Politechnika Śląska
Katedra Inżynierii i Biologii Systemów

Alicja Płuciennik

Algorytmy integracji danych i uczenia maszynowego w diagnostyce nowotworów

Promotor: prof. dr hab. inż. Krzysztof Fujarewicz
Promotor pomocniczy: dr hab. Małgorzata Oczko–Wojciechowska
Opiekun przemysłowy: dr inż. Michał Bachorz

Gliwice 2021

Abstract of Doctoral Dissertation

Alicja Płuciennik, MSc, BEng

Data Integration and Machine Learning Algorithms in Cancer Diagnostics

Cancer diagnostics is one of the key issues in contemporary medicine, as cancer is affecting a growing number of people. Currently, aside from the detection of cancer itself, it is also crucial to assess its nature, which may have an effect on decisions regarding treatment. The popular direction in the development of oncological diagnostics is molecular testing. However, its availability in routine diagnostics is insufficient, the common reason being the high cost of a test employing molecular multi-gene classifiers.

This work aims to present examples and strategies of integration algorithms application in the development of diagnostic tools employing molecular data and other clinical factors used to reduce molecular features and decrease the cost of a test.

The application of data fusion as a combination of feature sets for each classified observation is extensively analyzed in the work. The work encompasses the study of the effects of an application of two data fusion strategies on the quality of the classifier, recognized in the context of feature selection as *early data fusion* and *late data fusion* – expressed as classification accuracy, which is the percentage of correctly classified samples. Due to the introductory nature of the work, studies were conducted on the real data sets from the collected endocrine tumor samples. Data from the TCGA database were used for luminal breast cancer subtypes. Due to the study conditions resulting from the relatively small sets, the analyses were carried out using a bootstrap method. Additionally, the effects of the data fusion were studied in the context of either application of feature extraction methods for clinical data or with no extraction. Dependencies between features that underwent data fusion were thoroughly studied by using markers such as Spearman's correlation and mutual information. Feature selection stability analysis utilizing Kuncheva's index was suggested as an option for comparing the impact of fusions on feature selection.

Results of the conducted studies indicate that the use of data fusion allows for a reduction in the number of molecular features. However, the results are linked to the occurrence of feature dependencies. Statistically significant differences were demonstrated (p-value < 0.05) in the favor of data fusion use in the case of endocrine tumors. As a conclusion of the conducted analyses, one may state that features demonstrating high quality of classification on their own may be used in late data fusion, and feature dependencies are present in the primary set. However, if one is unsure of the sufficient strength of dependence and sufficient classification quality for the introduced features, it is beneficial to use early fusion for the algorithm of feature selection, allowing it to pick the best variables from the subset. The presented results of feature dependencies analysis show that in order for the fusion to be successful, the dependencies do not need to be strong – an average correlation and mutual information lower than 0.4 are sufficient.

Due to the introductory nature of the work, the exemplary results of the industrial research are shown that illustrate an application of the presented data fusion methods for data sets of endocrine neoplasms and breast cancers.

Methods, algorithms, and results shown in this work present a possible application of data fusion and may be used in the future in a broader context, utilizing other subsets of features for the creation of new, hybrid diagnostic methods.