Politechnika Śląska

Wydział Automatyki, Elektroniki i Informatyki

Dimensionality reduction and feature selection methods in the problems of grouping and classification of molecular imaging data.

Doctoral Dissertation

by

Katarzyna Frątczak

Supervisors

prof. dr hab. inż. Joanna Polańska

dr hab. Monika Pietrowska, prof. NIO

Gliwice 2022

Katarzyna Frątczak: "Dimensionality reduction and feature selection methods in the problems of grouping and classification of molecular imaging data."

Streszczenie

Wobec stale wzrastającej liczby zachorowalności na różne rodzaje nowotworów, uwaga naukowców skupia się na poznaniu przyczyn zjawiska nowotworzenia. Obrazowanie molekularne techniką Spektrometrii Mas jest obecnie jednym z podstawowych narzędzi analitycznych umożliwiających analizę składu molekularnego różnych tkanek, zarówno zdrowych, jak nowotworowych. Należy zauważyć, że duży rozmiar danych i wysoka złożoność sygnału stanowią wyzwania w analizie statystycznej i uczeniu maszynowym. Zwłaszcza, że w standardowym podejściu do klasyfikacji każda klasa jest z reguły wysoce jednorodna. Obecnie wiadomo jednak, że zarówno nowotwór, jak i zdrowa tkanka to zbiór komórek o dużej heterogeniczności, co stanowi wyzwanie w zbudowaniu wiarygodnego systemu klasyfikacji. Niestety komercyjne narzędzia nie umożliwiają wielopoziomowego przetwarzania i analizy wielkoskalowych danych MSI. Dlatego głównym celem rozprawy było opracowanie innowacyjnych algorytmów i narzędzi umożliwiających analizę statystyczną i klasyfikację wysoce niejednorodnych, wielowymiarowych danych obrazowania molekularnego.

Zaproponowany został schemat przetwarzania danych obejmujący adaptacyjne przetwarzanie wstępne, ekstrakcję i selekcję cech do stworzenia systemu decyzyjnego rozróżniającego tkankę zdrową od nowotworowej. Zastosowane metody przetestowane zostały na danych z kilku eksperymentów MALDI MSI, w których obrazowane były różne rodzaje tkanek. W badaniu skoncentrowano się również na problemie modelowania bioróżnorodności wewnątrztkankowej i międzyosobniczej. Zaproponowano adaptacyjną korekcję linii bazowej i ekstrakcję cech przy użyciu modelu mieszanin Gaussa, aby zmniejszyć wymiarowość danych. Wprowadzono inteligentny system konstruowania zbioru uczącego z wykorzystaniem informacji o niejednorodności podtypów tkanek, co przełożyło się na poprawę jakości klasyfikacji. Z kolei zaproponowana metoda selekcji cech z wykorzystaniem adaptacyjnego rankingu, pozwala zamodelować heterogeniczność zbioru danych i znaleźć wiarygodną ostateczną sygnaturę raka.

Podsumowując, niniejsza rozprawa proponuje rozwiązania umożliwiające szybkie przetwarzanie i wydajną analizę danych MSI, pozwalając tym samym na lepsze zrozumienie biologicznych podstaw rozwoju nowotworów.