Clinical Interpretation of Omics Clustering Results

Omics data e.g. gene expression data provide a molecular characterization of diseases. However, the resulting clusters building disease subgroups often lack interpretability due to missing clinical data annotations. These annotations would help clinicians, researchers and doctors tremendously to identify the causes and potential patterns for specific diseases.
We focus on a dataset of 29 real-world patients with heart failure consisting of omics and clinical data. The clinical data incorporates heterogeneous data points like categorical, numerical and Boolean values including missing data points. It has 33 common categories out of 199 total unique ones. Two molecular subgroups were provided based on an omics data clustering result.
We present a comparison of different feature selection methodologies identifying relevant parameters in the clinical data of heart failure patients based on the provided omics data clustering result. Therefore, by annotating the provided omics clustering result with identified relevant clinical parameters, we enable clinicians to interpret their omics data.

Feature selection models can be divided into three main methods: filter, wrapper and embedded methods [1].
Filter methods select features based on criteria independent of any supervised learning algorithm and their performance may not be optimal. Popular filter methods are Pearson's Correlation and Analysis of Variance (ANOVA) [2].
Wrapper methods use a predictive model to evaluate the relative usefulness of parameter subsets. For that purpose, one needs to define a) how to search the space of all possible parameter subsets, b) how to evaluate the prediction performance and c) the predictive model. One approach to evaluate the prediction performance of the model is to count the number of errors made on a subset defining its score. Nonetheless, because wrapper methods train a new model for each subset, they tend to be computationally expensive. Wrapper methods include approaches such as decision trees [1][2].
In our survey, we especially focus on decision trees and their various adaptions as they can handle heterogeneous data, missing values, different parameter scales and nonlinearities [2].
Embedded methods select features using the information gained from training a learning algorithm instead of treating it as a black box. Popular methods are Lasso regression using the L1 regularization norm and Ridge regression using the L2 regularization norm [2].

Several different imputation strategies of how to deal with missing values and their applicability for our task will be presented as well.

Our survey compares the following feature selection methodologies: decision tree, decision tree feature importance using Gini importance, Randomized Logistic Regression with L1 regularization, normal Logistic Regression with L2 regularization and Pearson Correlation.

We conclude with an evaluation of the usefulness of the results of the applied feature selection methods and their selected features on the technical side as well as the medical soundness of the selected features.

[1] Guyon, Isabelle, and André Elisseeff. "An introduction to variable and feature selection." Journal of machine learning research 3.Mar (2003): 1157-1182.

[2] Deng, Houtao, and George Runger. "Feature selection via regularized trees." Neural Networks (IJCNN), The 2012 International Joint Conference on. IEEE, 2012.