Ewelina PIOTROWSKA, Włodzimierz STANISŁAWSKI
POLITECHNIKA OPOLSKA, WYDZIAŁ ELEKTROTECHNIKI, AUTOMATYKI I INFORMATYKI, INSTYTUT AUTOMATYKI I INFORMATYKI
ul. Sosnkowskiego 31, 45-272 Opole

Analiza danych niezrównoważonych we wstępnej diagnostyce raka pęcherza moczowego

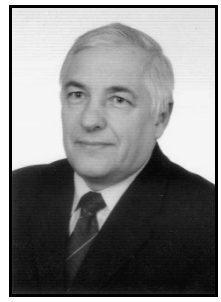
Dr inż. Ewelina PIOTROWSKA

Pracownik Wydziału Elektrotechniki, Automatyki i Informatyki Politechniki Opolskiej. Tytuł doktora uzyskała w 2012r. na tym samym wydziale. Prowadzi prace z zakresu diagnostyki procesów i systemów, systemów wspomagania decyzji, eksploracji danych, systemów baz danych, sztucznej inteligencji.

e-mail: e.piotrowska@po.opole.pl

Dr hab. inż. Włodzimierz STANISŁAWSKI

Studia i doktorat na Wydziale Automatyki, Elektroniki i Informatyki Politechniki Śląskiej. Habilitacja na Uniwersytecie Elektrotechnicznym w Sankt Petersburgu. Od 2003 stanowisko profesora Politechniki Opolskiej. Kieruje Katedrą Informatyki. Zainteresowania naukowe obejmują zagadnienia związane z modelowaniem i symulacją komputerową złożonych obiektów sterowania oraz sztuczną inteligencją. Od 1 września 2008 pełni funkcję prodziekana Wydziału ds. naukowych.



e-mail: w.stanislawski@po.opole.pl

Streszczenie

Artykuł przedstawia wyniki rozważań dotyczących klasyfikacji danych niezrównoważonych w obrazach mikroskopowych preparatów cytologicznych. Do klasyfikacji wykorzystano algorytmy uczenia nadzorowanego jak: naiwny klasyfikator Bayesa, analiza dyskryminacyjna, drzewa decyzyjne oraz zaproponowany przez autorów algorytm klasyfikacji będący połączeniem zbiorów przybliżonych i metody k-najbliższych sąsiadów. Do analizy wykorzystano opracowane przez autorów narzędzie Rough Sets Analysis Toolbox (RSA Toolbox) – przybornik dla środowiska MATLAB. Wykorzystane obrazy mikroskopowe uzyskano w procesie diagnostyki nowotworu pęcherza moczowego badając metodą FISH odpowiednio przygotowane preparaty moczu.

Słowa kluczowe: dane niezrównoważone, uczenie nadzorowane.

Analysis of imbalanced data using morphometric parameters in diagnosis of bladder cancer

Abstract

In the paper the results of imbalanced data classification based on microscope images are described. The images were acquired in the process of bladder cancer diagnosis using the FISH method. The conducted research were focused on the effectiveness of the initial cancer diagnosis using specimen radiation in a DAPI channel and supervised learning methods. The analyzed data set contains about 23,000 objects described by 212 morphometric features. Each object was classified to one of two classes: normal cells or cancers cells. Decisions about belonging objects to the corresponding classes were carried out by an expert. There were identified only 640 cancer cells in the analyzed data. Most of learning algorithms assume balance between classes. The class imbalance problem causes difficulties at a learning stage and reduces the predictive ability. Therefore, the classifier evaluation was performed using G-mean and F-value measures. The authors defined additional measure FMaxSen=sen2·spe which is the product of sensitivity and specificity coefficients. Use of the second power factor emphasizes the importance of sensitivity and allows searching the classifier with the maximum specificity at the maximum sensitivity. The analysis presented in the paper was performed with use of Rough Sets Analysis Toolbox (RSA Toolbox) for MATLAB implemented by the authors. The main part of the RSA Toolbox contains a module which supports the rough sets theory processing. Another part (RSAm module) is a wrapper for the proposed rough classification functions and others implemented in Matalab such as NaiveBayes, Discriminant Analysis, Decision Tree. The RSAm gives us possibility to use cross validation for measuring the classification accuracy. The RSAm also contains features reduction algorithms (correlation based feature selection, sequential feature selection, principal component analysis) as well as discretizations algorithms (EWD, CAIM, CACC). An important part of the RSAToolbox is implementation of distributed computations using Matlab Parallel Computing Toolbox and Distributed Computing Server.

Keywords: imbalanced data, supervised learning.

1. Wprowadzenie

Wiele chorób nowotworowych diagnozuje się dopiero, gdy zachodzące w organizmie zmiany widoczne są makroskopowo. Późne wykrycie choroby uniemożliwia przeprowadzenie leczenia, pozwalającego na skuteczną walkę z chorobą. Dlatego ważnym zagadnieniem jest poszukiwanie metod pozwalających na wczesną diagnozę. Prowadzone są w tym celu badania przesiewowe, którym poddawane są osoby będące w grupie zwiększonego ryzyka. W efekcie do badań otrzymuje się próbę o dużej liczebności przypadków, która charakteryzuje się znacznym niezrównoważeniem. Zdecydowaną większość próby stanowią przypadki osób zdrowych, natomiast znikomą przypadki osób chorych.

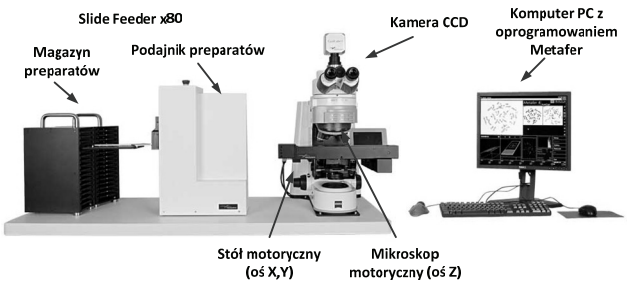
Zjawisko niezrównoważenia danych występuje także przy diagnostyce choroby na podstawie preparatów mikroskopowych badań cytologicznych. Problem zaobserwowano w diagnostyce nowotworu pęcherza moczowego przy użyciu metody FISH (ang. fluorescence in situ hybridization, fluorescencyjna hybrydyzacja in situ) w analizie cytologicznej moczu. Metoda FISH polega na zastosowaniu fluorescencyjnych markerów, które uwidaczniają się w komórkach zmienionych przez nowotwór. Do wykrywania markerów wykorzystywane są mikroskopy fluorescencyjne [1, 2, 3].

2. Diagnostyka mikroskopowa

W preparatach mikroskopowych nie są obserwowalne naturalne obrazy komórek. Pojedyncze komórki są prawie przezroczyste i muszą być odpowiednio uwidocznione, aby były dostrzegalne pod mikroskopem. Stosuje się w tym celu różne sposoby barwienia, obejmujące cały zakres światła widzialnego. Każdy z nich w odmienny sposób uwidacznia poszczególne składniki komórek, co wpływa na ich formę morfologiczną. W mikroskopii fluorescencyjnej powszechnie stosowanym barwnikiem jest DAPI. Jego niebieska emisja ułatwia analizę preparatów z wykorzystaniem różnych markerów fluoroscencyjnych. Metodyka wykonywania rozmazów cytologicznych powoduje, że komórki często tworzą zlepy, grupy, skupiska, wzajemnie nakładając się na siebie [4]. Największe z nich mają tendencję do rozmieszczania się na obwodzie preparatu. Dlatego przy analizie obrazów mikroskopowych powinno się brać pod uwagę wszystkie pola. Losowy wybór pól może prowadzić do zafałszowań statystycznych dotyczących wskaźników ilościowych opisujących badaną populację komórek.

W diagnostyce mikroskopowej wykorzystuje się tzw. systemy skaningowe (rys. 1). Analiza preparatu znajdującego się pod mikroskopem rozpoczyna się od jego podziału na pola skanowania. Obraz jest pobierany osobno dla każdego ze zdefiniowanych kanałów kolorów. Liczba komórek w analizowanych preparatach może być bardzo duża. Przy skanowaniu każdego pola preparatu różnymi kolorami światła mikroskopu, badanie trwałoby kilka, a nawet kilkanaście godzin. Dlatego ważnym etapem jest skanowanie wstępne w kanale DAPI, które pozwala na wybranie pól

preparatu w których mogą znajdować się komórki nowotworowe. Zasady rozpoznawania obrazów mikroskopowych opierają się o kryteria dotyczące struktury całego obrazu lub zespołów cech morfologicznych wyróżnionych obiektów. Nie można tutaj zastosować metody rozpoznawania polegającej na porównywaniu i subtrakcji obrazu wzorcowego i obrazu diagnozowanego [4, 5, 6, 7].



Rys. 1. Komponenty systemu skaningowego Metafer [8]
Fig. 1. Componens of Metafer scaning platform [8]

Do automatycznej oceny typu komórki wykorzystuje się parametry morfometryczne. Każdy parametr morfometryczny stanowi liczbowy opis obiektu morfologicznego. Może on dotyczyć takich właściwości jak liczebność obiektów, ich wielkość, kształt, właściwości optyczne, tekstura czy topologia. Wybór parametrów zależy od rodzaju badanego materiału genetycznego. Analiza komórek w polu preparatu pod kątem oceny wartości każdej z cech jest kosztowna czasowo. Im mniejsza liczba cech, będących kryterium oceny komórki, tym proces skanowania wstępnego jest szybszy. Dlatego poszukuje się metod pozwalających na wybór optymalnego zbioru cech [9].

3. Dane niezrównoważone na przykładzie analizy preparatów mikroskopowych

System skaningowy przedstawiony na rys. 1 zastosowano do analizy cytologicznej moczu z wykorzystaniem metody FISH (Fluorescence in situ hybridization). Prowadzone badania dotyczą skuteczności wstępnej diagnostyki nowotworu przy zastosowaniu metod uczenia nadzorowanego na poziomie analizy obrazów preparatów naświetlanych kanałem DAPI. W wyniku analizy obrazów pól preparatów uzyskano zbiór danych składający się z 22962 obiektów opisanych przez 212 parametrów morfometrycznych. Każdy z obiektów występujący w zbiorze przypisano do jednej z dwóch klas: komórki zdrowe lub komórki nowotworowe. Ocena przynależności obiektów do odpowiedniej klasy została przeprowadzona przez eksperta, który w badanym zbiorze wskazał tylko 640 komórek nowotworowych. Z tego powodu analizowany zbiór charakteryzuje się dużym niezrównoważeniem klas. Klasa poszukiwana, odpowiadająca komórkom nowotworowym, stanowi niecałe 3% wszystkich dostępnych w analizie przypadków.

Większość algorytmów uczących zakłada w przybliżeniu zrównoważenie klas. Dlatego opisany powyżej problem niezrównoważenia klas powoduje trudności w fazie uczenia i obniża zdolność predykcyjną. Niska jakość klasyfikacji może także wynikać ze złego uwarunkowania danych klasy mniejszościowej, jak: zbyt mała liczba przykładów, nakładanie się przypadków klasy większościowej na mniejszościową czy niejednoznaczność przykładów brzegowych [10, 11, 12, 13, 14].

W celu poprawienia jakości klasyfikatorów możliwe są dwa kierunki postępowania: modyfikacja algorytmów uczących lub/i modyfikacja zbioru danych [11].

Modyfikacja algorytmów wymaga wprowadzenia zmian, które mają wpływ na etap uczenia klasyfikatora. Zmiany mogą dotyczyć na przykład dostosowania prawdopodobieństw apriori czy wprowadzenia macierzy kosztów.

Modyfikacja zbioru danych polega na zmianie poziomu niezrównoważenia, zmianie liczebności zbioru czy dekompozycji klas na podobszary. Zmianę poziomu niezrównoważenia realizuje się poprzez zastosowanie takich metod jak [11]: nadpróbkowanie (ang. oversampling), podpróbkowanie (ang. undersampling). Metoda nadpróbkowania polega na zwiększaniu w zbiorze uczącym liczby przypadków klasy mniejszościowej. Metoda podpróbkowania polega na zmniejszaniu przypadków klasy większościowej. Metody te można ze sobą łączyć co pozwala na zachowanie liczebności zbioru.

4. Ocena klasyfikatorów w analizie danych niezrównoważonych

W diagnostyce mikroskopowej wiedza o sposobie klasyfikowania obiektów na podstawie parametrów morfometrycznych dostępna jest tylko w oparciu o analizę zbioru uczącego, dla którego znana jest prawidłowa klasyfikacja obiektów. Wynikiem wielu prac nad metodami klasyfikacji jest niezliczona liczba opracowanych algorytmów, które różnią się złożonością, jakością klasyfikacji, szybkością działania, szybkością uczenia, ograniczeniami pamięci komputerów. Skuteczne modele klasyfikujące to takie, które potrafią udzielać poprawnych odpowiedzi także dla danych, które nie były dostępne w czasie uczenia, a pochodzą z tej samej dziedziny.

W analizie danych niezrównoważonych miara klasyfikacji powinna uwzględniać szczególny wpływ klasy mniejszościowej. Algorytmy klasyfikacji powinny maksymalizować liczbę poprawnych wskazań w klasie mniejszościowej (TP, ang. true positive) i minimalizować liczbę błędnych wskazać klasy większościowej (FP, ang. false positive).

Jedną z najczęściej proponowanych w literaturze metryk do oceny jakości klasyfikacji jest miara $G-mean$ [12], opisana zależnością $G-mean=\sqrt{sen\cdot spe}$, gdzie sen – czułość klasyfikatora, spe – swoistość klasyfikatora. Wskaźnik G-mean osiągnie wysoką wartość, gdy oba wskaźniki sen i spe będą zbliżone do 1. Mniejsza wartość miary $G-mean$ pojawi się, gdy chociażby jeden ze wskaźników będzie posiadał małą wartość.

Inną miarą oceny jakości klasyfikacji w analizie danych niezrównoważonych jest miara $F-value$ [10], opisana wyrażeniem $F-value=\frac{\left(1+\beta^2\right)\cdot sen\cdot ppv}{\beta^2\cdot sen+ppv}$, gdzie ppv oznacza predykcyjną wartość dodatnią klasyfikatora, nazywaną także precyzją. Miara ta opisuje zależność pomiędzy trzema wartościami: TP, FP, FN (ang. false negative). Współczynnik β odpowiada relatywnej ważności wskaźnika ppv względem sen i najczęściej przyjmuje wartość $\beta=1$.

5. Zastosowanie algorytmów uczenia nadzorowanego we wstępnej diagnostyce raka pęcherza moczowego

Do wstępnej diagnostyki raka pęcherza moczowego zastosowano narzędzie Rough Sets Analysis Toolbox - autorski przybornik środowiska obliczeniowego MATLAB [15]). Głównym zadaniem realizowanym przez RSAToolbox jest wykorzystanie teorii zbiorów przybliżonych do redukcji przestrzeni cech i klasyfikacji danych. Przybornik umożliwia także przeprowadzenie klasyfikacji przy użyciu funkcji dostępnych w przyporniku Statistics Toolbox takich, jak: Naiwny klasyfikator Bayesa (NaiveBayes), Analiza dyskryminacyjna (classify), Drzewa decyzyjne (classregree).

Ocenę klasyfikatorów przeprowadzono z wykorzystaniem miar $G-mean$ oraz $F-value$. Zdefiniowano także dodatkową miarę $FMaxSen$ będącą iloczynem współczynników czułości i swoistości, jako $FMaxSen=sen^2\cdot spe$.

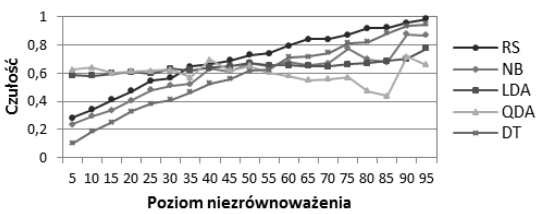
Zastosowanie drugiej potęgi współczynnika czułości podkreśla jego ważność i umożliwia poszukiwanie klasyfikatora charakteryzującego się maksymalną swoistością przy maksymalnej czułości.

Klasyfikację danych niezrównoważonych przeprowadzono pięcioma metodami: Zbiory przybliżone (RS), Naiwny Klasyfikator Bayesa (NB), Liniowa Analiza Dyskryminacyjna (LDA), Kwadratowa Analiza Dyskryminacyjna (QDA), Drzewa Decyzyjne (DT). Dla każdej z metod wyznaczono czułość, swoistość oraz precyzję – współczynniki niezbędne do obliczenia miar jakości klasyfikatora. Charakterystyki współczynników jakości klasyfikatorów zamieszczono na rysunkach 2-4.

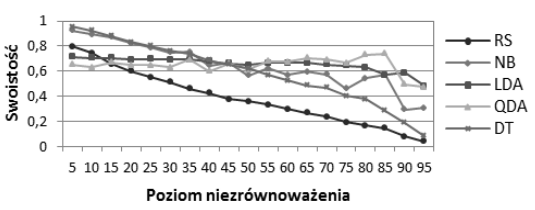
W analizie zastosowano metodę podpróbkowania. Liczbę przypadków komórek zdrowych zmniejszano w taki sposób, aby zawartość komórek nowotworowych w zbiorze była o 5% większa od poprzedniej liczebności. Liczba komórek rakowych w każdym kroku pozostała niezmieniona. Ponieważ w inicjalnym zbiorze uczącym komórki nowotworowe stanowią niecałe 3%, to zmiana poziomu niezrównoważenia prowadzi do zmniejszenia liczebności generowanego zbioru uczącego.

Obserwując wykresy czułości (rys. 2) i swoistości (rys. 4) można wyróżnić dwie grupy klasyfikatorów. Klasyfikatory RS, NB, DT są silnie zależne od poziomu niezrównoważenia. Na ocenę przynależności przypadków do klas wpływa rozkład klas zbioru uczącego. Klasyfikatory LDA i QDA nie zależą od poziomu niezrównoważenia, a podczas klasyfikacji uwzględniają faktyczne położenie punktów względem funkcji dyskryminacyjnych.

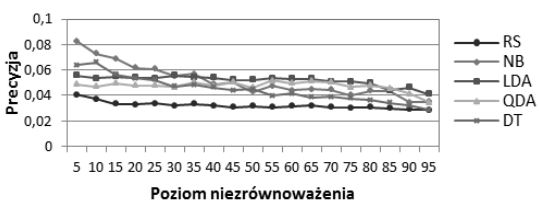
Niezależnie od zastosowanej metody klasyfikacji widać, że wraz ze wzrostem czułości maleje swoistość klasyfikacji. Zmniejszeniu ulega także precyzja.



Rys. 2. Charakterystyka czułości metod klasyfikacji dla różnych poziomów niezrównoważenia
Fig. 2. Sensitivity of classification methods on different levels of imbalance



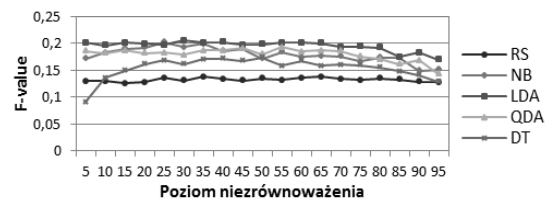
Rys. 3. Charakterystyka specyficzności metod klasyfikacji dla różnych poziomów niezrównoważenia
Fig. 3. Specificity of classification methods on different levels of imbalance



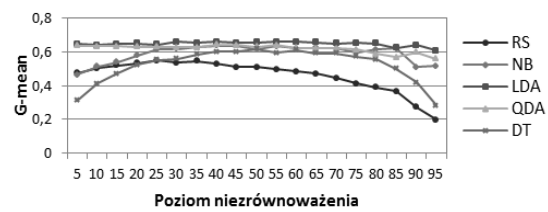
Rys. 4. Charakterystyka precyzji metod klasyfikacji dla różnych poziomów niezrównoważenia
Fig. 4. Precision of classification methods on different levels of imbalance

Przy poziomie niezrównoważenia równym 50%, poszczególne klasyfikatory osiągają zbliżoną czułość. Największą czułość, wynoszącą 0.72, osiągnięto dla klasyfikatora RS. Jednak dla tego poziomu niezrównoważenia klasyfikator charakteryzuje się najniższą swoistością, wynoszącą 0.35. Wartość ta odbiega od pozostałych klasyfikatorów dla których swoistość wynosi ok. 0.6.

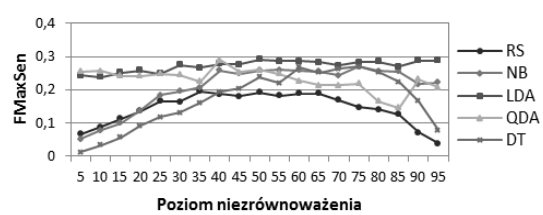
W rozpatrywanym zagadnieniu bardzo ważne jest uzyskanie jak największej czułości. Wysoka czułość wiąże się z możliwym wykryciem komórek nowotworowych. W sytuacji, gdy pole preparatu zostanie błędnie sklasyfikowane jako nowotworowe, jedynym kosztem jest czas poświęcony na wykonanie dokładnej analizy FISH [16]. W momencie, gdy pole zawierające komórki nowotworowe nie zostanie prawidłowo wskazane, traci się możliwość wczesnego wykrycia choroby. Dlatego ważny jest dobór takiej próby uczącej i takiej miary klasyfikacji, dla której możliwe jest uzyskanie jak największej czułości. W tym celu przeprowadzono ocenę miar, których charakterystyki zamieszczono na wykresach 5-7.



Rys. 5. Charakterystyka F-value metod klasyfikacji dla różnych poziomów niezrównoważenia
Fig. 5. F-value of classification methods on different levels of imbalance



Rys. 6. Charakterystyka G-mean metod klasyfikacji dla różnych poziomów niezrównoważenia
Fig. 6. G-mean of classification methods on different levels of imbalance



Rys. 7. Charakterystyka miary FMaxSen metod klasyfikacji dla różnych poziomów niezrównoważenia
Fig. 7. FMaxSen of classification methods on different levels of imbalance

Z przedstawionych charakterystyk wynika, że najmniej korzystną miarą przy ocenie klasyfikacji komórek nowotworowych jest miara $F-value$ (rys. 5). Miara ta pozwala na wybór klasyfikatora (najlepszym klasyfikatorem jest w LDA), lecz praktycznie nie zależy od poziomu niezrównoważenia zbioru uczącego i dlatego nie nadaje się do określenia optymalnego poziomu niezrównoważenia zbioru.

Miara $G-mean$ w większym stopniu nadaje się do wyboru poziomu niezrównoważenia zbioru uczącego. Dla większości metod klasyfikacji optymalny poziom niezrównoważenia zbioru uczącego wynosi ok. 50%. Jedynie dla metody RS optymalny poziom niezrównoważenia wynosi ok. 25%. Dla najlepszej metody klasyfikacji, jaką jest LDA, czułość oraz swoistość wynoszą ok. 0.65, a $G-mean$ przyjmuje wartość ok. 0.63. Jedynie dla metody RS maksymalna wartość $G-mean$ wynosi nieco ponad 0.5.

Miara *FMaxSen* , uwzględniająca w większym stopniu czułość, może być zastosowana tak do wyboru metody klasyfikacji, jak również do optymalnego doboru poziomu niezrównoważenia zbioru uczącego. Największą wartość tej miary uzyskuje się dla metody LDA przy poziomie niezrównoważenia zbioru uczącego na poziomie 50%. Wskaźnik, wynoszący 0.29 jest o 0,04 wyższy od miary *FMaxSen* uzyskanej dla zbioru inicjalnego.

Bardzo wysoki przyrost wartości funkcji osiągnięto dla charakterystyk RS, NB, DT. Zmiana poziomu niezrównoważenia do 50% pozwala na osiągnięcie funkcji celu o wartości 0.25. Wymienione klasyfikatory są silnie zależne od rozkładu danych, a przy klasyfikacji zbioru inicjalnego wyznaczone wartości funkcji *FMaxSen* były zbliżone do zera.

Dla każdej z przedstawionych miar najlepszymi klasyfikatorami okazały się klasyfikatory analizy dyskryminacyjnej. Dają wyniki klasyfikacji prawie niezależne od poziomu niezrównoważenia. Jednak zwiększenie poziomu niezrównoważenia zbioru uczącego korzystniej wpływa na klasyfikator LDA. W klasyfikatorze QDA obserwuje się pogorszenie się możliwości klasyfikacyjnych przy zwiększeniu poziomu niezrównoważenia zbioru uczącego.

6. Podsumowanie

Analizę danych niezrównoważonych przeprowadzono dla klasyfikacji zbioru komórek zdrowych i nowotworowych pęcherza moczowego. Ze względu na ważność współczynnika czułości w ocenie klasyfikatora, zaproponowano miarę *FMaxSen* w większym stopniu uwzględniającą czułość klasyfikacji niż swoistość. Przedstawione charakterystyki pokazują, że wpływ zmiany poziomu niezrównoważenia na jakość klasyfikacji zależy ściśle od typu klasyfikatorów. Największą poprawę otrzymano dla klasyfikatorów NB, DT, RS, które podczas klasyfikacji uwzględniają rozkład klas. Najlepsze rozróżnienie komórek otrzymano dla klasyfikatorów LDA i QDA. Dla klasyfikatora LDA odnotowano także pozytywny wpływ zmiany poziomu niezrównoważenia na klasyfikację komórek nowotworowych pęcherza moczowego.

7. Literatura

[1] Brown T.A.: Genomy, Wydawnictwo Naukowe PWN. Warszawa, 2001.

[2] Zając M., Wiśniewska M.: Zastosowanie fluoroscencyjnej hybrydyzacji in situ (FISH) w identyfikacji zmian materiału genetycznego u osób z niepełnosprawnością intelektualną. Nowiny Lekarskie 2003, 72, 1, s. 9-13.

[3] Oliveira A.M., French C.A.: Applications of Fluorescence in Situ Hybridization in Cytopathology. Acta Cytologica. Vol. 49,No. 6, 2005.

[4] Zieliński K., Strzelecki M.: Wybrane zagadnienia ocen ilościowych i przetwarzania obrazów. Biocybernetyka i inżynieria biomedyczna. T8. Obrazowanie biomedyczne. Red. L. Chmielewski, J. Kulilkowski, A. Nowakowski. Warszawa 2003. Akadem. Oficyna. Wydaw. Exit.

[5] Plesch A., Loerch T.: Metafer a Novel Ultra High Throughpt Scanning System for Rare Cell Detection and Automatic Interphase FISH Scoring. Early Prenatal Diagnosis, Fetal Cells and DNA in the Mother, Present State and Perspectives. 12th Fetal Cell Workshop, Prague, May 2001, pp. 329-339.

[6] Niemiewski M.: Rekonstrukcja i segmentacja obrazów w morfologii matematycznej . Biocybernetyka i inżynieria biomedyczna. T8. Obrazowanie biomedyczne. Red. L. Chmielewski, J. Kulilkowski, A. Nowakowski. Warszawa 2003. Akadem. Oficyna. Wydaw. Exit, s. 83-125.

[7] Zieliński K.: Parametry morfometryczne wykorzystywane w pomiarach biomedycznych. Biocybernetyka i inżynieria biomedyczna. T8. Obrazowanie biomedyczne. Red. L. Chmielewski, J. Kulilkowski, A. Nowakowski. Warszawa 2003. Akadem. Oficyna. Wydaw. Exit, s. 165-177.

[8] Guz T.: Poprawa efektywności klasyfikatora „Box Classifier" w systemie „Metafer". XIII Konferencja „Sieci i Systemy Informatyczne", Łódź, 2005.

[9] Szydłowska (Piotrowska) E. „Implementation of dimensionality reduction method in analysis of cell morphometric features", X International PhD Workshop, OWD'2008, 18-21 October 2008, s. 129-132.

[10] Chawla N.: Data Mining for Imbalanced Datasets: An Overview. Data Mining and Knowledge Discovery Handbook. Maimon O. Rokach L., 2010, Part 6, 875-886.

[11] Fernández A., García S., Herrera F.: Addressing the Classification with Imbalanced Data: Open Problems and New Challenges on Class Distribution. Hybrid Artificial Intelligent Systems. Lecture Notes in Computer Science, 2011, Vol.6678/2011, pp. 1-10.

[12] García V., Sánchez J.S., Mollineda R.A.: Exploring the Performance of Resampling Strategies for the Class Imbalance Problem. Trends in Applied Intelligent Systems Lecture Notes in Computer Science, 2010, Volume 6096/2010, pp. 541-549.

[13] Japkowicz N.: Learning from Imbalanced Data sets: A Comparison of Various Strategies. In Proceedings of the AAAI'2000 Workshop on Learning from Imbalanced Data Sets, Austin, TX.

[14] Stefanowski J., Wilk S.: Combining Rough Sets and Rule based Classifiers for Handling Imbalanced Data. In: Czaja L. (ed.) Proceedings of Concurrency, Specification and Programming CS&P 2005 Conference, vol. 2, 2005, 497-508.

[15] Piotrowska E., Stanisławski W.: Zastosowanie Rough Sets Analysis Toolbox pakietu MATLAB w zadaniach rozpoznawania wzorców. XVII Krajowa Konferencja Automatyki, Kielce, 2010, Streszczenia referatów, s. 99-100.

[16] Daniely M., Rona R., Kaplan T., Olsfanger S., Elboim L., Zilberstiena Y., Freibergera A.,Kidronc D. Lew S., Leibovitch I.: Combined analysis of morphology and fluorescence in situ hybridization significantly increases accuracy of bladder cancer detection in voided urine samples. Urology. Vol. 66, I.6, 2005, pp. 1354-1359.

otrzymano / received: 15.11.2011
przyjęto do druku / accepted: 02.07.2012 *artykuł recenzowany / revised paper*