AUTOMATYKA • 2006 • Tom 10 • Zeszyt 3

Artur Sierszeń*, Łukasz Sturgulewski*

Kondensacja zbioru odniesienia metodą punktów wzajemnie najdalszych jako system sterowania pomiędzy szybkością i jakością klasyfikacji

1. Wprowadzenie

Algorytmy rozpoznania obrazów są wciąż fascynujące tego powodu że maszyna (komputer, układ obliczeniowy) może uczyć się rozpoznawania obiektów – rozumianych bardzo ogólnie. Ważne, by były opisane zastawem cech związanych nieznaną relacją z rozważanymi klasami. Zestaw cech opisujących obiekt jest jego „obrazem" w przestrzeni cech, co usprawiedliwia nazwę dziedziny. Zasadniczym elementem tego uczenia jest uogólnienie: komputer otrzymuje tylko przykłady niektórych obiektów, natomiast orzekać musi o wszystkich nowych obiektach spoza otrzymanego zbioru przykładów [7].

W obecnych czasach przed dziedziną rozpoznawania obrazów stawia się wiele trudnych i odpowiedzialnych zadań. Wzrost przestępczości czy też eskalacja terroryzmu zwróciły baczną opinie społeczeństwa i rządów wielu krajów świata na kwestie rozpoznawania i automatycznego identyfikowania twarzy, linii papilarnych bądź analizy danych biomedycznych [5].

Już ponad 10 lat temu stosując metody bazujące na rozpoznaniu obrazów, udało się naukowcom z MIT (*Massachusetts Institute of Technologies*) pod kierunkiem prof. Pentlanda opracować bardzo dobry algorytm rozpoznania twarzy. Pod rozwagę warto wziąć fakt, iż system ten potrafi poprawnie sklasyfikować twarze lekko obrócone (kąt wychylenia w obie strony 15^{o}) oraz inaczej oświetlone [4].

Powyższy przykład ilustruje ważność automatycznego rozpoznawania, lecz jednocześnie formułuje problemy wdrożeniowe, np. podobne algorytmy użyte przez ochronę lotnisk w Stanach Zjednoczonych nie wykazały zbyt dużej przydatności. Powodem była zbyt duża liczba danych (liczba zdjęć twarzy). Algorytmy zaprojektowane do obsługi znacznie mniejszych ilościowo baz danych nie były w stanie poprawnie uczyć się nowych danych.

Wielu autorów przedstawiło metody kompromisu pomiędzy prawdopodobieństwem mylnej klasyfikacji a jej szybkością. Dotyczą one jednak tylko klasyfikatora typu najbliższy

* Katedra Informatyki Stosowanej, Politechnika Łódzka

sąsiad (1–NS). Przypisuje on obiektowi tę samą klasę, z której pochodzi najbardziej mu podobny obiekt w zbiorze odniesienia. Zbiorem odniesienia może być oryginalny zbiór uczący. Nie jest on najlepszym klasyfikatorem, ale można nim aproksymować bardziej złożony, ale również prosty klasyfikator oparty na regule *k* najbliższych sąsiadów (*k*–NS). Wystarczy w tym celu dokonać reklasyfikacji zbioru uczącego z użyciem klasyfikatora aproksymowanego, a następnie stosować regułę 1–NN ze zreklasyfikowanym zbiorem uczącym jako zbiorem odniesienia. Zbiór uczący jest zbiorem użytym do konstrukcji klasyfikatora, a zbiór odniesienia, to zbiór, który musi być pamiętany podczas klasyfikacji. Kondensacja zbioru uczącego, czyli oryginalnego zbioru odniesienia, wchodzi więc w zakres procesu uczenia, czyli operacji obliczeniowych prowadzących do konstrukcji klasyfikatora, tzn. wyznaczenia jego parametrów. Reguła 1–NS jest nieco szybsza niż reguła *k*–NS, ale stosowanie całego zbioru uczącego w charakterze zbioru odniesienia jest zbyt kosztowne obliczeniowo.

W przeciągu wielu lat od czasu wprowadzenia reguły *k* – najbliższych sąsiadów zaproponowano wiele jej modyfikacji (np. sieci klasyfikatorów, klasyfikatory podejmujące decyzje rozmyte) i algorytmów mających na celu podniesienie jakości i szybkości klasyfikacji.

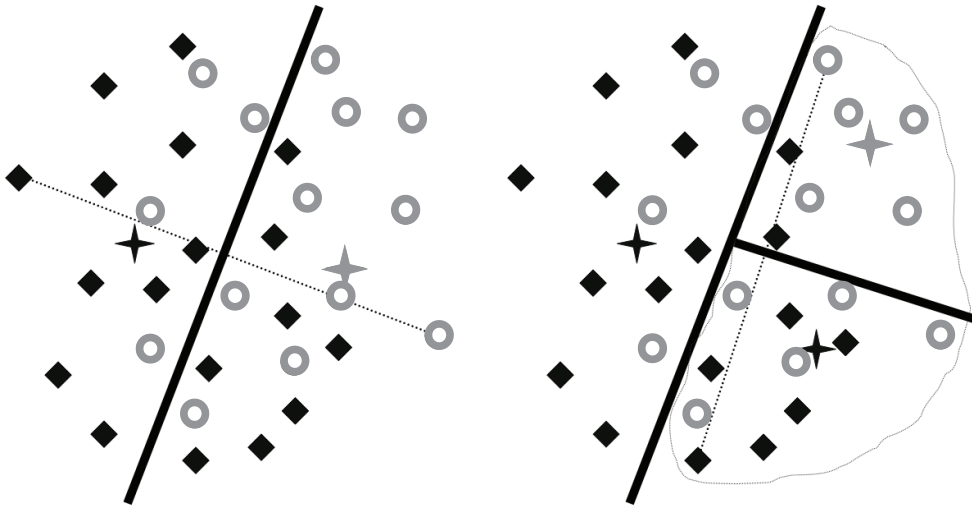
Trzeba zauważyć, iż bardzo mało jest rozwiązań przedstawiających przypadek, gdzie na klasyfikacje mamy określony limit czasu, co było powodem podjęcia tego zagadnienia przez autorów niniejszego opracowania.

2. Opis klasyfikatora wykorzystującego kondensację zbioru odniesienia metodą punktów wzajemnie najdalszych

Jednym z prostszych podejść jest podział zbioru odniesienia (uczącego) na podzbiory i zastąpienie każdego z tych podzbiorów jednym punktem w przestrzeni cech. Autorzy proponują dekompozycję uczenia na kilka cyklicznie powtarzających się podzadań, które można przerwać w dowolnym momencie, uzyskując najlepszą w tej chwili klasyfikacje.

W prezentowanej metodzie kondensacja jest rozwiązana poprzez przypisanie każdemu punktowi ze zbioru uczącego dokładnie jednej pary punktów wzajemnie najdalszych (z różnych klas) przy założeniu, że zawsze w przypadku najdalszych sąsiadów znajdujących się w równych odległościach wybieramy tego z mniejszym numerem. Ponieważ wielu obiektom może odpowiadać ta sama para punktów wzajemnie najdalszych lub para, w której dany punkt pokrywa się z innym, wybiera się tę parę z mniejszym numerem (tzn. szybciej znaleziona para). Problem pokrywania się punktów należących do tych samych klas i posiadających dokładnie te same cechy został rozwiązany poprzez ich pominięcie – nie wpłynęło to negatywnie na poziom błędu konkretnej iteracji. Para punktów służy do kolejnego podziału na nowy podzbiór poprzez wyznaczanie hiperpłaszczyzny rozdzielającej dokładnie te dwa punkty, przechodzi ona przez środek odcinka łączącego te punkty i jest do niego ortogonalna. Otrzymane w wyniku podziału nowe podzbiory są zastępowane środkami ciężkości, a przypisanie ich do konkretnej klasy następuje wg kryterium liczebności – do najliczniejszej. W każdej następnej iteracji, czyli kolejnym podziale jednego z podzbiorów,

brany jest pod uwagę zbiór o największej liczebności. Po każdym ustaleniu pary punktów wzajemnie najdalszych, wyliczano wielkość błędu klasyfikacji (metodą 1–NN) dla wielkości zbioru skondensowanego uzyskanego do danego etapu. Rysunek 1 przedstawia na przykładzie klasy dwuwymiarowej działanie (pierwsze dwie iteracje) algorytmu.



Rys. 1. Pierwsze dwie iteracje przykładowego działania przedstawianego algorytmu (z lewej strony – pierwsza iteracja, z prawej – druga)

Na rysunku przedstawiono dwie klasy (szarych kółek i czarnych rombów). Przerywaną linią zaznaczono wzajemnie najdalszą odległość pomiędzy punktami należącymi do przeciwnych klas. Czarna pogrubiona linia wyznacza podział na kolejne podzbiory, zaś czteroramienne gwiazdki symbolizują środki ciężkości zbiorów (ich kolor określa przynależność do klasy). W drugiej iteracji wybrano liczebniejszy zbiór zaznaczony także cienką obwódką.

3. Wyniki testów

Podczas przeprowadzania testów algorytmów wykorzystano zbiory (tab. 1) należące do repozytorium Uniwersytetu Kalifornijskiego w Irvine (Machine Learning Repository, University of California, Irvine) [3].

Są to następujące zbory:

– Glass. Jest to zestaw próbek różnego rodzajów szkła, różniących się na podstawie stwierdzenia występowania szczególnych pierwiastków chemicznych. Zbiór został zgromadzony przez kryminologów z Home Office Forensic Science Service w Reading w Wielkiej Brytanii.

– Iris. Jest to zbiór próbek trzech podgatunków kosaćca, klasyfikowanych na podstawie czterech geometrycznych cech [2, 8].
– Pima. Jest to zbiór odnoszący się do zadania rozpoznania symptomów cukrzycy w oparciu o kryteria przyjęte przez Światową Organizację Zdrowia (WHO). Dane zostały zgromadzone na podstawie badań populacji Indianek w wieku powyżej 21 lat z plemienia Pima (okolice Phoenix w Arizonie, USA) [6].
– Wine. Jest to zbiór dotyczący rozpoznania na podstawie cech wyekstrahowanych w wyniku analizy chemicznej jednego z trzech gatunków win włoskich [1].

Tabela 1
Charakterystyki zbiorów należących do repozytorium Uniwersytetu Kalifornijskiego w Irvine

Nazwa zbioru	liczba klas	liczba cech	liczba próbek
Glass	6	9	214
Iris	3	4	150
Pima	2	8	768
Wine	3	13	178

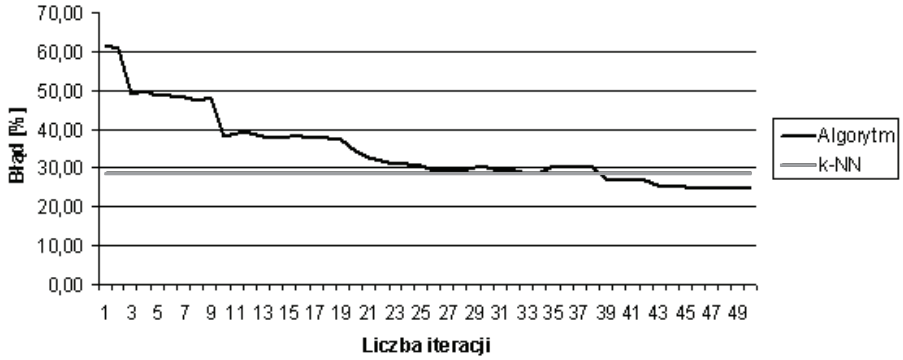
Wybranie tych konkretnych zbiorów testowych było podyktowane ich powszechnym stosowaniem i wykorzystywaniem w literaturze przedmiotu. Umożliwia to proste porównanie z innymi metodami rozpoznawania obrazów.

Wynik testów przedstawiono następująco:

1. przedstawiono nazwę testowanego zbioru (nazwa podrozdziału);
2. przedstawiono wykres ilustrujący wyniki uzyskane na konkretnym zbiorze testowym z zaznaczonym, dla łatwości porównania, błędem uzyskanym dla metody k–NN (dla $k = 1$) na całym zbiorze testowym;
3. przedstawiono tabelę, w której zawarto łączny czas obliczeń algorytmu klasyfikacji (uwzględniono czas klasyfikacji, czas potrzebny do redukcji zbioru odniesienia oraz czas potrzebny na ustalenie błędów metodą minus jednego elementu) w funkcji liczby iteracji potrzebnych; przedstawiono także czas obliczeń uzyskanym dla metody k–NN (dla $k = 1$).

Metoda minus jednego elementu polega na klasyfikacji każdego obiektu ze zbioru uczącego na podstawie reguły wyprowadzonej z pozostałych obiektów zbioru uczącego. Stosunek liczby obiektów mylnie w ten sposób zaklasyfikowanych do liczby wszystkich obiektów zbioru uczącego aproksymuje prawdopodobieństwo mylnych decyzji.

3.1. Glass (rys. 2, tab. 2)

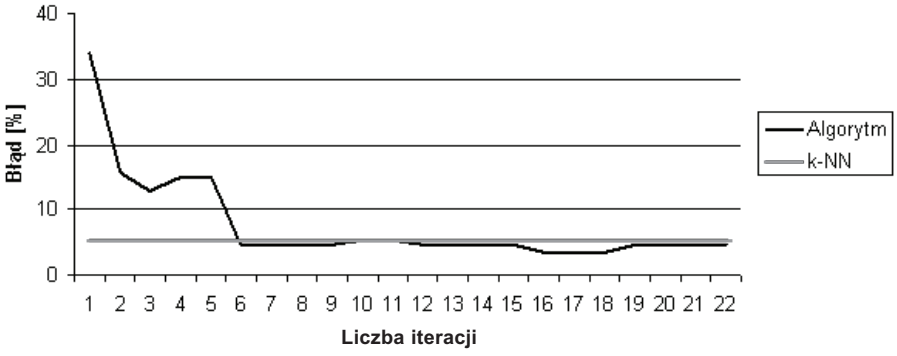


Rys. 2. Wyniki działania algorytmu uzyskane na zbiorze testowym Glass (dla porównania z błędem uzyskanym dla metody 1–NN na całym zbiorze testowym – oznaczona jako *k*–NN)

Tabela 2
Czas uzyskania wybranych wyników działania algorytmu uzyskane na zbiorze testowym Glass

Czas obliczeń błędu klasyfikacji [ms]		Błąd [%]
1–NN (cały zbiór)	6420	28,8
1–NN (zredukowany zbiór po 10 iteracjach)	360	38.51
1–NN (zredukowany zbiór po 28 iteracjach)	801	29.5
1–NN (zredukowany zbiór po 46 iteracjach)	1451	24,88

3.2. Iris (rys. 3, tab. 3)

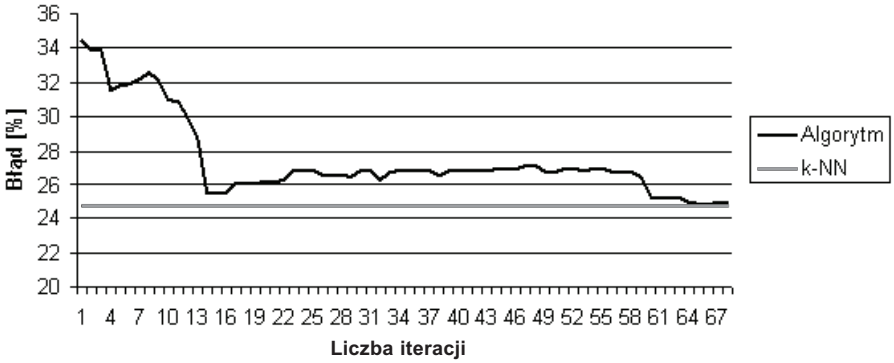


Rys. 3. Wyniki działania algorytmu uzyskane na zbiorze testowym Iris (dla porównania z błędem uzyskanym dla metody 1–NN na całym zbiorze testowym – oznaczona jako *k*–NN)

Tabela 3
Czas uzyskania wybranych wyników działania algorytmu uzyskane na zbiorze testowym Iris

Czas obliczeń błędu klasyfikacji [ms]		Błąd [%]
1–NN (cały zbiór)	300	5,47
1–NN (zredukowany zbiór po 3 iteracjach)	70	12,93
1–NN (zredukowany zbiór po 6 iteracjach)	120	4,76
1–NN (zredukowany zbiór po 16 iteracjach)	240	3,41

3.3. Pima (rys. 4, tab. 4)



Rys. 4. Wyniki działania algorytmu uzyskane na zbiorze testowym Pima (dla porównania z błędem uzyskanym dla metody 1–NN na całym zbiorze testowym – oznaczona jako *k*–NN)

Tabela 4
Czas uzyskania wybranych wyników działania algorytmu uzyskane na zbiorze testowym Pima

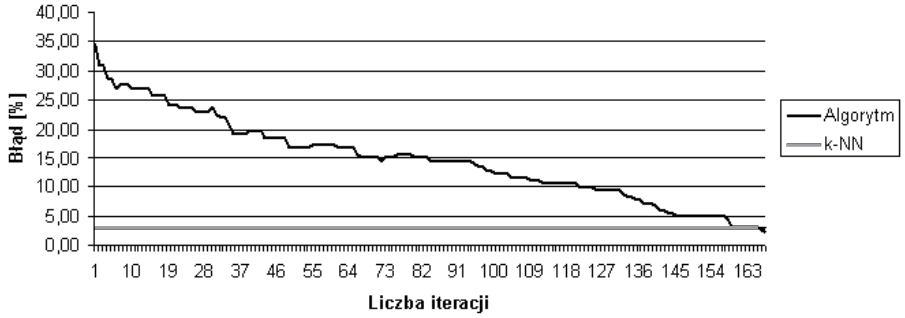
Czas obliczeń błędu klasyfikacji [ms]		Błąd [%]
1–NN (cały zbiór)	5276	24,7
1–NN (zredukowany zbiór po 4 iteracjach)	681	31,51
1–NN (zredukowany zbiór po 14 iteracjach)	1182	25,5
1–NN (zredukowany zbiór po 65 iteracjach)	4426	24,86

3.4. Wine (rys. 5, tab. 5)

Czas obliczeń został mierzony na komputerze klasy PC wyposażony w procesor Intela Pentium 4 HT 3 GHz i wyposażony w 512 MB pamięci operacyjnej.

Wyniki zawarte w tabelach pokazują, że przedstawiona konstrukcja klasyfikatora wykorzystująca opracowywaną metodę redukcji zbioru odniesienia jest skutecznym narzędziem sterowania kompromisem pomiędzy szybkością klasyfikacji i prawdopodobień-

stwem mylnej klasyfikacji z użyciem metody 1–NS. Jedynym wyjątkiem jest tu zbiór Wine – lecz uzyskany błąd 3.03 uzyskano w czasie przeszło cztery razy gorszym. Można zaradzić temu problemowi poprzez rezygnację z obliczeń błędu w każdej iteracji, a np. w co drugiej. Powoduje to znacznie skrócenie całkowitego czas obliczeń (jest on porównywany z czasem obliczeń pełnej klasyfikacji).



Rys. 5. Wyniki działania algorytmu uzyskane na zbiorze testowym Wine (dla porównania z błędem uzyskanym dla metody 1–NN na całym zbiorze testowym – oznaczona jako *k*–NN)

Tabela 5
Czas uzyskania wybranych wyników działania algorytmu uzyskane na zbiorze testowym Wine

Czas obliczeń błędu klasyfikacji [ms]		Błąd [%]
1–NN (cały zbiór)	3560	3,04
1–NN (zredukowany zbiór po 7 iteracjach)	150	27,53
1–NN (zredukowany zbiór po 67 iteracjach)	3184	15,16
1–NN (zredukowany zbiór po 161 iteracjach)	14481	3,37

Poprzez odpowiedni dobór liczby iteracji i oszacowanie dopuszczalnego błędu można otrzymać klasyfikator spełniający zadane wymagania. Uzyskany w ten sposób klasyfikator znakomicie można dopasować do konkretnego systemu przetwarzania danych.

4. Przyszłe kierunki badań

W przedstawionym artykule zaprezentowano obiecujące podejścia do redukcji kompletu odniesienia, oferujące jednocześnie niski odsetek błędów i równocześnie umożliwiające kontrolowanie kompromisem między jakością i szybkością klasyfikacji. Jednak relatywnie małe zbiory danych (poniżej 800 próbek) nie uwypuklają znacząco korzyści stosowania opracowanego algorytmu. Autorzy rozpoczęli badania nad wykorzystaniem metody do badań dużych zbiorów, mających nawet 80 000 próbek. Natrafiono tutaj na problem olbrzymiego zapotrzebowania na moc obliczeniową podczas ustalenia błędu metodą minus jednego elementu, co bardzo niekorzystnie wpływało na efektywność końcową.

Literatura

[1] Aeberhard S., Coomans D., deVel O.: *Comparison of Classifiers in High Dimensional Settings*. Tech. Rep. nr. 92–02, Department of Computer Science and Department of Mathematics and Statistics, James Cook University of North Queensland, 1992
[2] Fisher R.A.: *The use of multiple measurements in taxonomic problems*. Annual Eugenics, USA, nr 7, t. II, 1936, 179–188
[3] Merz Ch., Murphy P.M.: *UCI repository of machine learning databases*. [http://www.ics.uci.edu/~mlearn/MLRepository.html], 1996
[4] Moghaddam B. Pentland A.: *An automatic system for model based coding of faces*. IEEE International Conference on Image Processing, Washington DC, USA, 1995
[5] Skrabek W.: *Multimedia – Algorytmy i Standardy Kompresji*. AOWPLJ, Warszawa, 1998
[6] Smith J.W., Everhart J.E., Dickson W.C., Knowler W.C., Jonannes R.S.: *Using the ADAP learning algorithm to forecast the onset of diabetes mellitus*. Proceedings of the Symposium on Computer Applications and Medical Care, IEEE Computer Society Press, USA, 1988, 261–265
[7] Tadeusiewicz R., Flasiński M.: *Rozpoznawanie obrazów*. Warszawa, PWN 1991
[8] Wiley J.: *Contributions to Mathematical Statistics*. NY, USA, 1950