie selekcji. Autorzy rozważali liczby sąsiadów k ze zbioru $\{1, 3, 5, 7, 9\}$, nie dając jednak jasnej rekomendacji co do optymalnego wyboru tego parametru.

Tabela 1. Algorytm FASBIR

Ustal parametry modelu:
M – liczbę klasyfikatorów bazowych,
t – wartość progową kryterium oceny ważności zmiennych,
q – liczbę zmiennych losowanych do pojedynczego klasyfikatora,
k – liczbę najbliższych sąsiadów,
P – zbiór wartości potęg w metryce Minkowskiego.
1. Przeprowadź dobór zmiennych, wykorzystując cały zbiór uczący, przyjmując jako kryterium przyrost informacji (*information gain*) i wartość progową t. Uzyskany podzbiór zmiennych oznaczony będzie przez S, $S \subset \boldsymbol{X}$.
2. Z podzbioru S wylosuj q zmiennych. Uzyskaną w ten sposób losową podprzestrzeń oznaczymy przez S_q.
3. W podprzestrzeni S_q wylosuj próbę bootstrapową B.
4. W zbiorze B znajdź k najbliższych sąsiadów rozpoznawanego obiektu $\boldsymbol{x}$ i klasyfikuj go. Zastosuj metrykę Minkowskiego z potęgą losowaną ze zbioru P.
5. Kroki 2-4 powtarzaj M razy, zapamiętując wszystkie klasyfikacje z kroku 4.
6. Dokonaj agregacji za pomocą głosowania większościowego.

Źródło: opracowanie własne na podstawie [Zhou, Yu 2005b].

Prototypem FASBIR jest wcześniej proponowany przez tych samych autorów *Bag-In-Rand* (BIR) [Zhou, Yu 2005a], w którym losowo wybierane są obiekty (bootstrap), zmienne oraz potęgi w metryce Minkowskiego, ale brakuje etapu selekcji zmiennych.

2.3. Algorytm ESKNN

Zupełnie odmienny pomysł na uniknięcie wpływu zmiennych zakłócających mieli A. Gul i in. [2014]. W ich algorytmie ESKNN (*Ensemble of a Subset of kNN classifiers*) zmienne wybierane są losowo, ale jakość klasyfikatorów bazowych jest kontrolowana. Cała procedura przebiega dwuetapowo. Najpierw buduje się na próbach bootstrapowych M klasyfikatorów bazowych w q-wymiarowych losowych podprzestrzeniach. Klasyfikatory są porządkowane według dokładności klasyfikacji szacowanej na zbiorach obiektów, które nie zostały wylosowane do prób uczących (*out of bag*). W następnym kroku wybiera się h najlepszych klasyfikatorów bazowych, które przechodzą do kolejnego etapu sprawdzania jakości modelu. Polega on na ocenie modeli agregowanych, w skład których wchodzi kolejno m najlepszych klasyfikatorów bazowych, gdzie $m \in \{1, ..., h\}$. Klasyfikator bazowy jest włączany do modelu agregowanego, jeśli nastąpi poprawa wartości indeksu Briera:

$$B = \frac{\sum_{i=1}^{n_t} \left(y_i - \hat{p}(y_i | \mathbf{x}_i) \right)^2}{n_t}, \tag{1}$$

gdzie n_t jest liczbą obiektów w zbiorze walidacyjnym. Autorzy w swych badaniach przyjęli: $M = 1000$, $q = p/3$ (gdzie p jest liczbą zmiennych objaśniających), $h = 40\%M$ oraz agregowanie przez głosowanie większościowe. Liczba sąsiadów wybierana była ze zbioru $k \in \{1, ..., 10\}$ przez 10-częściowe sprawdzanie krzyżowe.

2.4. Algorytm AR-1NN

Kolejny algorytm agregujący klasyfikatory kNN zaproponował M. Kubus [2016b]. Jego kluczowym elementem jest selekcja zmiennych. Od opisanego wcześniej algorytmu FASBIR różni się on w następujących punktach. Po pierwsze, selekcja zmiennych jest wykonywana dla każdej próby bootstrapowej oddzielnie. Po drugie, przyjęto inne kryterium ważności zmiennych. Wykorzystano tu algorytm ReliefF [Kononenko 1994], dedykowany metodzie k najbliższych sąsiadów. Po trzecie, wartość progowa tego kryterium jest ustalana na podstawie analizy statystycznej, a nie arbitralnie. W pracy [Kubus 2016a] naświetlono problematykę ustalenia wartości progowej oraz zaproponowano, by zmienne porządkować według oceny ważności algorytmem ReliefF, a następnie metodą sprawdzania krzyżowego oceniać dokładność klasyfikacji modeli zagnieżdżonych. Na podstawie minimalnego błędu klasyfikacji wybierana jest optymalna liczba zmiennych. Należy podkreślić, że w agregowanym klasyfikatorze kNN z selekcją zmiennych (AR-1NN) przyjmuje się liczbę

sąsiadów równą jeden. Wybór ten jest intuicyjny, gdyż w przypadku $k = 1$ uzyskuje się największe dopasowanie do danych ze zbioru uczącego, co wpływa na zmniejszenie stabilności i powinno bardziej zróżnicować modele bazowe. Przyjęcie $k = 1$ powoduje też zmniejszenie czasu obliczeń oraz uniknięcie problemu wyboru optymalnej wartości tego parametru. W przedstawionych powyżej algorytmach k wybierane jest przez sprawdzanie krzyżowe (duży czas obliczeń) lub jego wybór jest arbitralny i brakuje wyraźnych rekomendacji co do jego optymalnego wyboru.

Słabszą stroną zaproponowanego w pracy [Kubus 2016b] algorytmu AR-1NN jest relatywnie duży czas obliczeń. Decyduje o tym przyjęcie liczby iteracji w ReliefF równej połowie liczebności zbioru uczącego [Kubus 2016a] oraz przedstawiony wyżej sposób ustalenia wartości progowej. W niniejszym opracowaniu proponuje się zatem następującą modyfikację. Po pierwsze, ustalenie mniejszej liczby iteracji w algorytmie selekcji zmiennych ReliefF. Wprawdzie w pojedynczym klasyfikatorze może to spowodować wprowadzenie zmiennej zakłócającej i obniżyć jego jakość, lecz w podejściu wielomodelowym może być pożądane ze względu na osiągnięcie większego zróżnicowania. Po drugie, w wyborze wartości progowej kosztowne obliczeniowo sprawdzanie krzyżowe zastąpione będzie oceną na zbiorze walidacyjnym. Takie podejście na ogół powoduje mniejszą stabilność modelu, lecz w kontekście agregacji może pozytywnie wpłynąć na dokładność klasyfikacji.

3. Badania empiryczne

Badania porównawcze algorytmów agregowania klasyfikatorów kNN przeprowadzono, wykorzystując zbiory z repozytorium Uniwersytetu Kalifornijskiego [Frank, Asuncion 2010]: *cardiotocographic* (2126, 21, 3), *ecoli* (336, 7, 8), *glass* (214, 9, 6), *ionosphere* (351, 33, 2), *segmentation* (2310, 19, 7) oraz *sonar* (208, 60, 2). W nawiasach podano kolejno liczby: obiektów, zmiennych objaśniających oraz klas. Do zbiorów rzeczywistych dołączono także zmienne zakłócające, które miały w klasach jednakowe rozkłady $N(0;1)$. We wszystkich eksperymentach wprowadzano ich 10 lub 50 procent oryginalnej liczby zmiennych objaśniających p. Błędy klasyfikacji estymowano 50 razy na zbiorach testowych, przy czym poszczególne klasyfikatory agregowane budowano dla tych samych podziałów na zbiory uczące i testowe. W obliczeniach wykorzystano własne kody oraz pakiet ESKNN programu R. W celu uzyskania porównywalności, we wszystkich badanych algorytmach ustalono jednakowe wartości parametrów. Przyjęto: liczbę najbliższych sąsiadów $k = 1$, liczbę losowanych zmiennych $q \approx \sqrt{p}$, potęgi w metryce Minkowskiego $\{1, 2, 3\}$ oraz liczbę klasyfikatorów bazowych $M = 100$. Wyjątek stanowił algorytm ESKNN, gdzie zgodnie z zaleceniami autorów przyjęto $M = 1000$. Wybór ten potwierdziły też badania własne, których wyniki przedstawiono w tab. 2. Zauważmy, że zdecydowanie mniejsze błędy klasyfikacji uzyskiwano dla jednego najbliższego sąsiada w porównaniu z k wyznaczonym wg sugestii G.G. Enasa i S.C. Choi [1986], to jest $k_N \approx N^{2/8}$.

W tab. 3 pokazano wyniki własnych propozycji algorytmów. Symbolem AR-1NN (1) oznaczono oryginalną wersję algorytmu zaproponowaną w pracy [Kubus 2016b]. Symbol AR-1NN (2) oznacza jego modyfikację przedstawioną w rozdziale 2 niniejszego opracowania. Test rangowanych znaków Wilcoxona wykazywał jej przewagę. Jedynie na zbiorze *ecoli* z jedną zmienną zakłócającą algorytm zmodyfikowany dawał istotnie większe błędy klasyfikacji. Podkreślmy też, że wersja zmodyfikowana jest zdecydowanie szybsza. Algorytm AR-1NN (2) stosowano także z dodatkowo wprowadzoną kontrolą jakości klasyfikatorów bazowych lub z losową potęgą w metryce Minkowskiego. Nie uzyskano jednak zwiększenia dokładności klasyfikacji. Zbadano też przypadek $k \approx N^{2/8}$, gdzie otrzymano na ogół większe błędy klasyfikacji.

Tabela 2. Błędy klasyfikacji (w %) w metodzie ESKNN dla różnych wartości parametrów. Błędy estymowano 50 razy na zbiorach testowych i uśredniono. Do oryginalnych zbiorów dołączono zmienne zakłócające z $N(0;1)$

Zbiory z dołączonymi zmiennymi zakłócającymi 10%	$M = 100$ $k = k_N$	$M = 500$ $k = k_N$	$M = 1000$ $k = k_N$	$M = 100$ $k = 1$	$M = 500$ $k = 1$	$M = 1000$ $k = 1$
cardiotocographic	10,9 (0,2)	9,1 (0,2)	8,9 (0,1)	9,8 (0,3)	8 (0,1)	7,6 (0,1)
ionosphere	10,1 (0,4)	9,3 (0,4)	8,2 (0,4)	9,4 (0,4)	7,7 (0,4)	7 (0,4)
segmentation	9,5 (0,5)	11,2 (0,6)	11,3 (0,6)	9,5 (0,8)	9,8 (0,7)	9,7 (0,6)
sonar	23,9 (0,8)	17,7 (0,7)	16,8 (0,7)	21,9 (0,9)	17,8 (0,7)	14,8 (0,7)
Zbiory z dołączonymi zmiennymi zakłócającymi 50%	$M = 100$ $k = k_N$	$M = 500$ $k = k_N$	$M = 1000$ $k = k_N$	$M = 100$ $k = 1$	$M = 500$ $k = 1$	$M = 1000$ $k = 1$
cardiotocographic	13,3 (0,4)	11,7 (0,3)	11,1 (0,2)	12,4 (0,3)	9,9 (0,2)	9,5 (0,2)
ionosphere	11,9 (0,5)	10,4 (0,4)	9,7 (0,4)	11 (0,5)	8,4 (0,3)	7,7 (0,4)
segmentation	18,6 (1,1)	17,6 (1,0)	18,2 (1,2)	13,9 (1,0)	16 (1,0)	16,4 (0,9)
sonar	28,4 (0,9)	24 (0,8)	20,5 (0,7)	29,2 (1,0)	21,2 (0,7)	17,9 (0,7)

Źródło: obliczenia własne.

Porównanie błędów klasyfikacji dla metod opisanych w punkcie 2 ilustruje tab. 4. Algorytm AR-1NN zastosowano w zmodyfikowanej wersji. W przypadku zbiorów *ecoli* oraz *glass* wystąpiły trudności implementacyjne w pakiecie ESKNN i nie uzyskano wyników. W algorytmie ENNWDS ważono zmienne za pomocą algorytmu ReliefF. W algorytmach, które nie wykorzystują ani selekcji zmiennych, ani ich losowania z różnymi prawdopodobieństwami (MFS, BIR, ESKNN) widoczny jest radykalny wzrost błędów klasyfikacji po wprowadzeniu większej liczby zmiennych zakłócających. Taka zależność występuje nawet w algorytmie FASBIR wykorzystującym selekcję zmiennych. Zastosowanie jednostronnego testu rangowanych znaków Wilcoxona pokazało, że dla wszystkich zbiorów, z wyjątkiem *segmenta-*

tion, wzrost błędu klasyfikacji dla algorytmu FASBIR był istotny na poziomie 0,05. Otrzymano następujące wartości *p*: *cardiotocographic* 0,00274; *ecoli* 0,00000; *glass* 0,00054; *ionosphere* 0,04859; *sonar* 0,00001.

Najmniejsze błędy klasyfikacji uzyskiwano przeważnie autorskim algorytmem AR-1NN. Wyjątek stanowiły: zbiór *cardiotocographic*, gdzie najmniejszy błąd zwracał algorytm ENNWDS z ważeniem zmiennych algorytmem ReliefF, oraz zbiór *ecoli* ze zmiennymi zakłócającymi w liczbie 10% oryginalnych zmiennych objaśniających, gdzie najmniejszy błąd zwracał algorytm FASBIR. Do oceny istotności różnic w błędach dla algorytmów AR-1NN, FASBIR oraz ENNWDS zastosowano test Friedmana. Hipotezy zerowej nie odrzucono jedynie w przypadku zbioru *glass*, w którym dołączono 10% zmiennych zakłócających (tab. 5). Dalsza analiza post-hoc testem Nemenyi (zob. [Demsar 2006]) pokazała wyższość algorytmu AR-1NN nad ENNWDS (z ReliefF) we wszystkich zbiorach z wyjątkiem *cardiotocographic* (oraz *glass*, gdzie nie było różnicy statystycznie istotnej). Porównanie z FASBIR także wychodzi na korzyść proponowanego algorytmu, zwłaszcza w przypadku, gdy wprowadzano więcej zmiennych zakłócających. Wówczas dla zbioru *cardiotocographic* różnica błędów klasyfikacji nie była istotna, a dla pozostałych zbiorów AR-1NN dawał istotnie mniejsze błędy klasyfikacji.

Tabela 3. Błędy klasyfikacji (w %) dla klasyfikatora AR-1NN (propozycja własna) i kilku jego modyfikacji. Błędy estymowano 50 razy na zbiorach testowych i uśredniono. Do oryginalnych zbiorów dołączono zmienne zakłócające z $N(0;1)$

Zbiory z dołączonymi zmiennymi zakłócającymi 10%	AR-kNN	AR-1NN (1)	AR-1NN (2)	AR-1NN (2) + ocena	AR-1NN (2) + los_M
cardiotocographic	10,2 (0,1)	8,4 (0,1)	8,4 (0,1)	8,4 (0,1)	8,9 (0,1)
ecoli	16,5 (0,5)	16,5 (0,5)	18,2 (0,5)	18,2 (0,5)	16,9 (0,5)
glass	23,8 (0,6)	22,6 (0,7)	22,6 (0,6)	22,5 (0,6)	23,6 (0,6)
ionosphere	7,8 (0,3)	6,2 (0,3)	5,7 (0,3)	5,7 (0,3)	6,7 (0,3)
segmentation	3,6 (0,1)	3 (0,1)	3 (0,1)	3 (0,1)	2,7 (0,1)
sonar	14,2 (0,6)	13,1 (0,6)	13,3 (0,6)	13,5 (0,6)	14,3 (0,6)
Zbiory z dołączonymi zmiennymi zakłócającymi 50%	AR-kNN	AR-1NN (1)	AR-1NN (2)	AR-1NN (2) + ocena	AR-1NN (2) + los_M
cardiotocographic	10 (0,1)	8,7 (0,1)	8,2 (0,1)	8,2 (0,1)	8,9 (0,1)
ecoli	16,5 (0,4)	17,3 (0,4)	17,9 (0,4)	17,9 (0,4)	17,6 (0,4)
glass	25,1 (0,6)	23,4 (0,6)	22,9 (0,6)	22,9 (0,6)	25,3 (0,6)
ionosphere	8,6 (0,4)	6,6 (0,3)	6 (0,3)	6 (0,3)	7,4 (0,3)
segmentation	3,6 (0,1)	2,9 (0,1)	3,1 (0,1)	3,1 (0,1)	2,8 (0,1)
sonar	14,3 (0,6)	13,8 (0,6)	12,1 (0,6)	12,2 (0,6)	14,7 (0,7)

Źródło: obliczenia własne.

Tabela 4. Błędy klasyfikacji (w %) dla agregowanych klasyfikatorów kNN estymowane 50 razy na zbiorach testowych. Do oryginalnych zbiorów dołączono zmienne zakłócające z $N(0;1)$

Zbiory z dołączonymi zmiennymi zakłócającymi 10%	MFS	BIR	FASBIR	ESKNN	ENNWDS ReliefF	AR-1NN
cardiotocographic	10 (0,2)	11,1 (0,2)	8 (0,1)	7,6 (0,1)	7,3 (0,1)	8,4 (0,1)
ecoli	18,9 (0,5)	18,3 (0,5)	17 (0,4)	–	21,8 (0,5)	18,2 (0,5)
glass	26,3 (0,5)	32 (0,6)	23,8 (0,6)	–	23,1 (0,7)	22,6 (0,6)
ionosphere	6,7 (0,3)	14,4 (0,4)	6,3 (0,3)	7 (0,4)	6,1 (0,3)	5,7 (0,3)
segmentation	4,2 (0,1)	7,4 (0,1)	4,1 (0,1)	9,7 (0,6)	3,7 (0,1)	3 (0,1)
sonar	14,1 (0,6)	15,3 (0,5)	14,2 (0,6)	14,8 (0,7)	15,4 (0,6)	13,3 (0,6)
Zbiory z dołączonymi zmiennymi zakłócającymi 50%	MFS	BIR	FASBIR	ESKNN	ENNWDS ReliefF	AR-1NN
cardiotocographic	14,3 (0,2)	13,5 (0,2)	8,4 (0,3)	9,5 (0,2)	7,4 (0,1)	8,2 (0,1)
ecoli	26,2 (0,5)	23 (0,4)	21,7 (0,7)	–	21,1 (0,5)	17,9 (0,4)
glass	33,7 (0,7)	45,4 (0,7)	26,3 (0,8)	–	23,1 (0,7)	22,9 (0,6)
ionosphere	8,7 (0,4)	15,7 (0,4)	6,5 (0,3)	7,7 (0,4)	6,5 (0,3)	6 (0,3)
segmentation	7,9 (0,2)	17,3 (0,2)	4 (0,1)	16,4 (0,9)	3,9 (0,1)	3,1 (0,1)
sonar	18,2 (0,6)	17,8 (0,6)	17,3 (0,6)	17,9 (0,7)	14,5 (0,6)	12,1 (0,6)

Źródło: obliczenia własne.

Tabela 5. Statystyczna istotność różnicy błędów klasyfikacji. W kolumnach przedstawiono różnice średnich rang między algorytmem AR-1NN a algorytmami FASBIR oraz ENNWDS z ReliefF. Wartości ujemne mówią o większej dokładności algorytmu AR-1NN. Ich moduły porównywane są z różnicą krytyczną 0,4686 dla poziomu istotności 0,05

AR-1NN	10% zmiennych zakłócających			50% zmiennych zakłócających		
	FASBIR	ENNWDS ReliefF	Test Friedmana wartość *p*	FASBIR	ENNWDS ReliefF	Test Friedmana wartość *p*
cardiotocographic	**0,68**	**1,36**	**0,00000**	0,03	**1,05**	**0,00000**
ecoli	0,46	**–0,88**	**0,00000**	**–0,81**	**–0,87**	**0,00000**
glass	–	–	0,27645	**–0,68**	–0,01	**0,00021**
ionosphere	**–0,70**	**–0,47**	**0,00017**	**–0,67**	**–0,68**	**0,00002**
segmentation	**–1,41**	**–0,75**	**0,00000**	**–1,22**	**–1,21**	**0,00000**
sonar	–0,31	**–0,53**	**0,01947**	**–1,19**	**–0,55**	**0,00000**

Źródło: obliczenia własne.

4. Podsumowanie

W artykule zaprezentowano przegląd algorytmów agregujących klasyfikatory kNN wraz z własną propozycją. Zaakcentowano, że głównym problemem dla tych metod są zmienne zakłócające, to jest bez mocy dyskryminacyjnej. Podkreślono też korzyści z ustalenia liczby sąsiadów równej jeden. Badania empiryczne wykazały, że selekcja zmiennych jest niezbędnym komponentem algorytmu agregowania klasyfikatorów kNN, decydującym o jego dokładności. Warto zwrócić uwagę na adekwatność wyboru kryterium oceniającego zmienne do metody kNN. Tylko w algorytmach AR-1NN oraz w ENNWDS-ReliefF nie wystąpił wzrost błędu klasyfikacji wraz z liczbą zmiennych zakłócających.

Literatura

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

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

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

Demsar J., 2006, *Statistical comparison of classifiers over multiple data sets*, Journal of Machine Learning Research, vol. 7, s. 1-30.

Domeniconi C., Peng J., Gunopulos D., 2002, *Locally adaptive metric nearest neighbor classification*, IEEE Trans on Pattern Analysis and Machine Intelligence, vol. 24(9), s. 1281-1285.

Domeniconi C., Yan B., 2004, *Nearest neighbor ensemble*, IEEE Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004), vol. 1, s. 228-231.

Dudek A., 2009, *Tworzenie zagregowanych modeli dyskryminacyjnych dla obiektów symbolicznych: wybrane problemy*, Studia i Prace Uniwersytetu Ekonomicznego w Krakowie, nr 3, s. 33-40.

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*, University of California, School of Information and Computer Science, Irvine, CA, http://archive.ics.uci.edu/ml/ (2.02.2016).

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.

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

Górecki T., Krzyśko M., 2015, *Regression Methods for Combining Multiple Classifiers*, Communications in Statistics-Simulation and Computation, vol. 44(3), s. 739-755.

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, DOI 10.1007/s11634-015-0227-5.

Ho T.K., 1998, *Nearest neighbors in random subspaces*, Proceedings of the Second International Workshop on Statistical Techniques in Pattern Recognition, Sydney, Australia, s. 640-648.

Kononenko I., 1994, *Estimating attributes: Analysis and extensions of RELIEF*, Proceedings European Conference on Machine Learning, s. 171-182.

Kubus M., 2016a, *Lokalna ocena mocy dyskryminacyjnej zmiennych*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 427, Taksonomia 27, s. 143-152, DOI: 10.15611/pn.2016.427.15.

Kubus M., 2016b, *Propozycja agregowanego klasyfikatora kNN z selekcją zmiennych*, Ekonometria, nr 3 (53), s. 32-41, DOI: 10.15611/ekt.2016.3.03.

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

Rozmus D., 2008, *Wykorzystanie podejścia zagregowanego w taksonomii*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, nr 7 (1207), Taksonomia 15, s. 330-336.

Trzęsiok M., 2006, *Łączenie równoległe modeli klasyfikacji otrzymanych metodą wektorów nośnych*, Studia Ekonomiczne, Akademia Ekonomiczna w Katowicach, nr 39, s. 129-137.

Tumer K., Ghosh J., 1996, *Error correlation and error reduction in ensemble classifiers*, Connection Science, vol. 8(3-4), s. 385-403.

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

Zhou Z.H., Yu Y., 2005b, *Ensembling local learners trough multimodal perturbation*, IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics, vol. 35(4), s. 725-735.