POLITECHNIKA ŚLĄSKA

KATEDRA INŻYNIERII I BIOLOGII SYSTEMÓW

Rozprawa doktorska

Opracowanie nowych algorytmów uczenia maszynowego dla heterogenicznych danych biomedycznych

Autor: mgr inż. Agata Wilk

Promotor: prof. dr hab. inż. Krzysztof Fujarewicz

Gliwice, czerwiec 2024

Summary

Cultures of cells desynchronised in cell cycle phase, tumour tissues composed of cells of different types and molecular profiles, cohorts of patients differing in disease status, genetic and environmental factors, diverse populations of regions or countries. Clinical, genomic, transcriptomic, proteomic and imaging data, from different sources, with different types, structure and dimensionality. Heterogeneity is an inherent aspect of biomedical research, the cause of many biological phenomena and the basis of personalised therapy. However, the analysis of heterogeneous data presents a number of challenges — from poor numerical conditioning and difficult estimation of model parameters, to the presence of different numbers of feature vectors for individual objects.

The key research problem addressed in this dissertation is the use of data representing heterogeneous objects in machine learning. The dissertation consists of a series of seven research articles, illustrating the author's own experience.

The first part addresses heterogeneity occurring at different levels. In particular, the dilemma between building a common model for heterogeneous groups or independent analysis of subgroups or individual objects is discussed. As a compromise combining these strategies, an original individualised approach to parameter estimation (using an epidemiological model as an example) is proposed, which involves estimating some parameters as common to all objects and some as independent. The use of individualised modelling helped overcome the problem of poor numerical conditioning, while allowing elements of individual characteristics to be retained, as reflected in the distribution of prediction errors.

The second part of the dissertation describes the problem of heterogeneity of the data structure, i.e. the presence of different numbers of feature vectors. Original strategies for the use of such data in classification (for proteomic imaging) and survival modelling (for

multiple uptakes in PET/CT imaging), based on aggregation of feature vectors or aggregation of modelling results are presented. The obtained results show that the use of information from all available feature vectors through aggregation improves the predictive ability of models relative to the use of a single feature vector.

Keywords: heterogeneous data, machine learning, classification, individualized model, survival analysis, aggregation