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."

Abstract

As global cancer incidence rates continue to increase, scientists focus on finding the causes of carcinogenesis. Molecular imaging of tissues with the Mass Spectrometry technique iscurrently one of the basic analytical tools that enable the analysis of the molecular composition of various tissues, both healthy and diseased. It should be noted that the large data size and the high signal complexity present challenges in statistical analysis and machine learning. Especially in the standard classification approach, each class is highly homogeneous. However, it is now known that both cancer and healthy tissue are a collection of very heterogeneous cells, which is a challenge in building a reliable classification system. Unfortunately, no commercial tools currently enable multi-level processing and analysis of large-scale MSI data. Therefore, it is necessary to create innovative methods and algorithms to extract relevant information from biological data. The main aim of the dissertation was to develop advanced algorithms and tools enabling statistical analysis and classification of highly heterogeneous multivariate cancer imaging data.

A data processing scheme was proposed that includes adaptive pre-processing, extraction and feature selection to create a decision-making system that differentiates between healthy and neoplastic tissue. The analysis was carried out on the MALDI-MSI data, which measures the spectra of many cancers and healthy tissues. The study also focused on the problem of modelling inter-tissue and intra-individual biodiversity. Adaptive baseline correction and feature extraction using the Gaussian mixture model has been proposed to reduce the dimensionality of the data. The introduction of an intelligent system for constructing the training set using information about the heterogeneity of tissue subtypes improved the classification quality. In turn, the proposed trait selection method using adaptive scoring can cover the heterogeneity of the cancer dataset and find a reliable final cancer signature.

In conclusion, this dissertation proposes solutions enabling the rapid processing and efficient analysis of MSI data, allowing a better understanding of the biological basis of cancer development.