AUTOMATYKA • 2007 • Tom 11 • Zeszyt 3

Artur Sierszeń*

Modyfikacja algorytmu Changa z wykorzystaniem metody znajdowania punktów najbliższych

1. Wprowadzenie

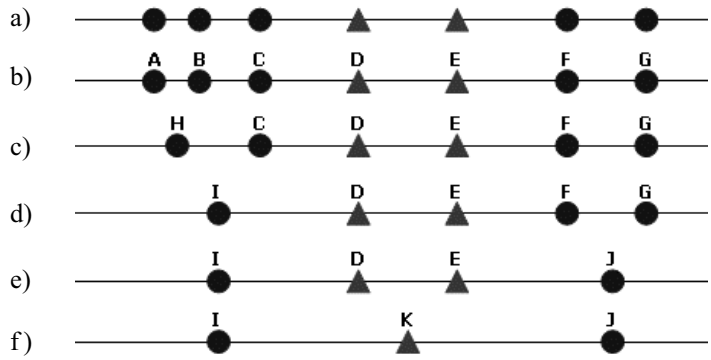
Algorytm Changa znany jest od ponad 30 lat [2]. Przez lata uważany był za klasyfikator dobry pod względem redukcji zbioru uczącego i jednocześnie mało użyteczny ze względu na prędkość działania. Klasyfikator ten jest klasyfikatorem najbliższego sąsiedztwa tzn. takim, który przypisuje obraz do klasy najbliższego prototypu. Idea polega na rozpoczęciu klasyfikacji z każdą próbką w zbiorze uczącym jako prototypem, a potem łączeniu ze sobą dwóch najbliższych prototypów tej samej klasy tak długo, aż poziom rozpoznania nie ulegnie pogorszeniu. Algorytm ten jest znakomity pod względem możliwości redukcji i stąd zainteresowanie autora dostosowaniem mechanizmu redukcji do budowy nowego klasyfikatora. Niniejsza publikacja stanowi wstępne studium poprawy szybkości działania metody.

2. Algorytm Changa i jego modyfikacja

Załóżmy, że zbiór uczący T jest dany jako $T = \{t1, \ldots, tm\}$. Zasada naszego algorytmu jest następująca: zaczynamy od każdego punktu w T jako prototypu. Następnie sukcesywnie łączymy ze sobą dwa najbliższe sobie prototypy $p1$ i $p2$ tej samej klasy (to znaczy zastępujemy $p1$ i $p2$ nowym prototypem p), jeżeli połączenie nie pogorszy klasyfikacji obrazów w T. Nowy prototyp p może być po prostu średnim wektorem z $p1$ i $p2$ lub średnim wektorem z ważonego $p1$ i $p2$. Klasa nowego prototypu jest taka sama jak klasa $p1$ i $p2$. Kontynuujemy proces łączenia, aż do momentu, kiedy zacznie wzrastać liczba nieprawidłowych klasyfikacji wzorów w zbiorze T. Podajemy prosty przykład ilustrujący tę zasadę. Załóżmy, że mamy dany zbiór uczący próbek pokazany na rysunku 1a. Zaczynamy od prototypów pokazanych na rysunku 1b, który jest taki sam, jak rysunek 1a. Zauważmy, że klasyfikator najbliższego sąsiedztwa, używając prototypów z rysunku 1b, może poprawnie sklasyfikować wszystkie obrazy z rysunku 1a. Teraz, ponieważ prototypy A i B są sobie najbliższe i są tej samej klasy, próbujemy je połączyć. Jeżeli A i B są zastąpione nowym prototypem H,

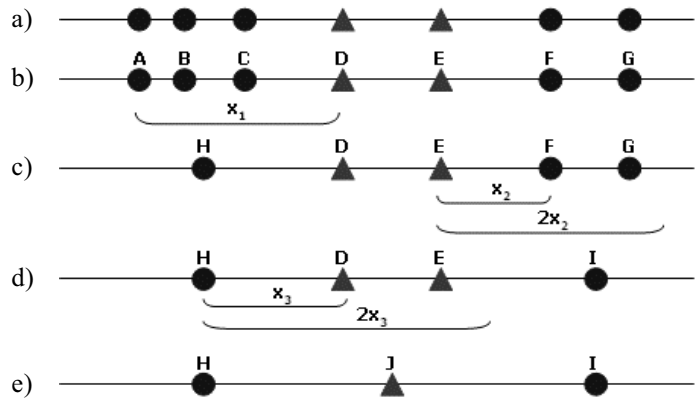
* Katedra Informatyki Stosowanej, Politechnika Łódzka w Łodzi

wszystkie obrazy z rysunku 1a są nadal poprawnie sklasyfikowane. Dlatego, zastępując A i B prototypem H, uzyskujemy nowy zbiór prototypów pokazany na rysunku 1c. Podobnie, zastępując H i C prototypem I, uzyskujemy stan pokazany na rysunku 1d. Łącząc F i G w prototyp J, otrzymujemy stan pokazany na rysunku 1e. Ostatecznie, zastępując D i E prototypem K, otrzymujemy stan pokazany na rysunku 1f. Używając prototypów pokazanych na rysunku 1f, każdy obraz z rysunku 1a nadal zostanie poprawnie sklasyfikowany. Jednak jeśli spróbujemy połączyć I i J, niektóre obrazy z rysunku 1a zostaną sklasyfikowane błędnie. Dlatego kończymy proces łączenia, a punkty pokazane na rysunku 1f zostaną użyte jako prototypy w klasyfikatorze najbliższego sąsiedztwa.



Rys. 1. Uproszczona graficzna interpretacja algorytmy Changa
Objaśnienia w tekście

Modyfikacja zaproponowana przez autora ma na celu przyspieszenie obliczeń poprzez zastąpienie ograniczenia łączenia ze sobą jedynie dwóch najbliższych sobie prototypów $p1$ i $p2$ tej samej klasy regułą określającą, iż łączone są ze sobą wszystkie prototypy tej samej klasy $p1...pN$ znajdujące się bliżej od jakiegokolwiek prototypu należącego do innej klasy. Uwzględniając poprzedni przykład opisujący algorytm Changa, możemy założyć, że dysponujemy dokładnie takim samym zbiorem uczącym próbek, co pokazany na rysunku 2a. Zaczynając od losowo wybranego prototypu (w tym przypadku od A), wyszukujemy odległość do najbliższego prototypu reprezentującego inną klasę (D). Odległość ta, oznaczona jako $x1$, umożliwia nam znalezienie wszystkich prototypów znajdujących się bliżej niż prototyp D, a więc zarówno B, jak i C – jak ilustruje to rysunek 2b. Łączymy znalezione w ten sposób prototypy w jeden oznaczony na rysunku 2c jako H. Nowy prototyp jest średnim wektorem z prototypów, które zastępuje. Następnym losowo wybranym punktem może być prototyp F. Wyznaczamy ponownie odległość do najbliższego prototypu reprezentującego inną klasę (E) – oznaczoną jako $x2$ i poszukujemy prototypów należących do tej samej klasy w promieniu $x2$, rysunek 2c. W tym kroku łączymy jedynie prototyp F i G, w wyniku czego powstaje prototyp J. Analogicznie postępujemy dalej, ponownie losowo wybierając punkt (tym razem prototyp D), rysunek 2d. Wyznaczamy odległość do najbliższego prototypu z innej klasy (prototyp H) – $x3$ i szukamy wszystkich prototypów leżących bliżej i należących do tej samej klasy (tylko prototyp E). Znaleziony prototyp E łączymy z D, w efekcie czego zostają one zastąpione prototypem J (rys. 2e).



Rys. 2. Uproszczona graficzna interpretacja modyfikacji algorytmu Changa
Objaśnienia w tekście

Po każdym ustaleniu nowego zbioru, wyliczano wielkość błędu klasyfikacji (metodą 1–*NN*) dla wielkości zbioru skondensowanego uzyskanego do danego etapu (iteracji) – czynność tę wykonywano przy analizie oryginalnego oraz zmodyfikowanego algorytmu Changa. Weryfikacja czasochłonności obliczeń prowadzona była poprzez porównanie czasu (mierzonego w cyklach pracy procesora) otrzymania kolejnych wyników błędu klasyfikacji.

3. Weryfikacja modyfikacji algorytmu Changa

Podczas przeprowadzania eksperymentów i testów algorytmy oraz podczas weryfikacji istniejących algorytmów rozpoznawania obrazów użyto zbiorów należących do repozytorium Uniwersytetu Kalifornijskiego w Irvine (Machine Learning Repository, University of California, Irvine). Uczyniono to przede wszystkim ze względu na ich powszechne stosowanie i wykorzystywanie w literaturze przedmiotu [3]. Są to następujące zbiory (tab. 1):

- 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 (odmian) kwiatu irysa, *Iris setosa*, *Iris versicolor* i *Iris virginica*, klasyfikowanych na podstawie czterech geometrycznych cech kwiatu – długości i szerokości płatków kielicha oraz długości i szerokości płatków okwiatu [5, 7].
- 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 plemienia Pima w wieku powyżej 21 lat z (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
Parametry małych liczebnie zbiorów użytych do weryfikacji modyfikacji algorytmu Changa

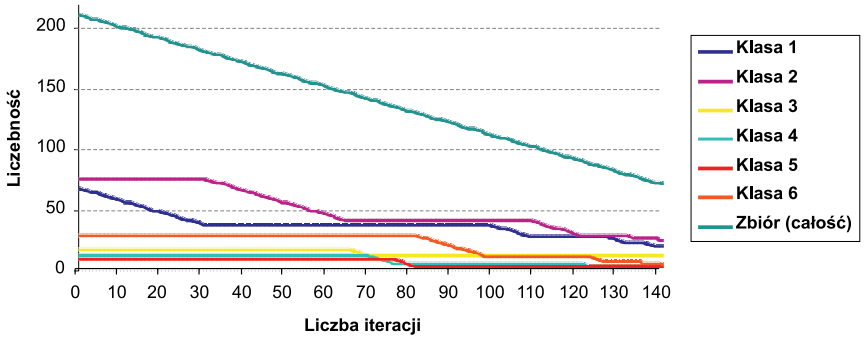
Nazwa zbioru	Liczba klas	Liczba cech	Liczba próbek	Liczebność poszczególnych klas w zbiorze					
				Klasa 1	Klasa 2	Klasa 3	Klasa 4	Klasa 5	Klasa 6
GLASS	6	9	214	70	76	17	13	9	29
IRIS	3	4	150	50	50	50	–	–	–
PIMA	2	8	768	500	268	–	–	–	–
WINE	3	13	178	59	71	48	–	–	–

Każdy test został poprzedzony dokładną analizą i testami konkretnego zbioru przy użyciu oryginalnej metody Changa. Następnie przeprowadzono serie testów (25) przy użyciu autorskiej modyfikacji algorytmu Changa (w niniejszym artykule przedstawione są wyniki serii będącej najbliższej średniej ze wszystkich serii pomiarowych). Wszystkie podane wyniki są wynikami uśrednionymi wszystkich serii pomiarowych.

4. Wyniki eksperymentów

4.1. Zbiór testowy GLASS

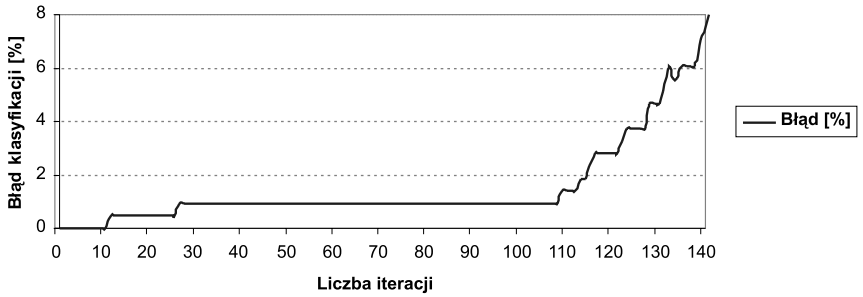
Na wykresach zaprezentowano wyniki działania oryginalnego algorytmu Changa, zarówno w aspekcie kondesacji zbioru GLASS (rys. 3), jak i jej wpływu na błąd klasyfikacji (rys. 4) oraz analogiczne wyniki działania zmodyfikowanego algorytmu Changa, zarówno w aspekcie kondesacji zbioru (rys. 5), jak i jej wpływu na błąd klasyfikacji (rys. 6).



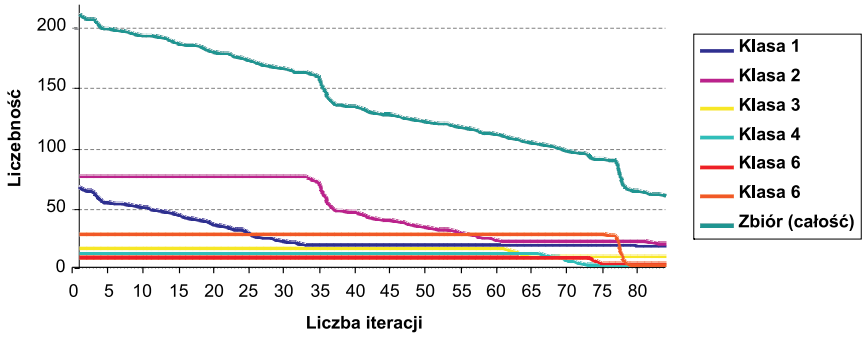
Rys. 3. Liczebność poszczególnych klas zbioru GLASS poddawanych kondensacji oryginalnym algorytmem Changa

Można od razu zauważyć znaczny wzrost błędu klasyfikacji metodą 1–*NN* dla zbioru kondensowanego zmodyfikowanym algorytmem. Jednak znaczna akceleracja obliczeń (czas trwania badania całego procesu analizy kondensacji zbioru GLASS był o 57,68%

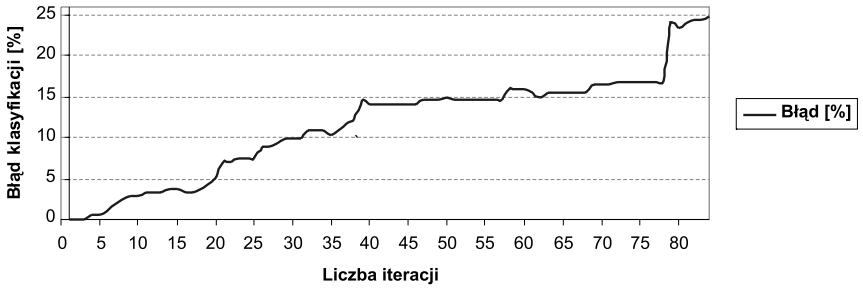
krótszy) skłania do szukania zastosowań, gdzie czas obliczeń jest pierwszorzędną cechą pracy systemu, a dokładność mniej istotną lub takich systemów, gdzie zakładany poziom błędu ma być osiągnięty w najlepszym, tzn. najkrótszym, czasie.



Rys. 4. Wpływ kondensacji zbioru GLASS oryginalnym algorytmem Changa na końcowy wynik klasyfikacji metodą 1–*NN*



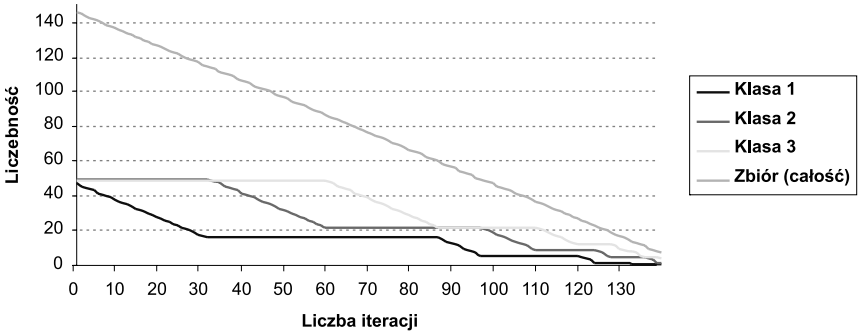
Rys. 5. Liczebność poszczególnych klas zbioru GLASS poddawanych kondensacji zmodyfikowanym algorytmem Changa



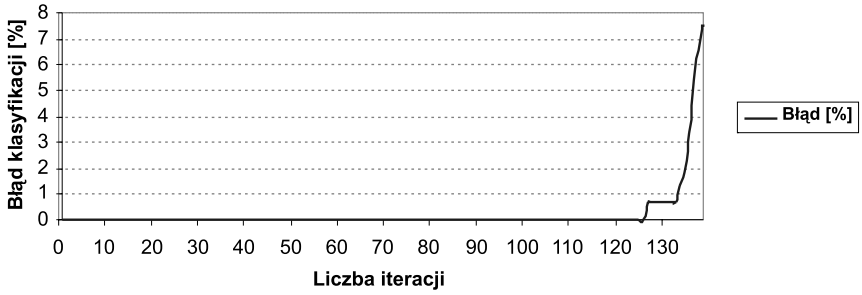
Rys. 6. Wpływ kondensacji zbioru GLASS zmodyfikowanym algorytmem Changa na końcowy wynik klasyfikacji metodą 1–*NN*

4.2. Zbiór testowy IRIS

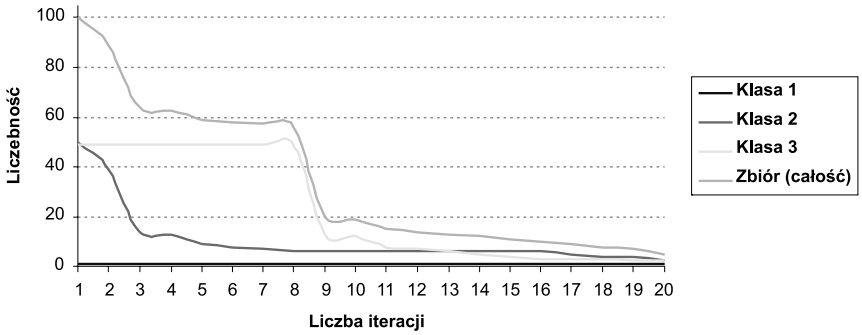
Na wykresach zaprezentowano wyniki działania oryginalnego algorytmu Changa, zarówno w aspekcie kondesacji zbioru IRIS (rys. 7), jak i jej wpływu na błąd klasyfikacji (rys. 8) oraz analogiczne wyniki działania zmodyfikowanego algorytmu Changa, zarówno w aspekcie kondesacji zbioru (rys. 9), jak i jej wpływu na błąd klasyfikacji (rys. 10).



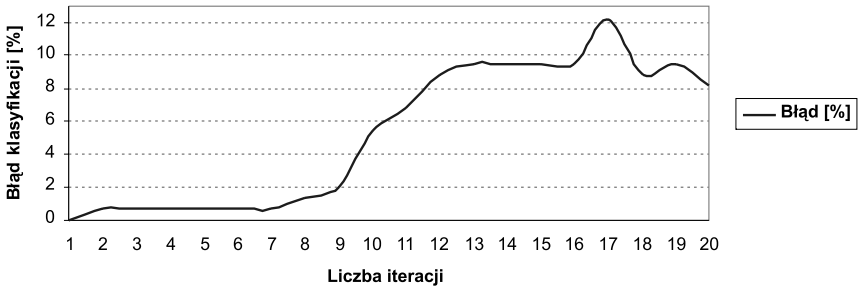
Rys. 7. Liczebność poszczególnych klas zbioru IRIS poddawanych kondensacji oryginalnym algorytmem Changa



Rys. 8. Wpływ kondensacji zbioru IRIS oryginalnym algorytmem Changa na końcowy wynik klasyfikacji metodą 1–*NN*



Rys. 9. Liczebność poszczególnych klas zbioru IRIS poddawanych kondensacji zmodyfikowanym algorytmem Changa



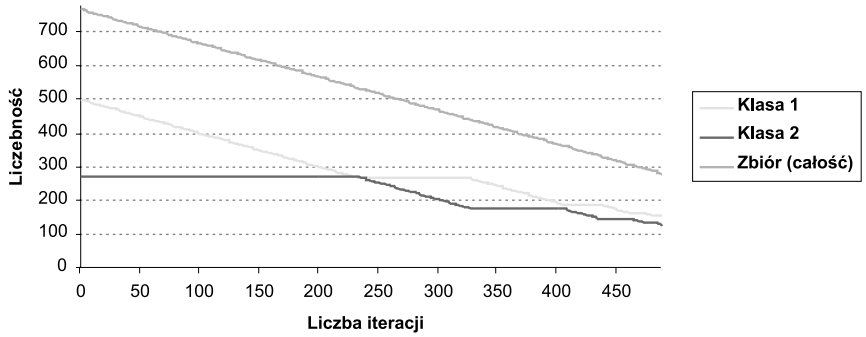
Rys. 10. Wpływ kondensacji zbioru IRIS zmodyfikowanym algorytmem Changa na końcowy wynik klasyfikacji metodą 1–*NN*

W porównaniu z wynikami z poprzedniego zbioru testowego widać wyraźnie niewielki stosunkowo wzrost błędu klasyfikacji metodą 1–*NN* dla zbioru kondensowanego zmodyfikowanym algorytmem. Ponadto, analogicznie jak w przypadku analizy zbioru GLASS, zaobserwowano przyspieszenie obliczeń (choć nie tak spektakularne jak we wcześniejszym przykładzie), gdyż czas trwania badania całego procesy analizy kondensacji zbioru IRIS był o 8,75% krótszy.

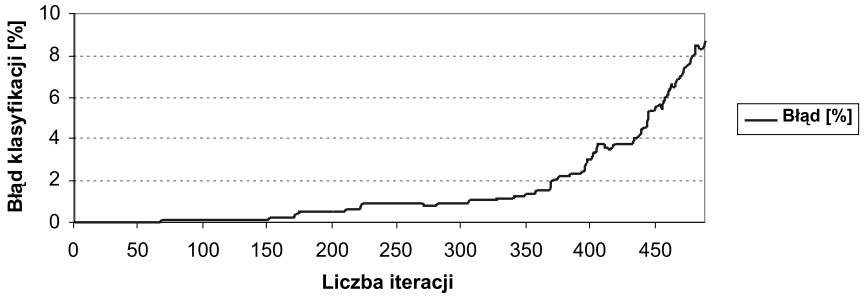
4.3. Zbiór testowy PIMA

Na wykresach zaprezentowano wyniki działania oryginalnego algorytmu Changa, zarówno w aspekcie kondesacji zbioru PIMA (rys. 11), jak i jej wpływu na błąd klasyfikacji (rys. 12) oraz analogiczne wyniki działania zmodyfikowanego algorytmu Changa, zarówno w aspekcie kondesacji zbioru (rys. 13), jak i jej wpływu na błąd klasyfikacji (rys. 14).

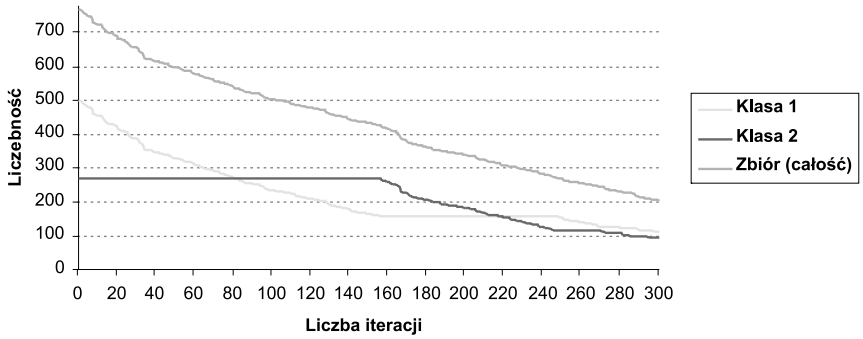
Analiza wykresów pozwala łatwo stwierdzić duże podobieństwo z testami przeprowadzonymi na zbiorze GLASS – można zaobserwować znaczny wzrost błędu klasyfikacji metodą 1–*NN* dla zbioru kondensowanego zmodyfikowanym algorytmem. Analogicznie zaobserwowano dużą akceleracje obliczeń – czas trwania badania całego procesu analizy kondensacji zbioru GLASS był o 53,74% krótszy.



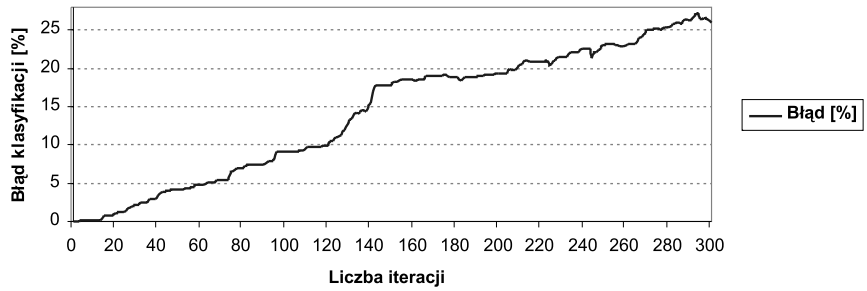
Rys. 11. Liczebność poszczególnych klas zbioru PIMA poddawanych kondensacji oryginalnym algorytmem Changa



Rys. 12. Wpływ kondensacji zbioru PIMA oryginalnym algorytmem Changa na końcowy wynik klasyfikacji metodą 1–*NN*



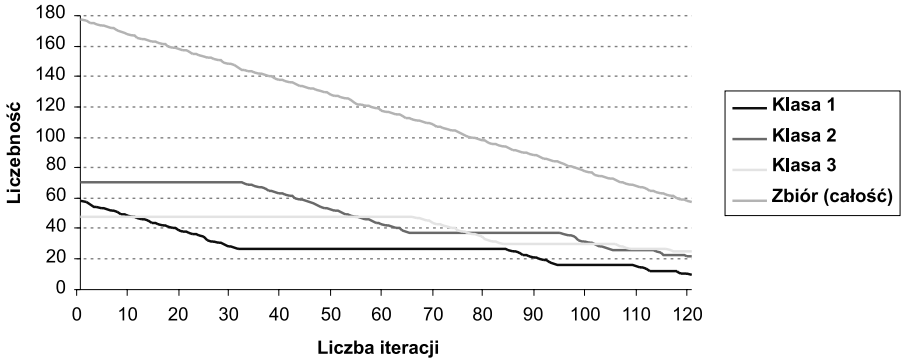
Rys. 13. Liczebność poszczególnych zbioru PIMA poddawanych kondensacji zmodyfikowanym algorytmem Changa



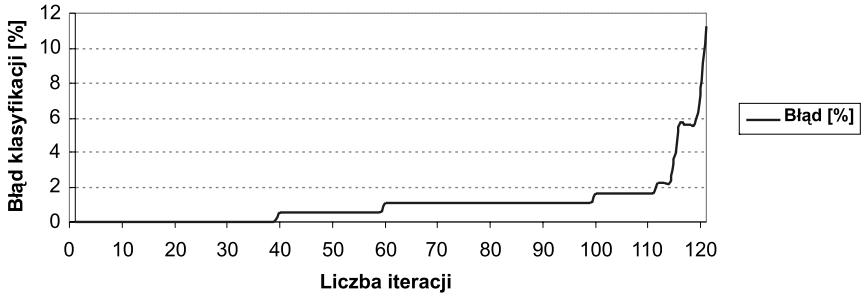
Rys. 14. Wpływ kondensacji zbioru PIMA zmodyfikowanym algorytmem Changa na końcowy wynik klasyfikacji metodą 1–*NN*

4.4. Zbiór testowy WINE

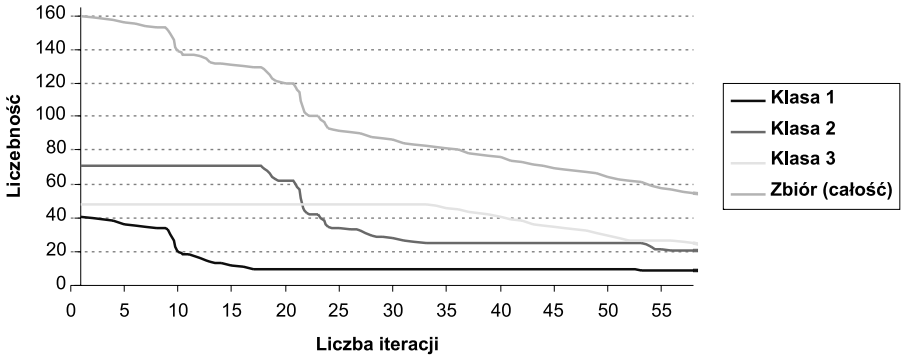
Na wykresach zaprezentowano wyniki działania oryginalnego algorytmu Changa, zarówno w aspekcie kondesacji zbioru WINE (rys. 15), jak i jej wpływu na błąd klasyfikacji (rys. 16) oraz analogiczne wyniki działania zmodyfikowanego algorytmu Changa, zarówno w aspekcie kondesacji zbioru (rys. 17), jak i jej wpływu na błąd klasyfikacji (rys. 18).



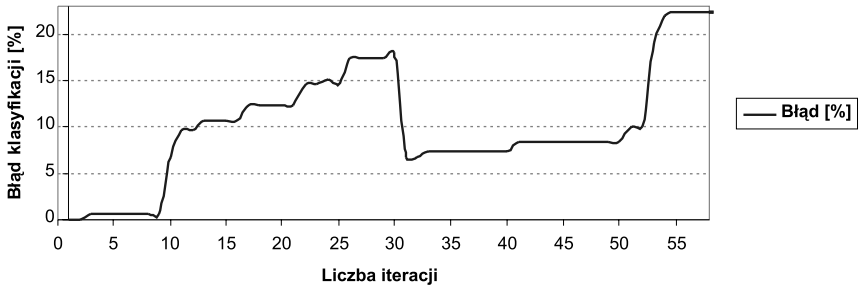
Rys. 15. Liczebność poszczególnych klas zbioru WINE poddawanych kondensacji oryginalnym algorytmem Changa



Rys. 16. Wpływ kondensacji zbioru WINE oryginalnym algorytmem Changa na końcowy wynik klasyfikacji metodą 1–*NN*



Rys. 17. Liczebność poszczególnych klas zbioru WINE poddawanych kondensacji zmodyfikowanym algorytmem Changa



Rys. 18. Wpływ kondensacji zbioru WINE zmodyfikowanym algorytmem Changa na końcowy wynik klasyfikacji metodą 1–*NN*

Analiza wyników testów przeprowadzonych na zbiorze WINE wykazała przede wszystkim bardzo nieznaczny wpływ jednej z klas kondensowanego zbioru na błąd klasyfikacji. Jednak globalny wzrost błędu klasyfikacji metodą 1–*NN* dla zbioru kondensowanego zmodyfikowanym algorytmem był duży. Zaobserwowano znaczną akceleracje obliczeń – czas trwania badania całego procesu analizy kondensacji zbioru WINE był o 43,08% krótszy.

5. Wnioski

W przedstawionym artykule zaprezentowano obiecujące podejścia do redukcji kompletu odniesienia. Znaczna akceleracja obliczeń jak i dokładne wyznaczenie błędu umożliwia kontrolowanie kompromisem między jakością i szybkością klasyfikacji, co było głównym kryterium rozpoczęcia badań w tym zakresie. Jednak relatywnie małe zbiory danych (poniżej 800 próbek) nie uwypuklają znacząco korzyści stosowania opracowanego algorytmu. Przedstawione wyniki badań wydają się być dobrym materiałem wyjściowym do opracowania klasyfikatora wykorzystującego przedstawiona modyfikację zbioru odniesienia. Wcześniej konieczne będzie zweryfikowanie i ewentualne dalsze modyfikacje algorytmu, tak aby był on wstanie obsługiwać duże zbiory danych.

Literatura

[1] Aeberhard S., coomans D., deVel O.: *Comparison of Classifiers in High Dimensional Settings*. Tech. Rep. no. 92–02, Department of Computer Science and Department of Mathematics and Statistics, James Cook University of North Queensland, 1992

[2] Chang C.L.: *Finding Prototypes for Nearest Neighbor Classifiers*. IEEE Transactions on Computers, t. C-23, nr 11, 1974, 1179–1184

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

[4] Prim R.C.: *Shortest connection networks and some generalizations*. Bell Syst. Tech, tom J, 1957, 1389–1401

[5] Fisher R.A.: *The use of multiple measurements in taxonomic problems*. Annual Eugenics, 7, Part II, USA, 1936, 179–188

[6] Smith J.W., Everhart J.E., Dickson W.C., Knowler W.C., Johannes 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, 1988, 261–265

[7] Wiley J.: *Contributions to Mathematical Statistics*. NY, USA, 1950