AUTOMATYKA • 2005 • Tom 9 • Zeszyt 3

Zofia Stawska*, Adam Jóźwik**

# Klasyfikatory z gradacją błędów wykorzystujące sumę rang

## 1. Wstęp

Statystyczne metody rozpoznawania obrazów są znane i wykorzystywane z powodzeniem od wielu lat. Jedną z najpopularniejszych jest metoda *k* najbliższych sąsiadów [1], która doczekała się wielu wariantów i modyfikacji. Swoją ciągłą popularność zawdzięcza ona kilku podstawowym zaletom:

- prostocie implementacji,
- na ogół wysokiej jakości klasyfikacji,
- szybkiemu procesowi uczenia.

Konstrukcja klasyfikatora *k*-NN polega na wyznaczeniu wartości *k*, która oferuje najmniejsze prawdopodobieństwo mylnej klasyfikacji. Prawdopodobieństwo to możemy oszacować za pomocą zbioru testującego lub estymować metodą „minus jednego elementu", polegającą na klasyfikacji każdego obiektu ze zbioru odniesienia przez regułę *k*-NN wyprowadzoną ze zbioru odniesienia pomniejszonego o obecnie klasyfikowany obiekt. Prawdopodobieństwo mylnej decyzji estymujemy, obliczając stosunek liczby punktów mylnie klasyfikowanych do liczebności zbioru uczącego. Jest ono liczone jako średni błąd klasyfikacji.

Z oszacowanego błędu mylnej decyzji wynika prawdopodobieństwo poprawnej decyzji, które jest zależne od wartości cech obiektów zbioru uczącego, a nie zależy od wartości cech obiektu klasyfikowanego. Z tego powodu dla pewnych obiektów klasyfikowanych będzie ono przeszacowane, a dla niektórych obiektów klasyfikowanych niedoszacowane, gdyż jest wartością oczekiwaną frakcją obiektów poprawnie klasyfikowanych. Dzieje się to z powodu niewykorzystania informacji o obiekcie klasyfikowanym.

Reguła *k*-NN daje w zasadzie możliwość oszacowania prawdopodobieństw przynależności klasyfikowanego obiektu do każdej z rozważanych klas, czyli prawdopodobieństw

* Katedra Informatyki Stosowanej, Politechnika Łódzka. Badania finansowane z grantu promotorskiego nr 3T11C05826

** Instytut Biocybernetyki i Inżynierii Biomedycznej, Polska Akademia Nauk; Katedra Informatyki Stosowanej, Politechnika Łódzka

$p(j/\boldsymbol{x}) \approx k_j/k$, gdzie $j$ oznacza numer klasy, $\boldsymbol{x}$ jest wektorem cech rozpoznawanego obiektu, a $k_j$ jest liczbą obiektów z klasy $j$ wśród jego $k$ najbliższych sąsiadów. Klasyfikator $k$-NN wskaże na klasę $i$, dla której $k_i/k = \max_j k_j/k$. Skoro $p(i/\boldsymbol{x}) = k_i/k$ jest oszacowaniem prawdopodobieństwa, że klasyfikowany obiekt pochodzi faktycznie z przyporządkowanej mu klasy $i$, to oszacowanie prawdopodobieństwa, że podjęta decyzja wskazująca na klasę $i$ jest mylna, wyraża się wzorem $er(i/\boldsymbol{x}) = 1–k_i/k$. Oszacowane prawdopodobieństwo mylnej klasyfikacji jest więc teraz zależne od klasyfikowanego obiektu. Tak więc reguła $k$-NN umożliwia teoretycznie oszacowanie wiarygodności $p(i/\boldsymbol{x})$ decyzji wskazującej na klasę $i$ lub ryzyka pomyłki $er(i/\boldsymbol{x})$ w zależności od cech klasyfikowanego obiektu. Od nich zależy, które z obiektów zbioru uczącego znajdą się wśród $k$ najbliższych sąsiadów obiektu klasyfikowanego.

Przedstawione wyżej podejście do oszacowania wiarygodności decyzji klasyfikatora miałoby praktyczne znaczenie pod warunkiem stosowania dużych wartości parametru $k$. Jednakże wartości $k$ dobierane są tak, by zminimalizować frakcję błędów, która stanowi oszacowanie prawdopodobieństwa mylnej decyzji, i zwykle są zbyt małe, by proporcję $k_j/k$ można było zaakceptować jako dostateczne przybliżenie prawdopodobieństwa mylnej klasyfikacji.

W artykule zaproponowano schemat podejmowania decyzji polegający na wprowadzeniu gradacji błędów, co umożliwi oszacowanie prawdopodobieństwa mylnej decyzji w zależności od cech klasyfikowanego obiektu.

## 2. Schemat podejmowania decyzji z gradacją błędów

Problem wprowadzenia do schematu klasyfikacji gradacji błędów był już rozważany przez autorów tej pracy we wcześniejszych publikacjach. Rozważana była wielostopniowa reguła oparta na wyznaczaniu obszarów klas w przestrzeni cech [2, 3] oraz metoda gradacji błędów oparta na mierze pozycyjnej i pewne jej modyfikacje [4].

Punktem wyjścia dla przedstawionego w niniejszej pracy klasyfikatora jest klasyczna reguła $k$-NN połączona ze stosowanym w testach statystycznych sumowaniem rang, czyli pewnych wag nadanych pomiarom. Metoda $k$-NN przyporządkowuje klasyfikowany obiekt do tej klasy, która jest najliczniej reprezentowana wśród jego $k$ najbliższych sąsiadów, nie uwzględniając jednak ich rozkładu w przestrzeni cech – o wyborze klasy decyduje większość. Przyjmując, że najbardziej znaczące dla procesu rozpoznawania są obiekty leżące najbliżej klasyfikowanej próbki, można nadać kolejnym sąsiadom klasyfikowanego obiektu, uporządkowanym według odległości, rangi od $k$ do 1 (pierwszy najbliższy sąsiad otrzymuje najwyższą rangę, każdy kolejny coraz niższą). Następnie obliczamy sumy rang dla poszczególnych klas i na tej podstawie klasyfikujemy rozpoznawany obiekt do tej klasy, która uzyskała największą sumę rang. Pozwala to na uwzględnienie dwóch kryteriów – bierzemy pod uwagę nie tylko przewagę liczebną pewnej klasy wśród $k$ najbliższych sąsiadów badanej próbki, ale również ich rozkład czyli ich „bliskość" względem klasyfikowanego punktu. Cechą przemawiającą dodatkowo na korzyść takiego podejścia jest możliwość określenia za pomocą sumy rang stopnia wiarygodności decyzji podjętej dla danego obiektu. Wraz ze wzrostem przewagi klasy przyporządkowanej nad pozostałymi klasami, pod względem sumy rang, maleje prawdopodobieństwo mylnej decyzji.

Algorytm z sumą rang można zdefiniować następująco:

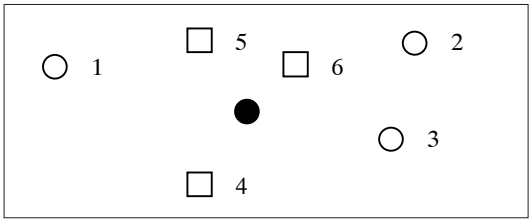
1. Dla każdego z klasyfikowanych obiektów $\boldsymbol{x}$ wyznaczamy jego $k$ najbliższych sąsiadów zgodnie z regułą $k$-NN, nadając im jednocześnie odpowiednie rangi $r_j$, dla $j = k, k–1, ..., 1$. Jak już wcześniej wspomniano najbliższy sąsiad otrzymuje najwyższą rangę równą $k$, kolejny sąsiad otrzymuje rangę $k$–1 itd.
2. Następnie dla każdej klasy $i$ obliczamy sumę rang $r_i$ spośród obiektów należących do $k$ najbliższych sąsiadów klasyfikowanego obiektu.
3. Obiekt $\boldsymbol{x}$ przypisujemy do tej klasy $i$, dla której suma rang $r_i$ ma największą wartość.

Przypisując obiekt do odpowiedniej klasy, możemy jednocześnie określić stopień wiarygodności podjętej decyzji jako

$$sw_x = \frac{r_i}{R}, \text{ gdzie } R = \sum_{j=1}^{k} r_j \tag{1}$$

Stopień wiarygodności jest tym większy, im liczniej reprezentowana jest klasa $i$ w najbliższym sąsiedztwie obiektu $\boldsymbol{x}$, ale jednocześnie zależy od rozmieszczenia reprezentantów tej klasy – może być większa dla mniejszej liczby bliżej położonych sąsiadów niż większości $k$ sąsiadów leżących dalej od rozważanego obiektu. Zaproponowany stopień wiarygodności nie ma sensu prawdopodobieństwa poprawnej decyzji, lecz został wprowadzony tylko jako miara służąca do podziału podejmowanych decyzji na grupy, w których *a priori* spodziewamy się różnych frakcji błędów klasyfikacji.

Rysunek 1 ilustruje działanie algorytmu z sumą rang dla $k = 6$.




**Rys. 1.** Reguła $k$-NN z sumą rang ($k = 6$).
Klasyfikowany obiekt zostanie przypisany do klasy □

## 2.1. Wybór optymalnego *k*

Do tej pory nie zostało sprecyzowane, w jaki sposób w prezentowanej metodzie określamy liczbę $k$. W zależności od przyjętego kryterium istnieją dwie możliwości wyznaczenia optymalnej wartości $k$.

Pierwsza z nich to (zgodnie z oryginalną regułą $k$-NN) wyznaczenie optymalnego $k$ w taki sposób, aby zminimalizować globalne prawdopodobieństwo mylnej decyzji. W fazie uczenia przeglądamy kolejno wszystkie wartości $k = 1 ,..., m$, gdzie $m$ – liczba obiektów zbioru uczącego, poszukując takiego $k$, dla którego osiągniemy minimum ze względu na liczbę błędnych decyzji. Faza ucząca jest wówczas identyczna jak w przypadku reguły $k$-NN.

Druga możliwość to wykorzystanie do wyznaczenia optymalnego $k$ sumy rang. W tym przypadku modyfikacji ulega nie tylko faza klasyfikacji, ale również faza ucząca reguły $k$-NN. Kryterium optymalności może być w tym przypadku dwojakie. Można minimalizować globalną liczbę błędnych decyzji lub przyjąć jako kryterium optymalności np. maksymalną liczbę decyzji pewnych lub obiektów klasyfikowanych z pewną zadaną wiarygodnością.

Przedział poszukiwań optymalnego $k$ nie może być dowolny. Dla $k = 1$ klasyfikator z sumą rang sprowadza się do standardowej reguły 1-NN. Dla niewielkich wartości $k > 1$ sumy rang dla poszczególnych klas przyjmują tak niewiele różnych wartości, że trudno na tej podstawie szacować stopień przynależności obiektu do danej klasy. Z kolei przy bardzo dużych wartościach $k$ (przekraczających liczebność danej klasy) nie można szacować stopnia wiarygodności decyzji za pomocą proponowanej reguły ze względu na zbyt duży udział, wśród $k$ najbliższych sąsiadów, obiektów należących do innych klas niż klasa przypisywana próbce. Na podstawie doświadczeń empirycznych jako dolną granicę przedziału poszukiwań optymalnego $k$ przyjęto pierwiastek z liczby elementów zbioru uczącego, górną granicą była średnia liczba elementów należących do poszczególnych klas.

## 3. Wyniki

Przedstawiony algorytm zostanie przetestowany na dwóch zbiorach medycznych Pima i Bupa zaczerpniętych z repozytorium Uniwersytetu Kalifornijskiego w Irvine [5]. Są to zbiory powszechnie znane i często wykorzystywane w literaturze dotyczącej rozpoznawania obrazów. Pierwszy z nich dotyczy rozpoznawania symptomów cukrzycy – zbiór zawiera 768 obiektów opisanych ośmioma cechami, podzielonych na dwie klasy. Drugi zbiór związany jest z rozpoznawaniem chorób wątroby, zawiera 345 obiektów opisanych sześcioma atrybutami, podzielonych na dwie klasy.

Rozważane były trzy typy klasyfikatora wykorzystującego sumę rang.

Pierwszym badanym klasyfikatorem był klasyczny algorytm $k$-NN wykorzystujący sumę rang wyłącznie do określenia stopnia wiarygodności decyzji dla każdego z klasyfikowanych obiektów. Uczenie i klasyfikacja przebiegały w tym przypadku analogicznie jak w regule $k$-NN. Optymalna wartość $k$ była wybierana w taki sposób, aby zminimalizować globalny błąd klasyfikacji.

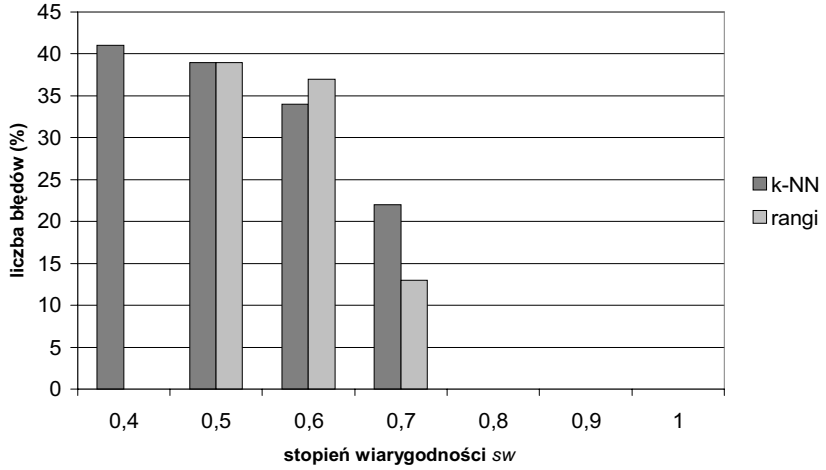
Drugim rozważanym rozwiązaniem był klasyfikator, w którym zarówno uczenie, jak i klasyfikacja zostały przeprowadzone przy użyciu sumy rang. Optymalna wartość $k$ była w tym przypadku, podobnie jak poprzednio, wybierana w taki sposób, aby zminimalizować globalny błąd klasyfikacji.

Trzeci zastosowany algorytm jest modyfikacją drugiej metody polegającą na wprowadzeniu innego kryterium doboru $k$. Jest ono wybierane tak, aby osiągnąć maksymalną liczbę decyzji pewnych.

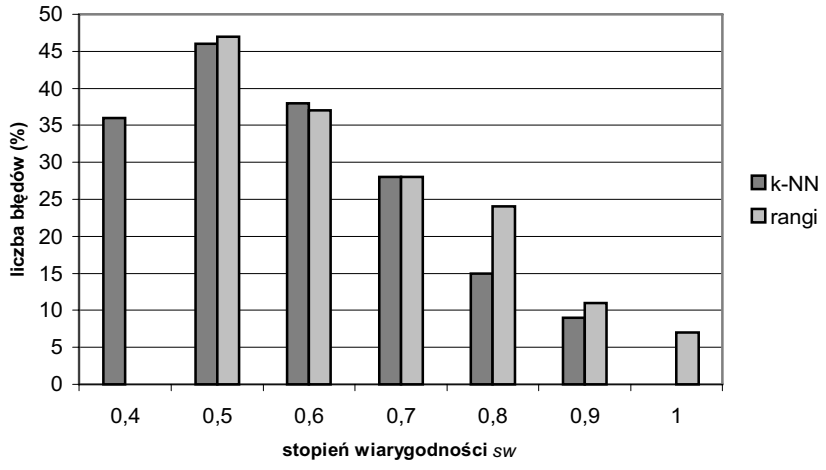
Na podstawie stopnia wiarygodności decyzji wyznaczonego za pomocą reguły (1) zbiór odniesienia został podzielony na dziesięć przedziałów wiarygodności i dla każdego z nich została osobno wyznaczona liczba błędnych decyzji. Oczywiście liczba przyjętych przedziałów wiarygodności może być dowolna. Przyjęty sposób podziału zależy wyłącznie od liczby stopni wiarygodności, które chcielibyśmy uzyskać.

Otrzymane wyniki prezentują rysunki 2 (zbiór Bupa) i 3 (zbiór Pima). Przedstawiają one rozkład błędnych decyzji w zależności od wartości $sw_x$ (stosunek sumy rang dla danej klasy do całkowitej sumy rang) dla dwóch pierwszych metod: klasyfikator oparty na regule *k*-NN (słupki ciemniejsze) i klasyfikator oparty na sumie rang (słupki jaśniejsze). Poszczególne słupki przedstawiają procentowy udział błędnych decyzji wśród wszystkich decyzji podjętych z danym stopniem wiarygodności.

Pierwsze trzy przedziały nie zawierały żadnych obiektów, zatem zostały one na wykresie pominięte.




**Rys. 2.** Zbiór Bupa. Udział błędów w poszczególnych przedziałach wiarygodności




**Rys. 3.** Zbiór Pima. Udział błędów w poszczególnych przedziałach wiarygodności

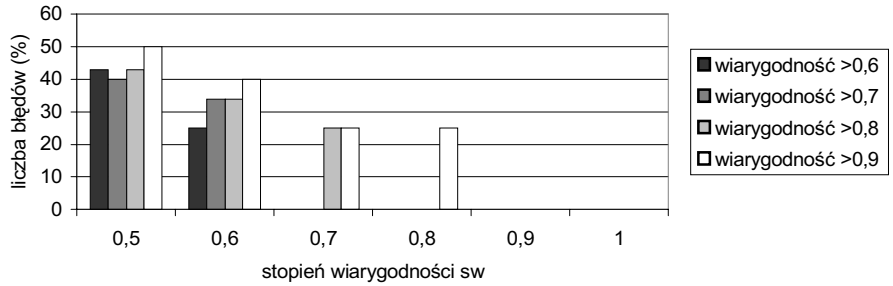
Zgodnie z oczekiwaniami liczba błędów maleje wraz ze wzrostem wartości stopnia wiarygodności $sw_x$. Dzięki tej własności możemy klasyfikować pewne obiekty z dużą pewnością, a inne z większym prawdopodobieństwem błędu. Jak łatwo zauważyć, na wspomnianych rysunkach, wśród obiektów o większej wartości stopnia wiarygodności, liczba błędów klasyfikacji jest znacznie mniejsza niż dla całego zbioru. Zwiększa się ona w miarę zmniejszania się wartości stopnia wiarygodności i dla wartości około 0,5 staje się większa niż globalna liczna błędów.

Dla dwóch prezentowanych metod otrzymaliśmy nieco inny rozkład liczby błędów w poszczególnych przedziałach. Dla reguły *k*-NN otrzymujemy nieco większy błąd dla stopnia wiarygodności z przedziału < 0,3;0,7> ale jednocześnie niższy błąd globalny (około 35,5%). Zastąpienie reguły *k*-NN sumą rang daje nieco lepszy rozkład błędów w poszczególnych przedziałach – otrzymujemy więcej decyzji z mniejszym prawdopodobieństwem błędów, ale dzieje się to kosztem błędu globalnego, który jest nieco wyższy (około 37%). Warto zauważyć, że wzrost błędu globalnego jest na tyle niewielki, że zastosowanie sumy rang zamiast reguły *k*-NN wydaje się w tym przypadku rozwiązaniem korzystniejszym (również ze względu na nieco mniejszy koszt obliczeniowy).

W przypadku zbioru Pima (rys. 3) różnice w otrzymanych wynikach są znacznie większe. Stosując regułę *k*-NN otrzymujemy błąd globalny około 25%, zastosowanie sumy rang zwiększa ten błąd do około 33%. Rozkład błędów w poszczególnych przedziałach wiarygodności jest również korzystniejszy w przypadku zastosowania standardowej reguły *k*-NN z sumą rang służącą jedynie do określenia stopnia wiarygodności decyzji.

Trzecim rozważanym rozwiązaniem było wprowadzenie innego kryterium doboru wartości optymalnego *k*. Było ono w tym przypadku wybierane w ten sposób, aby uzyskać maksymalną liczbę obiektów klasyfikowanych z zadaną wiarygodnością, czyli tak, aby błąd klasyfikacji dla tych obiektów nie przekraczał określonej wartości. Modyfikacja ta została wprowadzona do algorytmu opartego na sumie rang, ale można to kryterium zastosować również dla algorytmu opartego na klasycznej regule *k*-NN.

Na rysunku 4 przedstawiono zależność rozkładu błędów od przyjętego kryterium optymalizacji wartości *k* dla zbioru Bupa. Rozważane były cztery wartości zadanej wiarygodności decyzji. Współczynnik *k* optymalizowany był w każdym z przypadków w ten sposób, aby uzyskać maksymalną liczbę obiektów klasyfikowanych ze stopniem wiarygodności nie mniejszym niż zadana wartość.




**Rys. 4.** Zbiór Bupa. Zależność rozkładu błędów od przyjętego kryterium optymalizacji wartości *k*

Jak łatwo zauważyć, rozkład błędów w poszczególnych przedziałach zależny jest od przyjętego poziomu wiarygodności decyzji, optymalizującego *k*. Globalny błąd klasyfikacji wzrasta wraz ze wzrostem przyjętego poziomu wiarygodności decyzji od 38% dla wiarygodności większej od 0,6 do około 40%, dla wiarygodności większej od 0,9. Jednocześnie otrzymujemy jednak również wzrost liczby decyzji o większym stopniu pewności. Takie rozwiązanie jest znacznie elastyczniejsze niż poprzednie omawiane metody i pozwala dobrać taką wartość *k*, aby uzyskać optymalny kompromis między globalnym błędem klasyfikacji i liczbą obiektów, które zostały zaklasyfikowane z pewną zadaną wiarygodnością.

## 4. Podsumowanie

Przeprowadzone badania wskazują na celowość użycia zaproponowanej metody do wyznaczania stopnia wiarygodności decyzji podejmowanej przez klasyfikator dla każdego rozpoznawanego obiektu. Szczególną zaletą tego rozwiązania jest jego prostota i co za tym idzie – łatwość implementacji. Algorytm z sumą rang może nie tylko zastąpić standardową metodę *k*-NN, może również stanowić uzupełnienie lub modyfikację każdej metody opartej na regule *k* najbliższych sąsiadów. Wyniki przedstawione w niniejszej pracy, jak również wyniki z poprzednich prac przemawiają za celowością dalszego rozwoju metod określania stopnia wiarygodności zależnego od klasyfikowanego obiektu.

Zaprezentowana w pracy metoda wprowadzenia do procesu klasyfikacji gradacji błędów jest jednym z wielu możliwych rozwiązań. Prace nad nowymi sposobami określania stopnia wiarygodności decyzji będą w przyszłości kontynuowane.

## Literatura

[1] Fix E., Hodges J.L.: *Disriminatory Analysis: Nonparametric Discrimination Small Sample Performance.* From project 21-49-004, Report Number 11, USAF School of Aviation Medicine, Randolph Field, Texas, 1952, 280–322

[2] Jóźwik A., Stawska Z.: *Wielostopniowy klasyfikator typu najbliższy sąsiad z każdej klasy*. Materiały VIII Konferencji „Sieci i Systemy Informatyczne", Łódź, 2000, 339–346

[3] Stawska Z., Jóźwik A., Grabowski M., Filipiak K., Rudowski R., Opolski G.: *A multistage classifier based on distance measure and its use for detection of respiration pathology*. Materiały III Konferencji „Komputerowe Systemy Rozpoznawania KOSYR2003", Wrocław 2003, 369–375

[4] Stawka Z., Jóźwik A.: *Klasyfikatory z oceną wiarygodności decyzji i ich zastosowanie do wyboru strategii leczenia*. Półrocznik AGH Automatyka, 2004, 331–338

[5] Merz Ch., Murphy P.M.: *UCI repository of machine learning database*. 1996, [http://www.ics.uci.edu/~mlearn/MLRepository.html]