Biuletyn WAT
Vol. LXV, Nr 1, 2016

Porównanie liniowych metod PCA (*Principal Component Analysis*) i LDA (*Linear Discriminant Analysis*) zastosowanych do klasyfikacji matryc wzbudzeniowo-emisyjnych wybranych grup substancji biologicznych

MACIEJ LEŚKIEWICZ[1], MIRON KALISZEWSKI, ZYGMUNT MIERCZYK, MAKSYMILIAN WŁODARSKI

[1]PCO Spółka Akcyjna, 03-982 Warszawa, ul. Jana Nowaka-Jeziorańskiego 28, maciej.leskiewicz@pcosa.com.pl
Wojskowa Akademia Techniczna, Instytut Optoelektroniki, 00-908 Warszawa, ul. gen. S. Kaliskiego 2, miron.kaliszewski@wat.edu.pl

Streszczenie. W pracy porównano właściwości dwóch liniowych metod (PCA i LDA) pozwalających na redukcję wymiarów w trakcie analizy cech oraz zbadano wydajność tych dwóch algorytmów w procesie klasyfikacji wybranego materiału biologicznego na podstawie jego wzbudzeniowo-emisyjnych matryc fluorescencyjnych. Stwierdzono, że metoda LDA redukuje liczbę wymiarów (znaczących zmiennych) bardziej efektywnie niż metoda PCA. Za pomocą algorytmu LDA udało się uzyskać względnie dobre rozróżnienie badanego materiału biologicznego.
Słowa kluczowe: analiza cech, spektroskopia fluorescencyjna, klasyfikacja substancji biologicznych
DOI: 10.5604/12345865.1197960

1. Wstęp

Szybkie wykrywanie skażeń biologicznych stało się jednym z priorytetowych zadań zarówno służb działających w zakresie obronności, jak i instytucji zajmujących się monitoringiem środowiska. Obecnie do analizy drobnoustrojów obecnych w wodzie i powietrzu najpowszechniej stosowane są metody hodowlane oraz wykorzystujące analizę DNA lub specyficzności przeciwciał. Jednak techniki te są pracochłonne, a do analizy niezbędne jest pobranie próbki i przeprowadzenie

wieloetapowego procesu. Wraz z pojawianiem się nowych źródeł promieniowania UV i coraz czulszych detektorów, coraz większym zainteresowaniem cieszą się techniki wykorzystujące zjawisko autofluorescencji, co umożliwia scharakteryzowanie substancji na podstawie uzyskanego widma spektralnego. Zaletą tych metod jest wysoka czułość i brak konieczności przygotowania materiału do analizy. Z uwagi na bardzo dużą ilość danych opisujących rejestrowane widma, niezwykle istotnym etapem jest optymalizacja analizy danych. W pracy porównano efektywność dwóch liniowych metod statystycznych PCA i LDA w procesie obróbki danych i klasyfikacji charakterystyk fluorescencyjnych wybranych grup substancji.

2. Podstawy teoretyczne algorytmów i ich zastosowanie

Użyte algorytmy mają redukować wymiarowość przestrzeni danych (liczby zmiennych opisujących zjawisko) poprzez wykrycie zależności liniowych między nimi. Następuje to na drodze obrotu układu współrzędnych w taki sposób, aby zmaksymalizować zmienność danych według przyjętego kryterium dla każdego z algorytmów. W efekcie otrzymuje się składowe główne i wybiera te, które zawierają w sobie najwięcej zmienności danych pierwotnych. Metody te mogą również posłużyć do próby liniowej klasyfikacji, zwłaszcza w przypadku danych wielowymiarowych, gdzie w wyniku redukcji zmiennych następuje ekstrakcja cech w postaci otrzymanych czynników głównych. Dzięki temu można uzyskać naturalny podział na grupy względem cech, które niekoniecznie są łatwe do wychwycenia bez redukcji zmiennych. Dodatkowo opisywane metody dają możliwość wizualizowania wielowymiarowych danych w przestrzeni dwu- lub więcej wymiarowej.

2.1. PCA (*Principal Component Analysis*)

Analiza składowych głównych jest algorytmem opartym o rachunek macierzowy [1, 2]. Celem jest znalezienie macierzy składowych głównych Y reprezentujących w nowej przestrzeni macierz z danymi wejściowymi X, a niezbędne do tego jest policzenie macierzy transformacji W.

W pierwszej kolejności należy utworzyć macierz kowariancji K zmiennych odzwierciedlającą zależności liniowe między nimi:

$$K = \frac{1}{(N-1)} \sum_{i=1}^{N} (x_i - \bar{x}) \cdot (x_i - \bar{x})^T, \qquad \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i,$$

gdzie x_i są wektorami danych tworzącymi macierz X, a N jest ilością danych.

Kolejnym krokiem jest wyznaczenie wektorów własnych w_i i wartości własnych λ macierzy kowariancji.

$$Kw_{\mathbf{i}} = \lambda_i w_{\mathbf{i}}.$$

Po uporządkowaniu wartości własnych od największej do najmniejszej można przyporządkować odpowiadające im wektory własne i zapisać je w postaci macierzy W. Pierwsze z kolumn tej macierzy są wektorami definiującymi kierunki (osie) w przestrzeni najbardziej różnicujące dane. Macierz wektorów własnych W jest poszukiwaną macierzą transformacji, służącą do odnalezienia macierzy składowych głównych Y reprezentujących dane w nowej przestrzeni

$$Y = XW, \qquad W = [w_1, w_2, ..., w_m].$$

Wartości własne λ_i macierzy kowariancji są wariancjami składowych głównych. To na ich podstawie dokonuje się analizy, które składowe są najbardziej znaczące, a co za tym idzie, zawierają największy procent informacji o wejściowym zbiorze danych.

2.2. LDA (*Linear Discriminant Analysis*)

Metoda, pierwotnie zwana liniową dyskryminacją Fishera [1, 3], została skonstruowana dla problemu dwóch klas danych w celu ich rzutowania na jednowymiarowy kierunek, który najlepiej je separuje (rozróżnia). Później metoda została uogólniona dla przypadku wielu klas i wymiarów. Podstawową jej cechą jest sformułowanie problemu, gdzie w przeciwieństwie do metody PCA stosuje się wstępny podział danych ze względu na przynależność do klasy.

Liniowa analiza dyskryminacyjna jest algorytmem obliczeniowo zbliżonym do analizy składowych głównych. Zamiast jednak posługiwać się jedną macierzą kowariancji, definiuje się macierze charakteryzujące zmienność wewnątrzgrupową S_w i międzygrupową S_b.

$$S_w = \sum_{i=1}^{g} \sum_{j=1}^{N_i} (x_{i,j} - \overline{x_i}) \cdot (x_{i,j} - \overline{x_i})^T, \qquad \overline{x_i} = \frac{1}{N_i} \sum_{i=1}^{g} x_i$$

$$S_b = \sum_{i=1}^{g} N_i (\overline{x_i} - \overline{x}) \cdot (\overline{x_i} - \overline{x})^T, \qquad \overline{x} = \frac{1}{N} \sum_{i=1}^{g} \sum_{j=1}^{N_i} x_{i,j},$$

gdzie $x_{i,j}$ jest wektorem j-tej danej pochodzącej z grupy i, a N_i jest ilością danych w tej grupie.

Celem jest znalezienie takiego kierunku w przestrzeni, który maksymalizuje zmienność międzygrupową, jednocześnie minimalizując zmienność wewnątrzgrupową danych, co definiuje szukaną macierz transformacji jako maksymalizację stosunku wyznaczników macierzy S_b i S_w.

$$W_{LDA} = \arg_{\max} \frac{\left|W^T S_b W\right|}{\left|W^T S_w W\right|}, \qquad W_{LDA} = [w_1, w_2, ..., w_{g-1}].$$

Okazuje się, że znalezienie najlepiej różnicujących kierunków sprowadza się do rozwiązania problemu wektorów w_i i wartości własnych λ_i spełniających równanie:

$$S_b w_{\mathbf{i}} = \lambda_i S_w w_{\mathbf{i}}.$$

Rzutowanie (projekcja) macierzy z danymi wejściowymi na nowe kierunki odbywa się tak samo jak w przypadku metody PCA, za pomocą wymnożenia danych przez znalezioną macierz transformacji składającą się z wektorów własnych.

2.3. Graficzna prezentacja różnicy metod LDA i PCA jako liniowych separatorów

Rozpatrując najprostszy przypadek przestrzeni dwuwymiarowej z dwiema grupami, łatwo zwizualizować różnicę w działaniu obydwu metod poprzez odnalezienie takiego kierunku w tejże przestrzeni, który najlepiej różnicuje dane wejściowe. Wówczas klasyczna reguła dyskryminacyjna jest zdefiniowana poprzez rzutowanie wektora danych x na policzony kierunek w i zaklasyfikowanie tego elementu do klasy pierwszej, jeżeli:

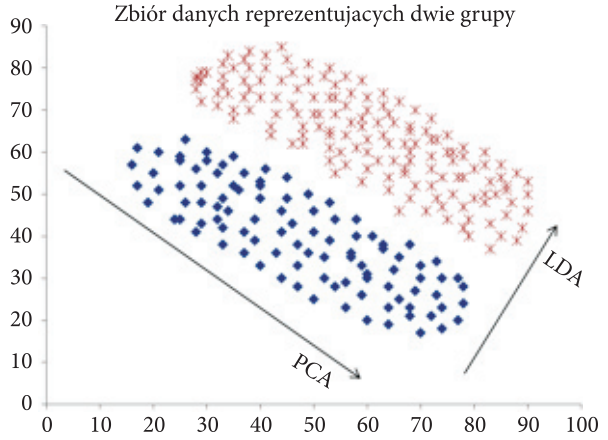
$$\left|w^T x - w^T \overline{x_1}\right| < \left|w^T x - w^T \overline{x_2}\right|,$$

gdzie $\overline{x_1}, \overline{x_2}$ są wektorami średnich dla każdej z grup.

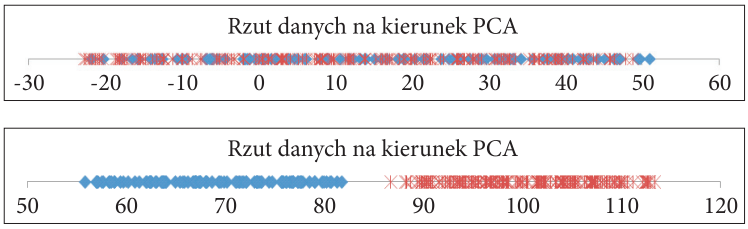
Poniżej przedstawiony jest przykładowy zbiór reprezentujący dwie grupy, które na pozór powinny być w łatwy sposób sklasyfikowane, oraz rzeczywiste obliczenia kierunków wraz z projekcją danych obydwiema metodami.

Jest to dość szczególny przypadek, gdyż obydwie metody prowadzą do rzutowania danych na kierunki niemal do siebie prostopadłe. Algorytm PCA opiera swoje obliczenia na jednej macierzy kowariancji liczonej na całym zbiorze danych i prowadzi do wybrania kierunku ich największego rozproszenia. Z kolei metoda LDA, posiadając informację o przynależności grupowej, odnajduje kierunek minimalizujący rozproszenie wewnątrzgrupowe, jednocześnie oddalając grupy od siebie. Skuteczność metod jako liniowego klasyfikatora można ocenić po zrzutowaniu wszystkich danych na odnaleziony kierunek.

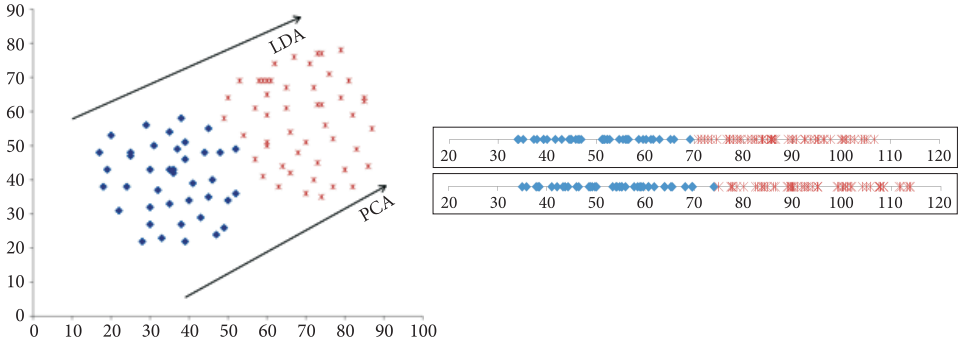
Okazuje się, że w tym przypadku metoda PCA jest bezużyteczna jako klasyfikator, natomiast LDA klasyfikuje dane bezbłędnie. Jest to jednak szczególny przypadek, nieświadczący o bezwzględnej wyższości jednej metody nad drugą. Ogólnie obie metody mogą wykazywać podobną zdolność klasyfikacji liniowej grup bądź jej kompletny brak [4].



Rys. 1. Dwuwymiarowy przykład danych zawierających dwie grupy

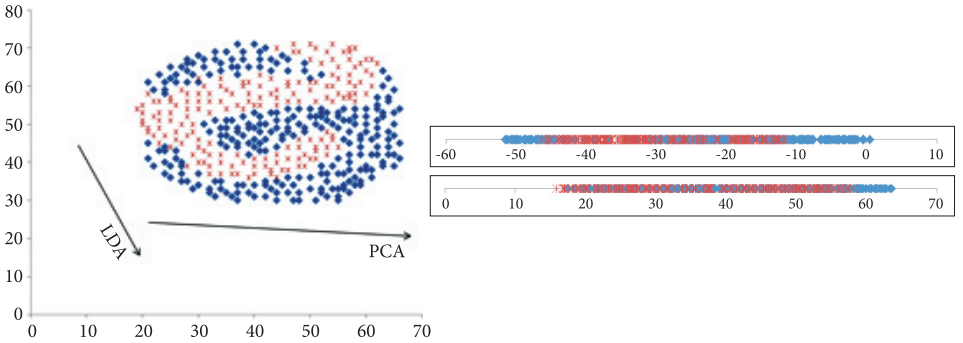


Rys. 2. Rzut danych z rysunku 1 na wybrane kierunki



Rys. 3. Przykład grup, dla których obydwie metody dają podobny wynik

Na przykładach dwóch zmiennych wynik działania algorytmów znajdowania kierunku najbardziej różnicującego dane jest dość prosty do przewidzenia. Sytuacja się komplikuje w momencie wprowadzenia większej ilości zmiennych, gdy nawet liniowe zależności nie są już tak łatwo dostrzegalne.



Rys. 4. Przykład grup, gdzie obie metody (PCA i LDA) nie umożliwiają separacji

3. Materiały i metodyka

3.1. Materiał biologiczny

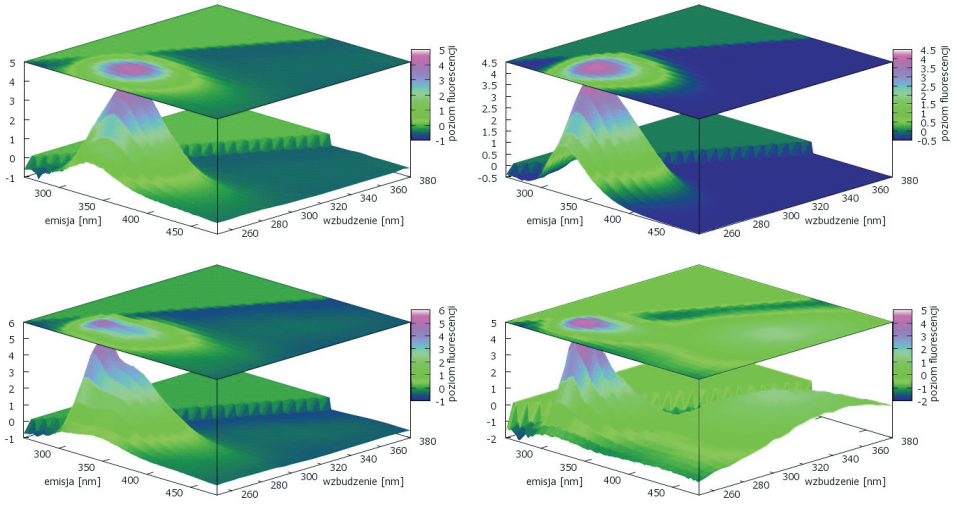
Próbki bakterii, przetrwalników bakterii i grzybów przygotowane zostały w Wojskowym Instytucie Higieny i Epidemiologii przez zespół prof. dr hab. Elżbiety Trafny. Robocze kultury posiewano ze szczepów referencyjnych na odpowiednim podłożu płynnym i inkubowano przez 18 godz. w temperaturze 37°C [5]. Szczepy były zbierane i odwirowywane przy 4000 obr/min w czasie 10 min. Endospory pięciu użytych gatunków *Bacillus* hodowano i czyszczono zgodnie z procedurą opisaną we wcześniejszej publikacji [6]. Do eksperymentów przygotowywano zawiesiny przetrwalników bakterii w sterylnej dejonizowanej wodzie. Badania przeprowadzono również dla *Candida albicans* i czterech szczepów grzybów strzępkowych. Szczep *Candida albicans* wzrastał w płynnej pożywce YPD, a pozostałe grzyby w pożywce MEA. Próbki grzybów do eksperymentów zawieszane były w sterylnej soli fizjologicznej. Pyłki roślinne pochodziły z firm Sigma-Aldrich i Duke Scientific (Fisher). Białka zostały zakupione w firmie Sigma-Aldrich. Zawiesiny pyłków roślin o stężeniu ok. 1-2 mg/ml przygotowane zostały z suchych preparatów i zawieszone w wodzie dejonizowanej. Pomiary fluorescencji wykonano w kuwetkach kwarcowych o drodze optycznej 1 cm.

3.2. Aparatura

Badania fluorescencji materiałów biologicznych przeprowadzono na spektrofluorymetrze FL 900 firmy Edingurgh Inst., stosując kuwety kwarcowe. W celu uniknięcia efektu filtra wewnętrznego, pomiary przeprowadzono, stosując przystawkę odbiciową umożliwiającą zbieranie sygnałów emisji ze wzbudzanej powierzchni (*front-surface measurements*).

3.3. Dane

W doświadczeniu zostały zebrane widma próbek biologicznych dla różnych długości promieniowania wzbudzającego i emitowanego, dzięki czemu powstały matryce wzbudzeniowo-emisyjne. Są to dane wielowymiarowe, co daje możliwość ich statystycznej analizy wieloczynnikowej. W niniejszym opracowaniu zostały opisane dwie liniowe metody analizy: Analiza Składowych Głównych PCA (*Principal Component Analysis*) oraz Liniowa Analiza Dyskryminacyjna LDA (*Linear Discriminant Analysis*). Dane spektralne użyte do analizy są przedstawione w postaci matryc emisyjno-wzbudzeniowych składających się z 26 wierszy odpowiadających długościom promieniowania wzbudzającego z przedziału 250÷380 nm oraz 115 kolumn odpowiadających długościom fal emitowanej fluorescencji z przedziału 260÷490 nm. Zostały one podzielone na cztery grupy: bakterie, białka i aminokwasy, grzyby, pyłki.



Rys. 5. Przykład danych spektralnych użytych do analizy

Tabela 1

Spis substancji użytych w analizie

Bakterie	Białka i aminokwasy	Grzyby	Pyłki
• *Bacillus anthracis spores* • *Bacillus anthracis vegetative in BHI* • *Bacillus atrophaeus spores* • *Bacillus atrophaeus vegetative* • *Bacillus cereus vegetative* • *Bacillus megaterium vegetative* • *Bacillus stearothermophilus vegetative* • *Bacillus subtilis spores* • *Bacillus subtilis vegetative* • *Bacillus thuringensis spores* • *Bacillus thuringensis vegetative* • *Escherichia coli vegetative* • *Micrococcus luteus vegetative* • *Pseudomonas aeruginosa vegetative* • *Staphylococcus aureus vegetative* • *Turex*	• Albumina z surowicy wołowej • Kazeina • Kolagen • Żelatyna • Globuliny • Albumina z jaja kurzego • Pepsyna	• *Alternaria alternata* • *Aspergillus flavus* • *Candida albicans* • *Cladosporium herbarum* • *Penicilium brevi compactum* • *Penicillium chryzogenum* • *Penicillium chryzogenum Notatum spores*	• *Bermuda Grass Pollen* • *Bermuda Grass Smut spores* • *Black Wallnut Pollen* • *Corn Pollen* • *Johnsons grass smut spores* • *Paper Mulberry Pollen* • *Pecan Pollen* • *Ragweed Pollen*

4. Analiza danych

W celu porównania substancji mających różny poziom natężenia emisji, wszystkie dane zostały znormalizowane w ten sam sposób poprzez:

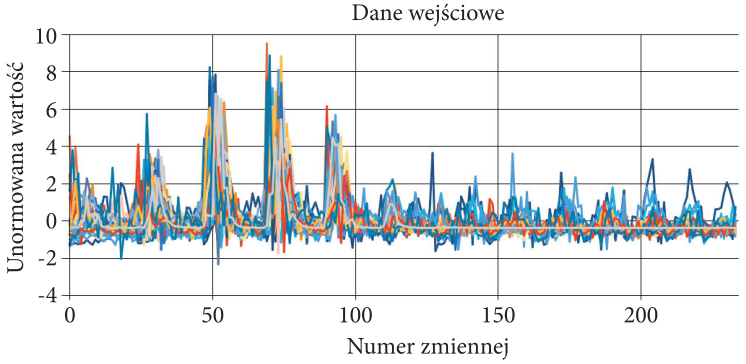
— eliminację składowej stałej stanowiącej średnią wartość fluorescencji policzoną po wszystkich elementach matrycy,
— unormowanie do odchylenia standardowego matrycy σ.

W wyniku takiego sposobu normalizacji wszystkie matryce mają wartość średnią równą zeru oraz wariancję równą jedności, co pozwala porównać kształt charakterystyk wszystkich substancji, zaniedbując różnice w poziomie intensywności ich fluorescencji.

W przypadku analizy metodą LDA istnieje potrzeba zapewnienia odpowiednio dużej liczebności zbioru danych. Istotny przy tym jest stosunek liczby elementów do liczby zmiennych (> 20). W tym celu konieczne było sztuczne zwielokrotnienie liczby matryc poprzez wprowadzenie szumu, tzn. do każdej wartości matrycy została dodana zmienna losowa o rozkładzie Gaussa wymnożona przez wartość sygnału. Takie postępowanie pozwala zwielokrotnić dane, jednocześnie zachowując informację zawartą w niskich poziomach intensywności fluorescencji. Do wygenerowania liczby losowej o rozkładzie normalnym użyto transformacji Boxa-Mullera korzystającej z dwóch zmiennych o rozkładzie jednostajnym z przedziału (0, 1>.

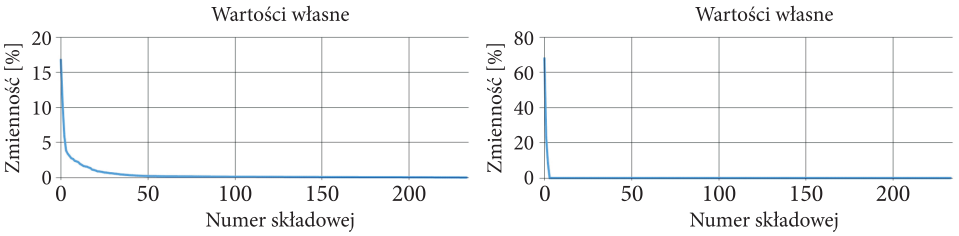
4.1. Analiza matryc wzbudzeniowo-emisyjnych

Do analizy matryc użyto danych spektralnych z krokiem 10 nm zarówno dla fali wzbudzenia, jak i emisji, co ostatecznie daje 234 zmienne dla każdej substancji. Dane zostały zaszumione ($\sigma^2 = 0{,}5$) i zwielokrotnione 125 razy, aby zachować odpowiedni stosunek liczby elementów do zmiennych dla metody LDA. W celu porównania obu metod użyto tych samych danych wejściowych dla obydwu algorytmów:



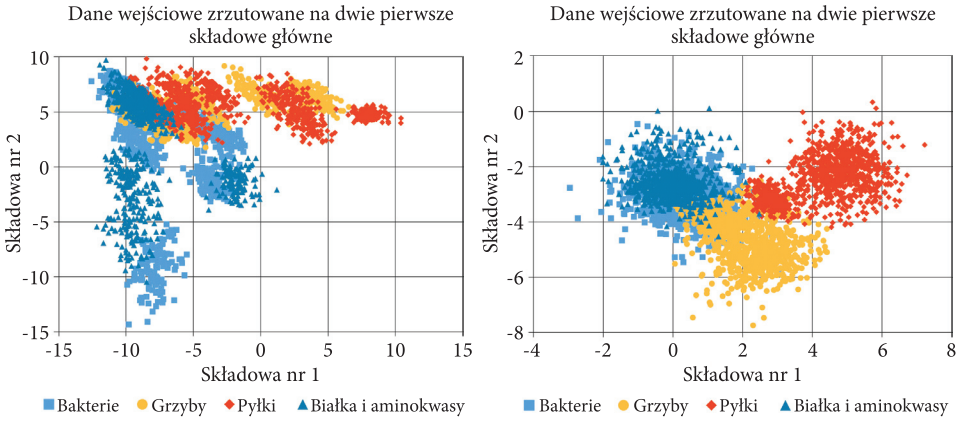
Rys. 6. Dane wejściowe składające się z 234 zmiennych stanowiących elementy matrycy (rys. 5)

Wynik działania algorytmów można prześledzić na podstawie policzonych wartości własnych (rys. 7). Są one wariancjami składowych głównych i można je przedstawić w postaci całkowitej zmienności wyrażonej procentowo (suma po wszystkich wartościach daje 100% zmienności).



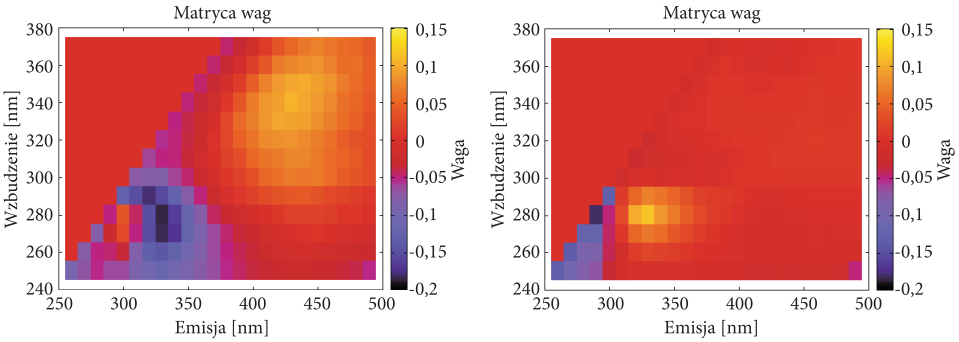
Rys. 7. Wartości własne uporządkowane od największych do najmniejszych dla PCA (po lewej) i LDA (po prawej) wyrażone w procentach zmienności

Można zauważyć, że metoda LDA pokazuje mniejszą liczbę kierunków istotnie separujących. Następnie w celu wizualizacji dokonano projekcji danych na dwa wektory własne odpowiadające dwóm największym wartościom własnym. Poniższe diagramy potwierdzają, że metoda LDA zapewnia lepszą separację.

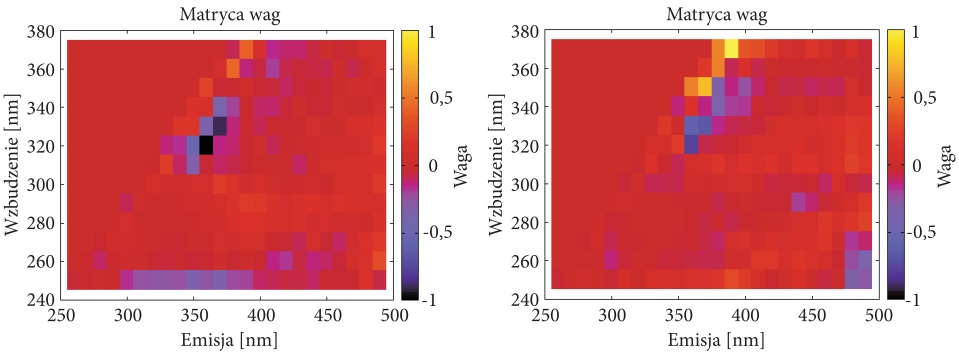


Rys. 8. Wynik działania algorytmów dla dwóch pierwszych składowych głównych dla PCA (po lewej) i LDA (po prawej)

Aby przekonać się, które zmienne pierwotne najbardziej różnicują dane, przedstawiono graficznie również dwa pierwsze wektory własne składające się z 234 cech (wag) odpowiednio dla metody PCA i LDA.



Rys. 9. Współczynniki wagowe pierwszego (po lewej) i drugiego (po prawej) wektora dla metody PCA



Rys. 10. Współczynniki wagowe pierwszego (po lewej) i drugiego wektora dla metody LDA (po prawej)

Wagi bliskie zeru nie wnoszą wkładu do projekcji, natomiast im większa wartość bezwzględna, tym większe znaczenie zmiennej pierwotnej. W metodzie PCA widać korelację pomiędzy wagami wektorów a pierwotną matrycą wzbudzeniowo-emisyjną, gdzie fragmenty matryc o niskiej intensywności wnoszą najmniejszy wkład w zmienność. Z kolei trudno doszukać się takiej analogii w metodzie LDA, która wybiórczo traktuje zmienne pierwotne. Wybiera te, które tworzą najlepszą liniową kombinację dla z góry założonego podziału na grupy.

W celu dokonania klasyfikacji użyto liniowej reguły dyskryminacyjnej zmodyfikowanej dla separacji kilku grup przy użyciu nie jednego, a pięciu najbardziej różnicujących kierunków (wektorów własnych) w_k, gdzie wartości własne λ_k pełnią rolę współczynników. Klasyfikacja każdego elementu x do danej grupy g następuje na drodze policzenia wszystkich odległości d_g *i* wybraniu tej grupy, dla której była ona najmniejsza według wzoru:

$$d_g = \sum_{k=1}^{5} \lambda_k \left| w_k^T x - w_k^T \overline{x_g} \right|,$$

gdzie $\overline{x_g}$ są wektorami średnich dla każdej z grup.

Skuteczność klasyfikacji przyjęto jako stosunek prawidłowo rozpoznanych elementów do ich całkowitej liczby w grupie i przedstawiono w tabeli z rozróżnieniem na poszczególne grupy i metody.

Tabela 2

Wynik liniowej separacji przedstawiony jako procent poprawnie sklasyfikowanych elementów w ramach danej grupy substancji

	PCA [%]				LDA [%]			
Grupy	Bakterie	Białka	Grzyby	Pyłki	Bakterie	Białka	Grzyby	Pyłki
Bakterie	**72**	11	14	3	**69**	27	4	0
Białka i aminokwasy	57	**29**	1	13	34	**64**	2	0
Grzyby	51	2	**18**	29	12	1	**82**	5
Pyłki	30	1	25	**45**	0	0	21	**79**

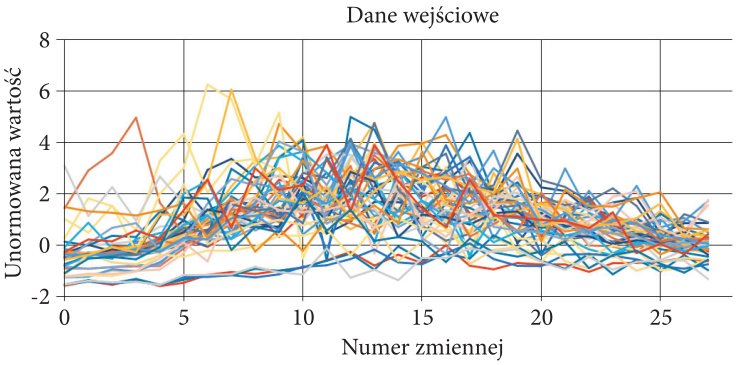
W metodzie PCA trudno wyróżnić grupę skutecznie separującą choćby jedną substancję od pozostałych. W metodzie LDA separacja grup jest dużo skuteczniejsza, zwłaszcza dla pyłków (grupa III). Niestety metoda ta zauważa również duże podobieństwo bakterii i białek (grupa I i IV), co jest bardzo dobrze widoczne na wykresie dwóch pierwszych składowych (rys. 8). Należy jednak podkreślić, że skuteczne odróżnianie pyłków i grzybów od bakterii i białek może pozwolić odróżnić biologiczne komponenty aerozoli pochodzenia naturalnego od generowanych sztucznie (intencjonalnie) i mogących stanowić zagrożenie.

4.2. Analiza pojedynczych linii wzbudzenia fluorescencji

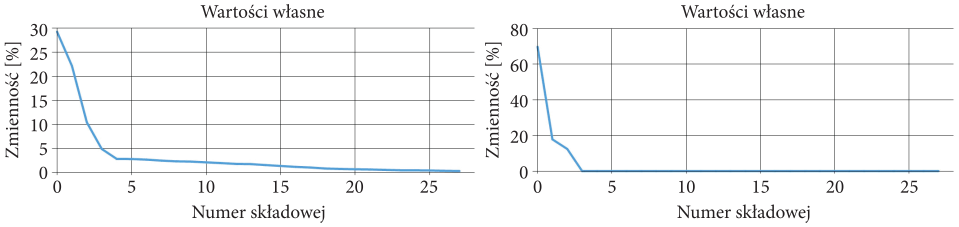
Matryce można traktować jako zbiór osobnych linii wzbudzenia, z których każda ma 28 zmiennych odpowiadających pierwszym 112 nm długości fali emisji. Tak jak uprzednio wprowadzono szumy do danych ($\sigma^2 = 0{,}3$) i jednocześnie zwielokrotniono je 25 razy, zatem macierz danych wejściowych złożona jest teraz z 28 kolumn i 975 wierszy.

Analogicznie w celu analizy matryc uporządkowano wartości własne od największych do najmniejszych. Wartości te przedstawiają procent całkowitej zmienności danych wejściowych. Na ich podstawie dokonano projekcji danych, ale ze względu na mnogość linii wzbudzenia do przykładowej prezentacji wybrano długość fali 265 nm.

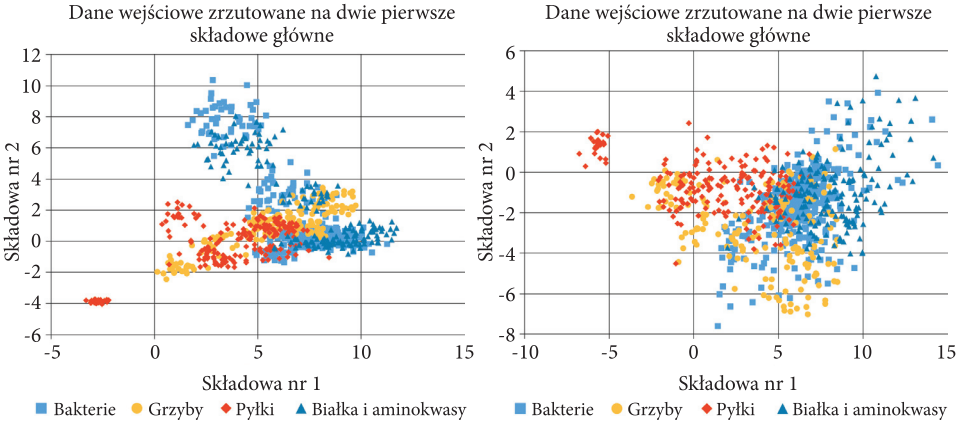
Jak widać, ograniczenie się do jednej długości fali wzbudzenia (zmniejszenie liczby informacji) zdecydowanie pogarsza separację wybranych czterech grup, czego należało się spodziewać. W celu wyłonienia linii wzbudzenia najlepiej separującej wszystkie grupy substancji wykonano procedurę dyskryminacji na podstawie tej samej liniowej reguły dyskryminacyjnej co w przypadku matryc, z tą różnicą,



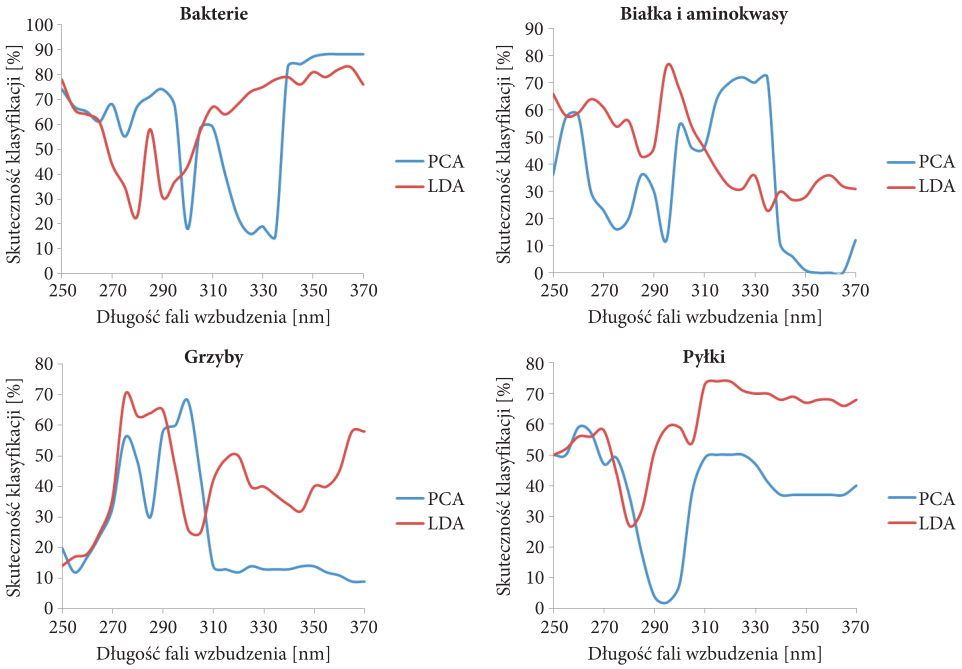
Rys. 11. Dane wejściowe składające się z 28 zmiennych odpowiadających pierwszym 112 nm długości fali emisji. Przykład dla długości fali wzbudzenia 265 nm



Rys. 12. Wartości własne uporządkowane od największych do najmniejszych, dla PCA (po lewej) i LDA (po prawej), wyrażone w procentach zmienności. Przykład dla długości fali wzbudzenia 265 nm



Rys. 13. Wynik działania algorytmów dla dwóch pierwszych składowych głównych dla PCA (po lewej) i LDA (po prawej). Przykład dla długości fali wzbudzenia 265 nm



Rys. 14. Skuteczność klasyfikacji różnych materiałów biologicznych dla pojedyńczych linii wzbudzenia

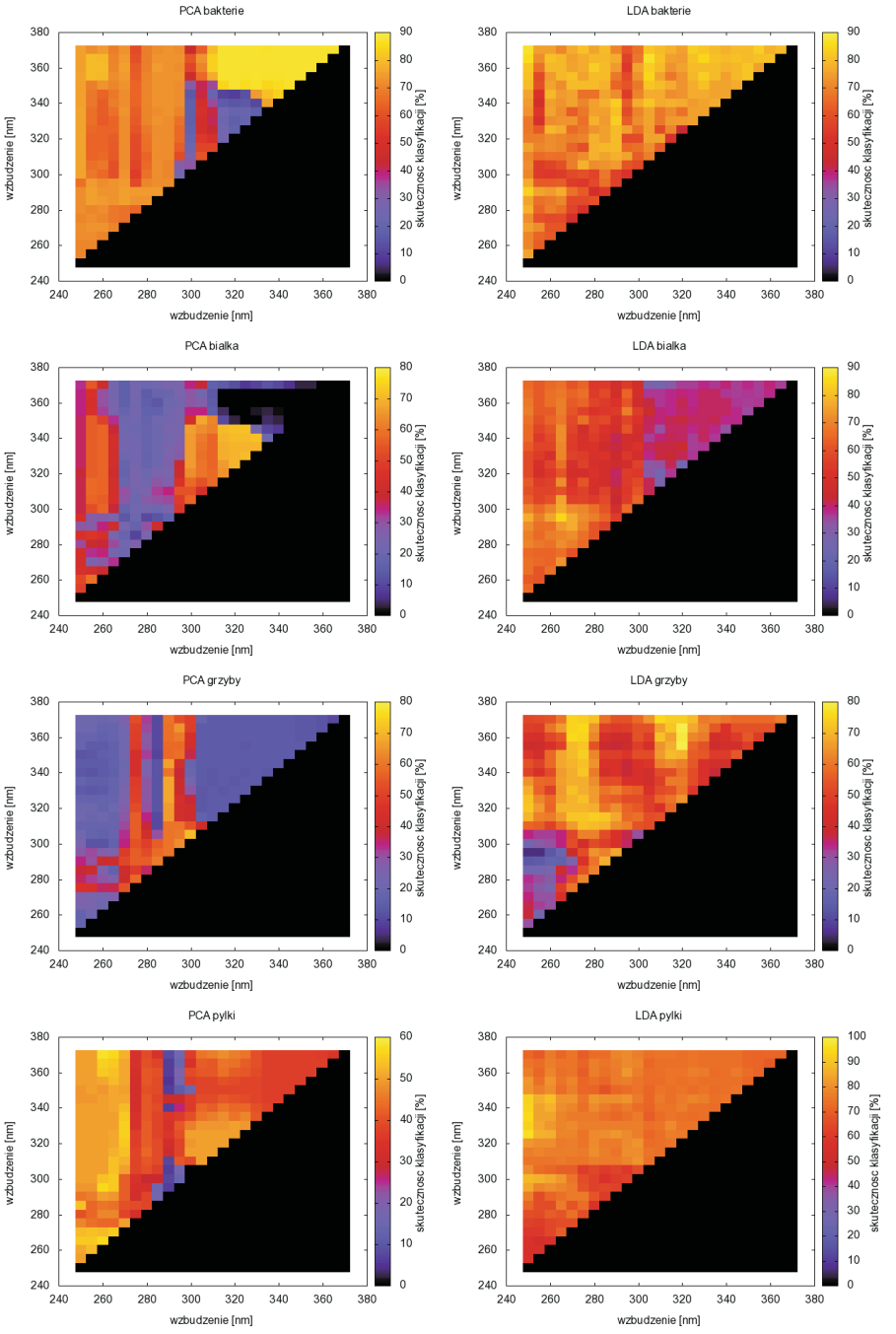
że powtórzono to dla wszystkich długości fal wzbudzenia osobno. Ponadto, jako skuteczność klasyfikacji uwzględniono tylko procentowe wartości poprawnie sklasyfikowanych elementów (znajdujące się na przekątnej tabeli 2).

Przyjęte kryterium klasyfikacji nie daje możliwości wybrania jednej długości fali wzbudzenia skutecznie separującej wszystkie grupy. Rejon krótkofalowy nie pozwala na separację grzybów, podczas gdy rejon długofalowy nie pozwala na separację białek i aminokwasów.

4.3. Analiza pary linii wzbudzenia fluorescencji

W tym przypadku analiza została przeprowadzona analogicznie do poprzedniej z tą różnicą, że dane wejściowe składały się nie z jednej, a z dwóch połączonych linii wzbudzenia. Dla wszystkich ich kombinacji zostały przeliczone skuteczności detekcji w ten sam sposób jak poprzednio.

Analiza par linii wzbudzenia potwierdza wyniki uzyskane za pomocą pojedynczych linii wzbudzenia. Nie ma jednego obszaru, który skutecznie klasyfikowałby wszystkie grupy. Wprowadzenie drugiej linii emisji daje pewną korzyść dla pyłków i białek, gdzie obszar maksimum zawiera w sobie zarówno krótko-, jak i długofalowe wzbudzenie. Ciekawy jest również długofalowy obszar maksimum dla bakterii,



Rys. 15. Skuteczność klasyfikacji określona dla wszystkich kombinacji dwóch długości fali wzbudzenia

które w tym obszarze mają bardzo niską fluorescencję i to właśnie dla algorytmów stanowi ich charakterystyczną ich cechę.

5. Wnioski

Zostały przeprowadzone analizy PCA i LDA charakterystyk fluorescencyjnych wybranych materiałów biologicznych. Analizowano trzy typy danych: dwuwymiarowe matryce emisyjno-wzbudzeniowe, pojedyncze linie wzbudzenia oraz kombinacje dwóch linii wzbudzenia. We wszystkich przypadkach korzystano z tej samej liniowej reguły dyskryminacyjnej.

Analiza matryc wykazała dużo większą skuteczność dyskryminacji metodą LDA niż PCA (tab. 2), skutecznie odróżnia ona pyłki i grzyby od bakterii i białek, co jest wartościowym wynikiem. Należy jednak zwrócić uwagę, że w powyższej analizie zaprezentowano jedynie potencjalną zdolność algorytmów do separacji substancji, nie została ona jednak w pełni zweryfikowana. W tym celu należałoby użyć większej liczby danych z podziałem na dane uczące i testowe. Przy liczbie i typie danych wykorzystanych w analizie (mało próbek mocno zróżnicowanych wewnątrz grupy) wyniki byłyby bardzo zależne od dokonanego podziału.

Bez wątpienia oba algorytmy bardzo skutecznie redukują liczbę zmiennych niezbędnych do opisu, o czym świadczy rozkład wartości własnych (rys. 7 i rys. 12). Metoda LDA wymaga mniejszej liczby składowych głównych niż PCA.

Żadna z zaprezentowanych metod nie pozwala wybrać pojedynczej linii emisji lub ich pary, za pomocą których można dokonać jednoczesnej i równie skutecznej separacji wszystkich grup substancji względem siebie, tak jak w przypadku użycia matryc. Spowodowane jest to przede wszystkim ograniczoną liczbą informacji. Przyczyny należy także szukać w przestrzennych rozkładach wyników, gdzie poszczególne grupy nie są wewnętrznie zwarte (składają się z kilku skupisk) i nie są też rozłączne względem siebie (występuje przekrywanie się obszarów z elementami pochodzącymi z różnych grup).

Podziękowania

Dziękujemy zespołowi pani prof. dr hab. Elżbiety Trafny z Wojskowego Instytutu Higieny i Epidemiologii za przygotowanie próbek materiału biologicznego, którego charakterystyki fluorescencyjne stały się bazą danych dla analiz przedstawionych w powyższym artykule.

Pragniemy również podziękować płk. dr. inż. Krzysztofowi Kopczyńskiemu oraz dr. inż. Zbigniewowi Zawadzkiemu z Instytutu Optoelektroniki WAT za wielokrotną merytoryczną dyskusję mającą bezpośredni wpływ na kształt artykułu.

Projekt finansowany ze środków Narodowego Centrum Badań i Rozwoju Nr Umowy ROB 0031 01/ID/31/1.

Artykuł wpłynął do redakcji 23.02.2015 r. Zweryfikowaną wersję po recenzjach otrzymano 18.02.2016 r.

LITERATURA

[1] Bishop C.M., *Pattern Recognition and Machine Learning*, Springer, Discriminant Functions, 181-196, Continuous Latent Variables, 2006, 559-570.

[2] Hardle W., Simar L., *Applied Multivariate Statistical Analysis*, Springer, Principal Component Analysis, 2003, 233-275.

[3] Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning*, Springer, Linear Discriminant Analysis, 2008, 106-119.

[4] Fukunaga K., *Introduction to Statistical Pattern Recognition*, Academic Press, Discriminant Analysis, 1990, 445-455.

[5] Isenberg H.D. (ed.), *Clinical Microbiology Procedures Handbook*, ASM Press, 2004.

[6] Włodarski M. et al., Proc. SPIE, 6398, 2006, 6-1-12.

M. LEŚKIEWICZ, M. KALISZEWSKI, Z. MIERCZYK, M. WŁODARSKI

Comparison of Principal Component Analysis and Linear Discriminant Analysis applied to classification of excitation-emission matrices of the selected biological material

Abstract. Quality of two linear methods (PCA and LDA) applied to reduce dimensionality of feature analysis is compared and efficiency of their algorithms in classification of the selected biological materials according to their excitation-emission fluorescence matrices is examined. It has been found that LDA method reduces the dimensions (or a number of significant variables) more effectively than PCA method. A relatively good discrimination within the examined biological material has been obtained with the use of LDA algorithm.

Keywords: Feature Analysis, Fluorescence Spectroscopy, Biological Material Classification
DOI: 10.5604/12345865.1197960