STUDIA INFORMATICA 2011
Volume 32 Number 2A (96)

Jolanta KAWULOK, Joanna POLAŃSKA
Politechnika Śląska, Instytut Automatyki

STATYSTYCZNA ANALIZA DANYCH PROTEOMICZNYCH

Streszczenie. Celem prac przedstawionych w niniejszym artykule była konstrukcja narzędzi do analizy danych proteomicznych, które pomogą w wykrywaniu choroby nowotworowej we wczesnym jej stadium. W prezentowanym artykule na początku została przeprowadzona wstępna obróbka widm masowych. Po zamodelowaniu ich za pomocą mieszanin gaussowskich wykorzystano klasyfikator SVM do rozdzielenia grupy z rakiem od grupy kontrolnej. Przedstawione wyniki badań potwierdziły skuteczność wykonanych operacji.

Słowa kluczowe: widmo masowe, klasyfikacja, modele mieszanin gaussowskich

STATISTICAL METHODS FOR ANALYSING PROTEOMIC DATA

Summary. The aim of the work reported in this paper was to develop statistical tools for mass spectra analysis. They would make it possible to detect cancer at its early stages. The main goal was to construct a classifier which would best distinguish people with cancer from a control group. First, the mass spectral signal is preprocessed. Next, the signals are modeled using Gaussian mixtures and they are later classified. The obtained results confirmed the effectiveness of the presented method.

Keywords: mass spectrometry, classification, Gaussian Mixture Models

1. Wprowadzenie

Międzynarodowy projekt poznania ludzkiego genomu (ang. *Human Genome Project*) pomógł poznać sekwencje nukleotydów w genomie człowieka. Jednak wiedza na temat, jakie białka są kodowane przez genom ciągle jest bardzo uboga. Należy też wspomnieć, że wszystkie komórki w organizmie mają wspólny genom, jednak w różnych typach komórek występuje ekspresja innych jego fragmentów. Istnieje wiele baz danych dotyczących poznanych już

białek w organizmach, które skupiają w sobie wiedzę na temat ich budowy przestrzennej oraz funkcji. Obecnie najbardziej znaną bazą danych jest PDB (ang. *Protein Data Bank*), w której znajduje się 70813 opisanych struktur, w tym 65545 samych białek[1]. Zestaw białek, który pochodzi z danego organizmu określa się mianem proteomu, a naukę o nim – proteomiką. Obecnie prężnie rozwijającą się nauką jest proteomika kliniczna, w ramach której są analizowane zmiany rozwoju choroby w czasie, jak również poszukuje się różnic w proteomie osób zdrowych i chorych [12].

Analizując proteom stosuje się najczęściej spektrometry mas (MS – ang. *mass spectrometry*). Urządzenie to, jako wynik zwraca zależność natężenia prądu jonowego odpowiadającego poszczególnym jonom od stosunku masy tych jonów do ich ładunku (m/z). Wynik ten przedstawia się jako tzw. widmo masowe. Następnie są wykorzystywane różne metody analizy matematycznej, pozwalające uzyskać istotną wiedzę o badanej próbce [9, 12]. Zbierając informacje o badanej tkance w postaci widm masowych od osób z określonym nowotworem, jak również od osób zdrowych, należących do podobnej grupy (m.in.: wiek, płeć, inne przebyte choroby), istnieje możliwość identyfikacji biomarkerów białek, których występowanie wskazuje na chorobę. Obecnie prowadzone są liczne badania nad metodami, które pomogą we wczesnym diagnozowaniu zmian nowotworowych [1, 9, 10]. Mając wyznaczone różnicujące piki w widmie, na podstawie wartości m/z oraz ogólnych informacji o sposobie wykonania badań, istnieje możliwość identyfikacji białek na podstawie informacji gromadzonych w bazach danych (np. EPO-KB, Mascot).

W przedstawionych badaniach skoncentrowano się na analizie widm masowych poprzez ich wstępną obróbkę oraz wykorzystaniu SVM w celu klasyfikacji obu grup. Dane użyte do analizy pochodzą z Zakładu Radiobiologii Doświadczalnej i Klinicznej z Centrum Onkologii w Gliwicach. Została tam pobrana i wyizolowana surowica z krwi od 49 zdrowych osób (jako kontrola) oraz od 38 pacjentów z rakiem płuc. W spektrometrii mas użyto metody MALDI-TOF. Procesy izolacji i analizy w spektrometrze surowicy zostały powtórzone kilkakrotnie dla każdej osoby, w celu eliminacji przypadkowych zakłóceń [11].

Dostarczone wyniki w postaci MS sygnałów zostały poddane wstępnej obróbce poprzez usunięcie wartości odstających, uśrednieniu, usunięciu tła, normalizacji (rozdział 2). Następnie użyto modelu mieszanin gaussowskich do opisu widm (rozdział 3), aby można było je klasyfikować. W rozdziale 4 przedstawiono ogólne informacje o budowie klasyfikatora, wyznaczaniu błędów, a w rozdziale 5 omówiono wyniki badań eksperymentalnych. Podsumowanie przeprowadzonej pracy zostało przedstawione w rozdziale 6.

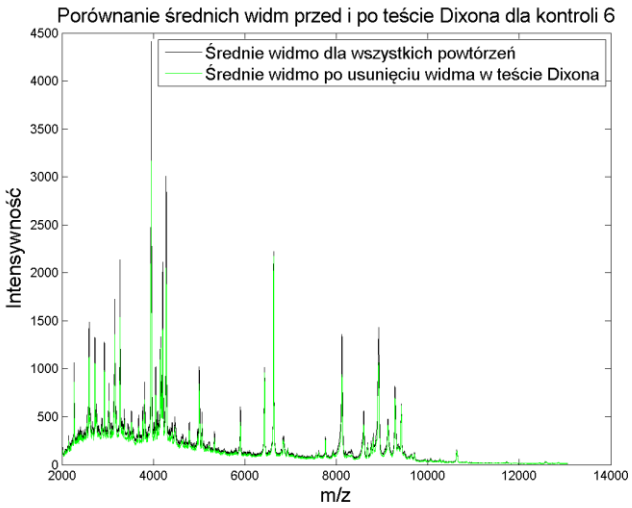
[1] Informacja z bazy danych PDB (www.pdb.org) z 25.01.2011.

2. Wstępne przetworzenie sygnału po spektrometrii mas

Podczas wykonywania eksperymentu na poziomie biologicznym lub technicznym (MS) mogą wystąpić różne zakłócenia, wpływające na otrzymane widmo. Należy zatem surowe wyniki pomiarów w odpowiedni sposób przetworzyć do dalszych analiz. W pracy posłużono się prostymi metodami, takimi jak m.in. uśrednienie prób z powtórzeń, usunięcie tła, normalizacja danych.

2.1. Uśrednienie wyników z powtórzeń

Pierwszym elementem przygotowania spektrów do dalszych analiz było ustalenie wspólnej osi m/z dla wszystkich widm. Różnice w próbach wynikają z tego, że podczas spektrometrii mas dla każdej próbki była osobno zliczana ich intensywność w detektorze. Wykorzystano tutaj interpolację danych do jednej ustalonej skali m/z.



Rys. 1. Średnie widmo przed testem Dixona oraz po usunięciu widma odstającego
Fig. 1. Average spectra before and after Dixon's test

Następnie przystąpiono do uśrednienia widm na podstawie powtórzeń. Dla każdej badanej osoby zostały przeprowadzone powtórzenia na poziomie izolacji surowicy z krwi, a także wykorzystania spektrometru mas. W ten sposób otrzymano po 4 powtórzenia dla grupy z rakiem oraz od 4 do 12 powtórzeń dla grupy kontrolnej. Przed uśrednieniem, na próbach został wykonany test Dixona wykrywający grube błędy. Test ten bada znaczne różnice między skrajnymi wartościami (najmniejszymi i największymi) a pozostałymi w próbce, przy małej liczbie powtórzeń (od 4 do 25) [6]. W niniejszym artykule badaną próbą była suma intensywności pików przy ustalonej osi m/z, a wielkością próby – liczba powtórzeń dla danego pacjenta. Jako poziom istotności przyjęto wartość 0.01. Wykorzystując tę metodę, odrzucono po jednym powtórzeniu od 5 osób zdrowych (u chorych nie wykryto odstępstw). Po

usunięciu wartości odstających, wyliczono średnie widmo z powtórzeń dla każdej osoby. Na rys. 1 przedstawiono przykładowe średnie widmo przed i po teście Dixona. Można zauważyć, że spektrum przed operacją ma część pików znacznie wyższych niż widmo średnie po usunięciu odstającego wyniku. Jest to spowodowane tym, że spektrum z jednego z powtórzeń zawyżyło wartość średnią. Dodatkowo, dla przyspieszenia późniejszych operacji obliczeniowych, zmniejszono liczbę punktów w widmie do 45 tys. dla każdego pacjenta, ośmiokrotnie zwiększając krok dyskretyzacji.

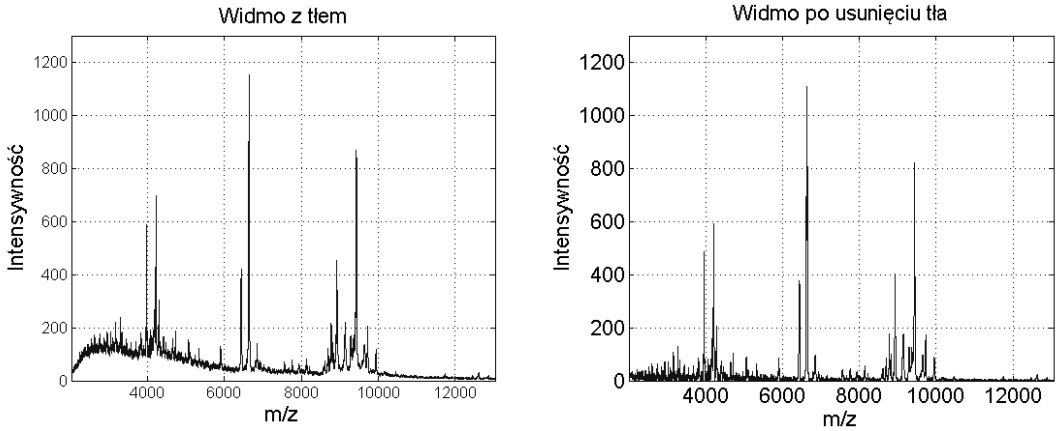
2.2. Usunięcie tła z widma oraz normalizacja

Jednym z artefaktów, który występuje w widmie masowym jest przesunięcie sygnału, określane jako tło. Zależy ono m.in. od przygotowania próbki, opóźnionego uwalniania ładunku, kokrystalizacji cząsteczek białka lub peptydu z cząsteczkami MS matrycy. Wartość poziomu tła może mieć wpływ na zwiększenie obszaru pików, sprawiając trudności w wyborze tych najbardziej znaczących. Celem podstawowych algorytmów jest usunięcie rozległej jego struktury w ten sposób, aby nie naruszyć linii kształtu i szczytów pików w widmie. Dokonuje się to poprzez znalezienie linii bazowej widma, jako niskich częstotliwości szumu w spektrum. Następnie odejmuje się je od pierwotnych pików w spektrum. Na rys. 2 przedstawiono przykładowe widmo przed i po usunięciu tła. Można zauważyć, że w ten sposób został usunięty spory garb na początku osi.

Obecnie istnieje pare algorytmów do wykrywania linii tła w spektrum, np.: średnia ruchoma, funkcja sklejana, transformacja falkowa lub poszukiwanie lokalnych minimów [4, 8, 17, 18]. Jednakże jakość i dokładność detekcji linii bazowej jest trudna do oceny.

W niniejszym artykule posłużono się algorytmem poszukiwania lokalnych percentyli (rząd 0,1), które szacują niskie częstotliwości tła ukrytych wśród wysokich częstotliwości rzeczywistych pików sygnałów. Jest to prosty sposób, a jednocześnie dający bardzo dobry efekt. Na znalezionych punktach w oknach (szerokość równa 200 m/z) dokonano regresji, posługując się interpolacją Hermite'a (ang. *Piecewise Cubic Hermite Interpolation*) [2]. Na koniec odjęto wyznaczone punkty tła od wartości danych pików.

W spektrometrii masowej matryca z badaną próbą może zostać w różnym stopniu zjonizowana. Oznacza to że, mogą występować duże różnice w liczbie jonów w MS, a zatem i w intensywności pików w spektrum. W celu wyeliminowania wspomnianych różnic, po usunięciu tła, sygnał został znormalizowany metodą TIC (ang. *total ion current*). W tej metodzie zakłada się, że suma wszystkich jonów w każdym porównywanym widmie jest taka sama [15, 16].



Rys. 2. Porównanie tego samego widma przed i po usunięciu linii tła
Fig. 2. The graph compares spectra before and after remove baseline

3. Przygotowanie danych do klasyfikacji

Obecnie istnieje bardzo wiele metod opisanych w literaturze, wykorzystywanych do dalszej analizy widm proteomicznych po ich wstępnym przetworzeniu [10, 15, 19]. Większość z nich służy do przygotowania danych do klasyfikacji poprzez zredukowanie wielkości próbki. W tym celu przeprowadza się detekcję pików (poszczególnych szczytów) oraz porównuje je dla różnych spektrów. Zakłada się wtedy, że jeden pik to jedno białko/peptyd, który został zarejestrowany w spektrometrze. Detekcja pojedynczych znaczących pików jednakże jest dość trudna, ponieważ pojawiają się problemy związane z odtworzeniem ich właściwej pozycji czy zagęszczenia (pomiędzy spektrami różnych osób występują przesunięcia).

W niniejszym artykule do opisu pików użyto modelu mieszanin gaussowskich (GMM – ang. *Gaussian Mixture Models*) o różnych parametrach i udziałach w widmie, wykorzystując algorytm EM (ang. *estimation-maximization*) [13, 14]. Rozkład na mieszaninę składowych gaussowskich powtórzono dla różnej liczby komponentów (K), równej 100, 150, … 850, 900.

Ponieważ nie było z góry narzuconej liczby składowych użytych do tworzonego modelu, czynnik ten należało uwzględnić przy ocenie grupowania. Jeśli liczba składowych (K) jest zbyt duża, niektóre elementy są zbędne, gdyż nie dają żadnej informacji. Jednak, jeżeli liczba ta jest zbyt mała, kilka różnych elementów może zostać połączonych w jedną składową. Wyboru liczby komponentów dokonano na podstawie Bayesowskiego Kryterium Informacji (BIC) [14], które porównuje zysk z liczbą składowych, użytych w GMM. Kryterium to dane jest wzorem:

$$L^{BIC} = l - \frac{1}{2}(3K-1)\ln\left(\sum_{n=1}^{N}\sum_{m=1}^{M} y_{m,n}\right), \tag{1}$$

gdzie: l – logarytmiczna funkcja wiarygodności, $y_{m,n}$ – wartość n-tego piku dla m-tej osoby, K – liczba składowych, N – liczba pików po wstępnej obróbce widm, M – liczba widm (licz-

ba badanych osób). Ze wzrostem liczby składowych, wartość L^{BIC} zwiększała się, a następnie ustabilizowała się przy około 300 składowych, stąd tę wartość wybrano do dalszych analiz.

Ustalone składowe Gaussowskie zostały wykorzystane do wyznaczenia wektora cech C dla każdej osoby (spektrum), według równania:

$$C_m(k) = a_k \sum_{n}^{N} y_{m,n} \cdot f(x_n, \mu_k, \sigma_k), \tag{2}$$

gdzie: a_k – waga k-tej składowej, $f(\cdot)$ – funkcja dystrybuanty gaussowskiej dla x_n (n-ta wartość m/z) z wartością oczekiwaną μ i odchyleniem standardowym σ dla k-tej składowej. W ten sposób otrzymano 87 (liczba osób) wektorów wejściowych do klasyfikacji, które zawierały K=300 (liczba składowych gaussowskich) elementów.

4. Klasyfikacja

Do klasyfikacji danych posłużono się maszyną wektorów podpierających (SVM ang. *support vector machines*) [3], gdzie dane dwóch klas są rozdzielane za pomocą hiperpłaszczyzny. Badano tutaj wpływ wyboru jądra oraz miękkiego marginesu na jakość klasyfikacji, przy uwzględnieniu różnej liczby cech.

4.1. Ocena klasyfikatora

Podstawowym wskaźnikiem oceny jakości klasyfikatora jest rzeczywiste prawdopodobieństwo błędnego przyporządkowania do danej grupy. Wartość ta jednak jest nieznana, dlatego należy ją oszacować eksperymentalnie. W przeprowadzonym badaniu wartość tę szacowano metodą oceny błędu, na podstawie wielokrotnego podziału próby (400 razy). W pierwszym etapie dane dzielono losowo na dwie części: próbę uczącą, zawierającą dane od 25 osób zdrowych i 19 chorych oraz próbę walidacyjną, zawierającą dane od 24 osób zdrowych i 19 chorych. Wykorzystując statystyczny test t-Studenta dla dwóch prób [5] wybierano c najlepszych cech (z 300), na podstawie grupy uczącej. Następnie klasyfikator o zadanych parametrach był trenowany na grupie uczącej z wyselekcjonowanymi cechami i testowany na zbiorze walidacyjnym. Otrzymane w ten sposób przyporządkowanie etykiety dla grupy testującej porównywano z prawdziwą przynależnością klasową (obliczano prawdopodobieństwo błędnej klasyfikacji). W ten sposób wyznaczono ogólny błąd przynależności – δ_O, a także błąd dla grupy chorej – δ_{FN} i grupy zdrowej – δ_{FP}. Przyjmując oznaczenia wyników klasyfikatora za pomocą tabeli oznaczeń, używanych w testach diagnostycznych (FP – liczba fałszywych pozytywów, FN – liczba fałszywych negatywów, TP – liczba prawdziwych pozytywów, TN – liczba prawdziwych negatywów), poszczególne błędy wyznaczono następująco:

$$\delta_O = \frac{FP + FN}{TN + TP + FN + FP}, \qquad \delta_{FN} = \frac{FN}{TP + FN}, \qquad \delta_{FP} = \frac{FP}{TN + FP}. \tag{3}$$

Dla każdego, losowego podziału na dwie grupy wyliczano wyżej wymienione błędy. Następnie na podstawie N=400 powtórzeń z losowego podziału na grupy uczące i testujące wyznaczono wartości oczekiwane prawdopodobieństwa złego zaklasyfikowania próby ($\bar{p}$) z 95% przedziałem ufności, według równań:

$$\bar{p} = \frac{\sum_i^N \hat{p}_i}{N}, \qquad \left[\bar{p} - z_{1-\alpha/2} \sqrt{\frac{\sum_{i=1}^N \hat{p}_i (1 - \hat{p}_i)}{nN^2}}, \bar{p} + z_{1-\alpha/2} \sqrt{\frac{\sum_{i=1}^N \hat{p}_i (1 - \hat{p}_i)}{nN^2}} \right], \tag{4}$$

gdzie: $\hat{p}_i$ – estymator dowolnego rodzaju błędu w i-tym losowaniu, n – wielkość grupy walidującej. Wyżej wymienione wzory na przedział ufności są stosowane zgodnie z założeniem, że proces klasyfikacji jest modelowany jako rozkład dwumianowy [5].

5. Badania i wyniki eksperymentalne

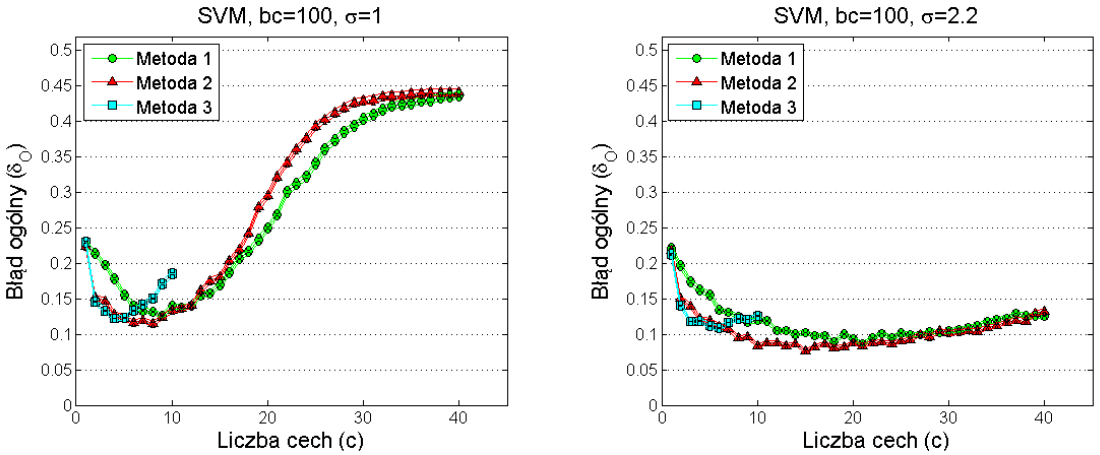
Jak wspomniano na początku rozdziału 4., do wyboru cech posłużono się testem *t*. Dla 300 początkowych cech wyznaczono *p*-wartość (graniczny poziom istotności) dla próby uczącej. Parametr ten ocenia stopień wiarygodności hipotezy, że średnia wartość danej cechy jest różna w grupie z nowotworem i grupie kontrolnej. Indeksy składowych ułożono według rosnącej *p*-wartości w wektorze $F=(f_1, f_2, ..., f_{300})$. W wektorze F' umieszczono c indeksów cech, które zostały wykorzystane w klasyfikacji. Do budowy klasyfikatora nie powinno się używać liczby cech większej niż rozmiar grupy trenującej. Ponieważ w przedstawionych badaniach wielkość próby uczącej wynosiła 44, przyjęto, że maksymalna wartość liczby cech (c) wynosi 40. Wektor F' (w przedstawionym artykule) został wyznaczony na podstawie wektora F, według trzech metod.

Pierwszą metodą wyboru cech do klasyfikatora było przyjęcie c pierwszych składowych wektora F. Zatem wektor wybranych cech wynosił $F'=(f'_1, f'_2, ..., f'_c)$, gdzie $f_j'=f_{j,.}$. Wybierając w ten sposób cechy można byłoby podejrzewać, że składowe, które są ze sobą skorelowane w teście *t* mogą zwracać podobny wynik, czyli nie wnosić dodatkowych informacji do klasyfikatora.

Druga metoda została zaproponowana w celu uniknięcia wyboru skorelowanych cech. Wykonano test na zależność danych, wyznaczając *p*-wartość dla współczynnika korelacji Pearsona [5]. Jako współczynnik α, poniżej którego przyjmowano powiązanie cech, przyjęto wartość 0.01. Początkowo badano tylko skorelowanie dwóch sąsiednich składowych F. Czyli, jako pierwszą cechę przyjmowano zawsze f_1 ($f'_1 = f_1$), a następnie sprawdzano, czy f_2 jest z nią

skorelowana, jeśli nie była, to przyjmowano ją ($f''_2 = f_2$), w przeciwnym razie testowano następną, aż do zaobserwowania braku powiązania ($f''_2 = f_i$). Po wyznaczeniu drugiej cechy do klasyfikatora następne skorelowania sprawdzano tylko z nią, już nie uwzględniając wcześniejszych cech. Podsumowując, poszukując cechę f''_{j+1} sprawdzano tylko, czy jest ona skorelowana z f''_j. W ten sposób wybierano c najlepszych składowych sprawdzając tylko sąsiednie zależności.

Do wyboru cech posłużono się jeszcze dodatkową, trzecią metodą. Polegała ona na sprawdzaniu wszystkich powiązań (f''_j nie może być skorelowana z f''_i, jeżeli i≠j). Na rys. 3 przedstawiono porównanie tych 3 sposobów wyboru cech dla błędu ogólnego. Zastosowany został klasyfikator SVM z radialnym jądrem dla dwóch różnych σ, przy parametrze ograniczenia pola miękkiego marginesu bc =100. Dla analizowanych danych maksymalna liczba wszystkich cech nieskorelowanych ze sobą była równa 10. Dlatego wykres dla trzeciej metody ma taką właśnie graniczną wartość na osi x. Z wykresu można zauważyć, że niezależnie od parametrów klasyfikatora najlepsze wyniki otrzymano przy sprawdzaniu sąsiednich korelacji (metoda 2) – minimalny błąd jest niższy oraz osiągany przy mniejszej liczbie cech w porównaniu, z pozostałymi metodami. Przy sprawdzaniu wszystkich zależności składowych istnieje ryzyko usunięcia jednak istotnych cech, zatem i minimalny błąd jest wyższy niż przy drugiej metodzie.



Rys. 3. Porównanie 3 sposobów wyboru cech do klasyfikacji dla jądra radialnego z σ=1 (a) oraz σ=2.2 (b)

Fig. 3. The graph compares 3 methods to select features for SVM with radial kernel (a: σ=1, b: σ=2.2)

Tabela 1 zawiera indeksy10 najczęściej wybieranych cech, na podstawie testu t, po 400 losowaniach. Wyodrębniono osobno najczęściej wybierane cechy, jako pierwsza, druga i trzecia z kolei (f''_1, f''_2, f''_3). Tabela ta zawiera porównanie częstości wyboru danej cechy, gdy nie badano korelacji (a) oraz gdy była ona sprawdzana dla 2 sąsiednich cech (b). Zgodnie z oczekiwaniami dla f''_1 otrzymano tę samą kolejność najczęściej wybieranych cech w obu przypadkach. Różnice pojawiają się jednak już przy cechach drugiej i trzeciej. Wprowadzenie

badania korelacji powoduje, że do klasyfikatora nie wprowadza się podobnych składowych, dzięki czemu przy mniejszej ich liczbie można osiągnąć minimalny błąd (rys. 3).

Tabela 1

Porównanie częstości wyboru cech do klasyfikatora bez oraz z sprawdzeniem korelacji sąsiednich cech przy teście *t*

a)

Dane bez badania korelacji					
Cecha 1 (f'_1)		Cecha 2 (f'_2)		Cecha 3 (f'_3)	
Nr skł.	Udział [%]	Nr skł.	Udział [%]	Nr skł.	Udział [%]
12	39,18	215	26,07	109	18,81
215	28,47	109	19,49	215	17,64
173	11,02	12	17,73	12	17,35
109	8,48	173	16,35	173	13,06
73	6,43	73	4,66	266	5,07
266	1,46	66	4,20	73	4,01
66	1,13	266	3,29	235	3,45
68	0,93	68	1,25	66	3,15
279	0,50	246	1,06	68	2,26
192	0,41	192	0,70	227	1,84

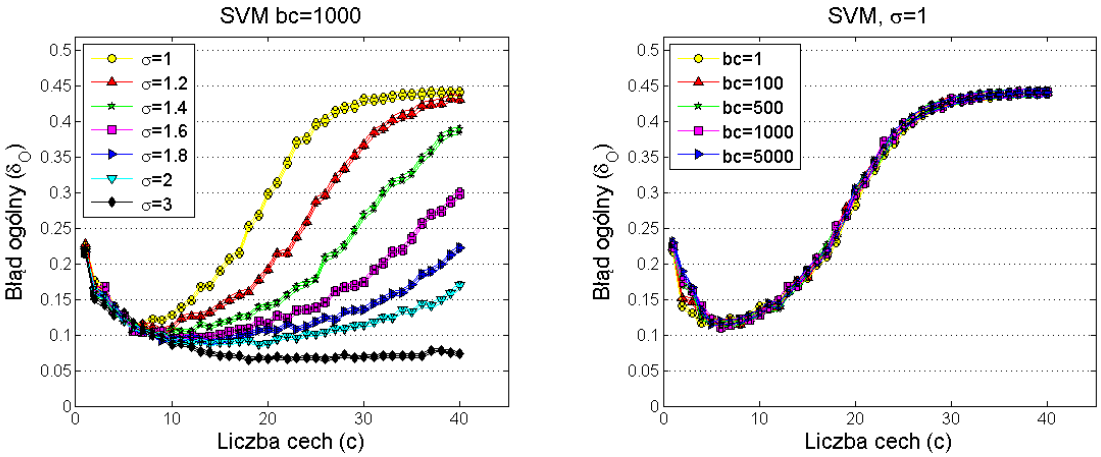
b)

Dane przy badaniu korelacji dwóch sąsiednich składowych					
Cecha 1 (f'_1)		Cecha 2 (f'_2)		Cecha 3 (f'_3)	
Nr skł.	Udział [%]	Nr skł.	Udział [%]	Nr skł.	Udział [%]
12	38,96	68	20,52	109	7,26
215	28,74	73	14,34	173	5,21
173	10,99	246	7,52	266	5,14
109	8,59	237	7,34	235	4,30
73	6,42	192	7,11	239	4,26
266	1,48	279	7,07	12	4,22
66	1,09	263	6,86	286	4,12
68	0,95	215	4,13	215	4,07
279	0,46	12	3,63	53	3,12
192	0,35	139	3,27	95	2,90

W przeprowadzonych badaniach sprawdzano nie tylko wpływ liczby cech wykorzystywanych w klasyfikatorze, ale również jak wybór parametru *bc* (służącego do określenia ograniczenia pola miękkiego marginesu) oraz jak wybór jądra SVM wpływają na jakość klasyfikacji. Na rys. 4 oraz w tabeli 2 przedstawiono wpływ tych parametrów na poprawność zaklasyfikowania do właściwej grupy przez SVM. Można zauważyć, że parametr określający właściwości miękkiego marginesu nie wpływa znacząco na wynik końcowy, różnice jednak widać przy różnych parametrach σ dla radialnego jądra. Wraz ze wzrostem jego wartości spada minimalny błąd klasyfikacji, ale jednocześnie to minimum pojawia się przy większej liczbie cech, branych do klasyfikatora. Zatem, im wyższa jest wartość σ, tym przy większej liczbie cech błąd dąży do stabilności równej losowemu przypisaniu etykiety, czyli do 19/43= 0,4419.

Na podstawie przedstawionych wyników mogłoby się wydawać, że najlepszym rozwiązaniem byłoby użycie klasyfikatora z jądrem radialnym o dużej wartości σ, przy dużej liczbie cech. Trzeba jednak wziąć pod uwagę, że uczenie na zbiorze o dużej liczbie parametrów może powodować, że otrzyma się dobry klasyfikator dla danej, pobranej próby. Jednak byłby on mało odporny na małe zmiany; w omawianym przypadku np. na pobieranie krwi przez inną osobę czy wykonywanie MS na innym sprzęcie. Można przyjąć, że dobrą klasyfikację otrzymuje się przy parametrze σ równym od 1 do 1,4, przy 6-8 cechach. Jak wspomniano wcześniej wartość ograniczenia na miękki margines (*bc*) nie ma znaczącego wpływu na wyniki

końcowe, warto jednak wybierać wyższe wartości, w celu uniknięcia zbyt dokładnego dopasowania się do danych.



Rys. 4. Wpływ parametrów klasyfikatora na błąd klasyfikacji

Fig. 4. Influence of classifier's parameters on the classification error

Tabela 2

Minimalny błąd klasyfikatora wraz z liczbą cech dla zadanych parametrów klasyfikatora

Minimalny błąd											
bc \ σ	1	1,2	1,4	1,6	1,8	2	2,2	2,4	2,6	2,8	3
1	0,116	0,106	0,103	0,101	0,094	0,086	0,085	0,080	0,074	0,069	0,071
100	0,115	0,107	0,102	0,098	0,089	0,085	0,076	0,076	0,070	0,063	0,065
500	0,116	0,106	0,102	0,097	0,090	0,084	0,081	0,074	0,070	0,067	0,061
1000	0,111	0,107	0,100	0,095	0,090	0,087	0,081	0,073	0,071	0,067	0,066
2000	0,117	0,108	0,098	0,092	0,088	0,084	0,081	0,075	0,070	0,067	0,064
5000	0,115	0,106	0,103	0,095	0,091	0,083	0,078	0,075	0,074	0,066	0,064
Liczba cech dla minimalnego błędu											
bc \ σ	1	1,2	1,4	1,6	1,8	2	2,2	2,4	2,6	2,8	3
1	6	6	7	4	9	11	15	10	18	18	16
100	8	6	7	10	12	11	15	21	21	24	25
500	6	8	10	10	11	18	19	23	17	24	29
1000	6	9	9	10	14	19	13	22	21	19	18
2000	8	7	8	9	13	12	15	22	19	28	23
5000	5	7	8	11	15	14	15	19	20	20	24

Tabela 3

Końcowa tabela błędnej klasyfikacji

c	bc=5000								
	σ=1			σ=1,2			σ=1,4		
	δ_O	δ_{FP}	δ_{FN}	δ_O	δ_{FP}	δ_{FN}	δ_O	δ_{FP}	δ_{FN}
6	0,119	0,105	0,138	0,113	0,101	0,129	0,109	0,089	0,134
7	0,118	0,099	0,142	0,106	0,091	0,124	0,109	0,090	0,132
8	0,121	0,101	0,145	0,110	0,091	0,134	0,103	0,087	0,122

Wyniki końcowe przedstawiono w tabeli 3 dla trzech różnych wartości parametru σ oraz wykorzystaniu w SVM 6, 7 i 8 cech wybieranych testem *t* z badaniem sąsiednich korelacji.

Oprócz analizy poprawności ogólnej klasyfikacji osób do poszczególnej grupy (zdrowy, chory), warto również zwrócić uwagę na wielkości błędów cząstkowych (δ_{FN} oraz δ_{FP}). Podczas testów diagnostycznych bardzo ważne jest, aby minimalizować błąd δ_{FN} polegający na kwalifikacji osoby chorej jako zdrowej.

5.1. Znajdowanie cech różnicujących

W celu znalezienia najlepszych, różnicujących składowych gaussowskich można byłoby przyjąć r najczęściej wybieranych indeksów cech występujących na pierwszej pozycji w wektorze cech użytych do klasyfikacji (f'_1) po 400 powtórzeniach – metoda 1. Innym sposobem wyboru najlepszych składowych jest wyliczenie sumy wszystkich częstości występowania danej cechy w wektorze F' do r-tego miejsca ($f'_1,...,f'_r$) – metoda 2 – lub sumowanie tych częstości występowania cech z uwzględnieniem wag, w zależności od kolejności wyboru na podstawie testu t (metoda 3). W przedstawionym badaniu wzór na częstość z wagą wygląda następująco:

$$W(k) = \sum_{i}^{r} \omega^{r-i} S(k,i), \tag{5}$$

gdzie: k jest indeksem składowej od 1 do 300, $S(k, i)$ jest macierzą częstości występowania po 400 powtórzeniach cechy j w wektorze F' na i-tej pozycji, W jest współczynnikiem, na podstawie którego zostały potem wyznaczone najczęściej wybierane cechy. W tabeli 4 pokazano jak – w zależności od przyjętego kryterium – kolejność najczęściej wybieranych cech może się różnić. Można jednak zauważyć, że wybierając zbiór 6 najczęściej wybieranych cech (oprócz metody 1) otrzymujemy taki sam zestaw – {12, 68, 73, 109, 173, 215}.

Tabela 4

Kolejność najczęściej wybieranych cech według W wyznaczonego 7 przykładowymi sposobami

	Metoda1	Metoda 2		Metoda 3			
	Na podstawie f'_1	Na podstawie f'_1:f'_3	Na podstawie f'_1:f'_6	r=3, ω=10	r=3, ω=3	r=6, ω=10	r=10, ω=10
1	12	12	12	12	12	12	12
2	215	215	68	215	215	215	215
3	173	68	215	173	73	73	68
4	109	73	109	109	173	68	73
5	73	173	173	73	68	173	173
6	266	109	73	68	109	109	109
7	66	237	279	266	246	279	279
8	68	192	192	66	237	192	192
9	279	246	237	279	279	237	237
10	192	279	227	246	192	246	246

Każda wybrana składowa reprezentuje określony zbiór pików w widmie. Na podstawie tej informacji jest możliwe sprawdzenie, jaki związek chemiczny jest reprezentowany przez te piki. Przykładem bazy danych, umożliwiającej wyszukiwanie białek na podstawie tego typu informacji jest Empirical Proteomics Ontology Knowledge Base [9] – dostępna na stronie http://www2.dbmi.pitt.edu/EPO-KB.

6. Podsumowanie przeprowadzonych badań

W przedstawionym artykule pokazano jeden ze sposobów analizy danych proteomicznych po spektrometrii masowej. Algorytmy wstępnej obróbki widm, a także przygotowanie cech wraz z ich selekcją pozwoliły na uzyskanie zadowalających wyników klasyfikacji. W jej trakcie wykonywano 400-krotny losowy podział na grupy uczącą oraz testującą, a na koniec estymowano prawdopodobieństwo błędnego przypisania do klasy na podstawie średniej z powtórzeń. Sposób ten z jednej strony pomaga w oszacowaniu rzeczywistego prawdopodobieństwa błędu, jednak z drugiej strony utrudnia wyciągnięcie z danych cech, które najbardziej różnicują. W artykule pokazano przykładowe sposoby, wraz z wynikami, wyznaczenia zestawu najlepszych cech do klasyfikatora.

Skupiając się na rozwinięciu przedstawionych algorytmów, można byłoby również przeanalizować i porównać między sobą proteomy osób cierpiących na inne nowotwory, a to mogłoby odpowiedzieć na pytania, jakie są wspólne cechy charakteryzujące te nowotwory i co rozróżnia poszczególne ich typy? Otwiera to zatem dalszą drogę do wzrostu znaczenia proteomiki w diagnostyce chorób nowotworowych.

Podziękowania

Dziękujemy prof. Piotrowi Widłakowi oraz dr Monice Pietrowskiej z Zakładu Radiobiologii Doświadczalnej i Klinicznej z Centrum Onkologii w Gliwicach za udostępnienie danych.

BIBLIOGRAFIA

1. Baggerly K. A., Morris J. S., Coombes K. R.: Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. Bioinformatics 20(5), 2004, s. 777÷758.
2. Catmull E., Rom R.: A class of local interpolating splines. [in:] Barnhill R. E., Reisenfeld R. F., (eds.): Computer Aided Geometric Design, Academic Press, New York 1974, s. 317÷326.

3. Cortes C., Vapnik V.: Support vector networks. Machine Learning 20, 1995, s. 1÷25.
4. Ferguson J.C.: Multi-variable curve interpolation. J. ACM 11(2), 1964, s. 221÷228.
5. Grzegorzewski P., Bobecka K., Dembińska A., Pusz J.: Rachunek prawdopodobieństwa i statystyka, Wyd. 4, WSISiZ, Warszawa 2003.
6. Kanji G. K.: 100 Statistical Tests. 3 edn. SAGE Publications Ltd, 2006.
7. Koronacki J., Ćwik J.: Statystyczne systemy uczące się. WNT, Warszawa 2005.
8. Liu Q., Krishnapuram B., Pratapa P., Liao X., Hartemink E., Carin L.: Identication of differentially expressed proteins using maldi-tof mass spectra. In: Asilomar Conference: Biological Aspects of Signal Processing, 2003.
9. Lustgarten. J. L., Kimmel Ch., Ryberg H., Hogan W.: EPO-KB: a searchable knowledge base of biomarker to protein links. Bioinformatics 24 (11), 2008, s. 1418÷1419.
10. Petricoin E. F., Ardekani A. M., Hitt B. A., Levine P. J., Fusaro V. A., Steinberg S. M., Mills G. B., Simone C., Fishman D. A, Kohn E. C., Liotta L. A.: Use of proteomic patterns in serum to identify ovarian cancer. The Lancet 359, 2002, s. 527÷577.
11. Pietrowska M., Marczak L., Suwinski R., Stobiecki M., Polanska J., Polanski A., Widlak P., Gawkowska-Suwinska M., Drosik A., Walaszczyk A.: Application of mass spectrometry-based serum proteome pattern analysis in identiffication of lung cancer patients. J Thorac Oncol 5(5, Suppl 1), S60, 2010. Abstract book, 2nd European Lung Cancer Conference, Geneva, Switzerland, 28 April-1 May 2010.
12. Pietrowska M., Marczak Ł., Widłak P.: Proteomika kliniczna – wykorzystywanie metod spektrometrii mas do analizy proteomu surowicy krwi w diagnostyce chorób nowotworowych. In: Na pograniczu chemii i biologii, t. XVII, 2007.
13. Polanska J., Widlak P., Rzeszowska-Wolny J., Kimmel M., Polanski A.: Gaussian Mixture Decomposition of Time-Course DNA Microarray Data. In: Mathematical Modeling of Biological Systems, vol. I, Modeling and Simulation in Science, Engineering and Technology. Springer, 2007, s. 351÷359.
14. Polanski A., Kimmel M.: Bioinformatics. Springer, 2007.
15. Morris J. S., Coombes K. R., Koomen J., Baggerly K. A., Kobayashi R.: Features extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. Bioinformatics 21(9), 2005, s. 1764÷1775.
16. Na S., Paek E.: Quality assessment of tandem mass spectra based on cumulative intensit normalization. J Proteome Res. 5(12), 2006, s. 3241÷3248.
17. Shin H., Sampat M. P., Koomen J. M., Markey M. K.: Wavelet-Based Adaptive Denoising and Baseline Correction for MALDI TOF MS. OMICS 14(3), 2010, s. 283÷295.
18. Wagner M., Naik D., Pothen A.: Protocols for disease classiffication from mass spectrometry data. Proteomics 3(9), 2003, s. 1692÷1698.

19. Yu W., He Z., Liu J., Zhao H.: Improving Mass Spectrometry Peak Detection Using Multiple Peak Alignment Results. Journal of proteome research 7(01), 2008, s. 123÷129.

Recenzent: Prof. dr hab. inż. Andrzej Świerniak

Wpłynęło do Redakcji 31 stycznia 2011 r.

Abstract

Mass spectra analysis is an important area in clinical proteomics. It has a great potential to contribute in terms of biomarker or therapeutic target discovery.

This paper shows one of the methods for analyzing mass spectra. The main aim of this study is to classify two groups of people – the group with lung cancer and the control group (healthy). The data were delivered by the Department of Experimental and Clinical Radiobiology, Cancer Center and Institute of Oncology in Gliwice. Serum was isolated from blood taken from 49 healthy people and 38 people with cancer. Next, they were used in mass spectrometry.

The first step, before classification, was to preprocess the data. This procedure included standardization of m/z axis, removing outliers using Dixon's test and determining average spectrum of repetitions for each person. Next, the baseline of the signal was corrected. After that, the TIC method was used in order to normalize data. At the following step, Gaussian Mixture Models were used for representing peaks in the spectra. Finally, using Bayesian information criterion, 300 Gaussians were chosen for the model.

Support vector machine (SVM) was used in the work reported here. The data were split into learning and validation sets. Features used for the classification were selected based on *t*-test performed for the learning set. By using *t*-test only on teaching group, there was chosen features to classifier. Next, the classification error was measured for the validation set. This process was repeated 400 times and, at the end, average error was computed with 95% confidence interval (CI). In this process of classification it was investigated how the SVM's parameters, number of features (which were used for the classification) influenced the validation scores. Moreover it was checked whether the classification quality would have been improved if the correlated features had been eliminated. Also, it was investigated which features were most frequently chosen. The final results show that the overall error (for 6 – 8 features in SVM) equals about 10%, which is encouraging for future works in this area.

Adresy

Jolanta KAWULOK: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16, 44-101 Gliwice, Polska, jolanta.kawulok@polsl.pl.

Joanna POLAŃSKA: Politechnika Śląska, Instytut Automatyki, ul. Akademicka 16, 44-101 Gliwice, Polska, joanna.polanska@polsl.pl.