Silesian University of Technology
Faculty of Automatic Control, Electronics and Computer Science
Institute of Informatics

Doctor of Biomedical Engineering Dissertation

Machine learning-based workflow for the analysis of MALDI-TOF mass spectrometry cancer data

mgr inż. Wojciech Sikora

Supervisor: Prof dr hab. inż. Joanna Polańska

mgr inż. Wojciech Sikora

Zastosowanie technik uczenia maszynowego do kompleksowej analizy danych obrazowania molekularnego MALDI-TOF w badaniach nad rakiem

Streszczenie

Przedmiotem pracy doktorskiej jest analiza danych otrzymanych za pomocą obrazowania spektrometrią mas próbek pobranych od pacjentów z nowotworem głowy i szyi. W ramach pracy postawiono następujące hipotezy. Pierwsza hipoteza twierdzi, że identyfikacja pików w spektrach masowych może być skutecznie przeprowadzona za pomocą modelowania całego spektrum, poprzez podzielenie go na części oraz zamodelowaniu ich mieszaninami normalnymi. Druga hipoteza twierdzi, że informacja o przestrzennej dystrybucji danych pozyskana z obrazowania spektrometrią mass skutecznie usuwa redundancje i znacznie zmniejsza wymiarowość danych, przy jednoczesnym zachowaniu jakości danych. Ostatnia hipoteza twierdzi, że identyfikacja najważniejszych cech, dla danych heterogenicznych jest możliwa i skuteczna dzięki wnioskowaniu na podstawie wielu modeli jednostkowych.

Pierwsze rozdziały skupiają się na podstawowych zagadnieniach związanych z proteomiką i spektrometrią mas. W pierwszej kolejności przedstawiono ogólny opis spektrometrii mas oraz obrazowania tkanek za pomocą tej techniki. Opisane są najważniejsze metody jonizacji oraz analizatory mas, które są powszechnie wykorzystywane do analizy próbek pochodzenia biologicznego, w szczególności próbek pochodzących od pacjentów z nowotworem. Następnie po krótce opisano proces przygotowania próbek oraz pozyskiwania danych, ich charakterystykę, a także pierwsze działania mające na celu przygotowanie danych do dalszej analizy, tj. zastosowane metody korekty linii bazowej oraz normalizacji.

Następny rozdział porusza temat agregacji spektrów masowych oraz przedstawia aktualny wiedzy na temat detekcji pików. Na zagregowanych danych przeprowadzono detekcję pików przy pomocy najczęściej wykorzystywanych w tym celu algorytmów.

W pierwszej kolejności piki zostały zidentyfikowane za pomocą najbardziej podstawowej i popularnej metody bazującej na określeniu współczynniku sygnału do szumu, wykorzystując intensywność pików jako sygnał. Następną wypróbowaną metodą jest, bazująca na transformacie falkowej, metoda modelowania pików.

W następnym rozdziale przedstawiona została metoda identyfikacji pików polegająca na modelowaniu całego spektrum masowego za pomocą modelu mieszanin rozkładów normalnych. Na początku opisana została metoda dzielenia spektrum masowego na mniejsze fragmenty w sposób odmienny od oryginalnie proponowanej. Przedstawiono szczegółowy schemat działania, wraz z pseudokodem oraz porównaniem wyników z oryginalną metodą.

Kolejna część pracy porusza temat dopasowywania mieszanin rozkładów normalnych do podzielonego spektrum. Zawiera ona matematyczny opis dopasowywania mieszanin z wykorzystaniem własnej implementacji algorytmu expectation-maximization (EM). W szczególności opisano metodę wybierania optymalnej liczby elementów w mieszaninach oraz wpływu losowej natury algorytmu EM na wyniki modelowania spektrum. Na końcu rozdziału przedstawiono wyniki procesu identyfikacji pików. Wyniki potwierdzają prawdziwość pierwszej postawionej hipotezy.

Następny rozdział skupia się na wykorzystywaniu metod statystycznych oraz przestrzennej dystrybucji cech w celu usunięcia redundancji i redukcji wymiarowości danych. W tym celu przeprowadzona została filtracja szumów wykorzystując parametry dystrybucji normalnych opisujących elementy modelu spektrum. Dalej inżynieria cech jest kontynuowana z wykorzystaniem informacji, których dostarcza obrazowanie. Przestrzenna dystrybucja pobliskich cech jest wykorzystana do zmniejszenia liczby cech. Porównanie jest wykonywane z pomocą testu statystycznego na podobieństwo dystrybucji Peacock'a. Jest to rozszerzenie testu Kołmogorov'a-Smirnov'a do dwóch wymiarów. Po wyznaczeniu eksperymentalnie wartości krytycznych, pobliskie cechy o statystycznie identycznej dystrybucji przestrzennej są łączone. Proces redukcji wymiarowości kończy się na detekcji obwiedni izotopowych, które są redukowane do pojedynczej cechy. Obwiednie izotopowe są wykrywane na podstawie odległości między pikami, kształtu pików oraz dystrybucji przestrzennej. Wyniki pokazują znaczne zmniejszenie wymiarowości danych, potwierdzając drugą postawioną hipotezę.

W kolejnych rozdziałach opisano proces uczenia klasyfikatorów na przetworzonych danych. Nauczono dwie grupy klasyfikatorów. Pierwsza grupa to klasyfikatory trenowane z wykorzystaniem wielomianowej regresji logistycznej, gdzie klasyfikator jest trenowany przez iteracyjne wykonywanie regresji logistycznej,

wybierając za każdym razem najistotniejszą cechę ze zbioru i dodając ją do listy predyktorów końcowego modelu. Druga grupa klasyfikatorów to proste w pełni połączone sieci neuronowe z dwoma ukrytymi warstwami, każda z liczbą węzłów równą liczbie cech. Klasyfikatory zostały ocenione między innymi za pomocą miar obliczanych na podstawie macierzy błędów takich jak dokładność, precyzja, czułość, swoistość, wartość predykcyjna ujemna, miara F1, a także krzywych ROC oraz krzywych dokładność-czułość. Ostatnia część eksperymentów bada słuszność trzeciej postawionej tezy. Jest ona poświęcona badaniom ogólnej ważności cech obu modeli, wykorzystując ważności cech w modeli jednostkowych. W szczególności, dla sieci neuronowych do określenia ważności cech wykorzystano takie metody jak LIME oraz wartości Shapley'a.

Ostatni rozdział pracy to dyskusja na temat przeprowadzonych badań, wniosków jakie zostały wyciągnięte podczas ich przeprowadzania, ich wyników oraz planów na dalsze badania.