mgr inż. Wojciech Sikora

Machine learning-based workflow for the analysis of MALDI-TOF mass spectrometry

Abstract

The subject of the dissertation is the analysis of data acquired by mass spectrometry imaging of samples obtained from patients with head and neck cancer. The following hypotheses were made in the thesis. The first hypothesis states that peak identification in mass spectra can be successfully performed using a spectrum modeling approach by fragmenting the spectrum into parts and then modeling them with Gaussian mixture models. The second hypothesis states that the spatial distribution information obtained through imaging can be used to remove redundancy and reduce the dimensionality of the data, while maintaining the quality of the data. The final hypothesis states that evaluating the importance of features in heterogeneous data is possible and effective through the use of multiple unit models. The first chapters of the thesis address the basic issues related to proteomics and mass spectrometry. First, the general description of mass spectrometry and mass spectrometric imaging of biological samples is described. This is followed by a description of the main ionization methods and mass analyzers commonly used for the analysis of biological samples, especially samples from cancer patients. Then, there is a brief description of sample preparation, as well as data acquisition, its characteristics, and the initial steps taken to prepare the data for further analysis. These steps are baseline correction, normalization and alignment of the spectra.

The next chapter deals with the aggregation of mass spectra and the state of the art in peak detection. Peak detection was performed on the aggregated data using the most commonly used for this purpose methods. First, peaks were identified using a simple method based on the signal-to-noise ratio of peak intensities. Then peaks were identified with a peak modeling method based on the continuous wavelet transform.

In the following chapter, a more complicated method of peak identification was described in detail. With this method, peaks are identified by splitting the spectrum into smaller fragments and modeling them with Gaussian mixture models. First, a new signal

splitting method is described that differs from the method proposed in the original paper. A detailed operation scheme is described, and compared with the original method as well as the pseudocode for the algorithm implementation. The next section of the paper deals with the process of fitting the parts of the spectrum with Gaussian mixture models, with the general and mathematical description of the custom implementation of the expectation-maximization (EM) algorithm used for the fitting of Gaussian mixtures. The thesis also describes the selection of the optimal number of elements in the mixture and the influence of the stochastic nature of the EM algorithm on the results. All peak identification methods are compared to each other. The results of proposed peak identification method confirm the validity of the first hypothesis.

The sixth chapter describes the entire process of feature engineering. The deals with the use of statistics and spatial distribution to remove redundancy in the data and reduce the dimensionality of the data. To this end, noise was filtered using the parameters of the normal distributions that make up the spectrum model. Feature engineering is then continued by using the information provided by the imaging. The spatial distributions of nearby elements of the spectrum model are compared. The comparison is made using Peacock's statistical test for similarity of distributions. This statistical test is an extension of the Kolmogorov-Smirnov test to two dimensions. The critical values are calculated experimentally, and then the nearby elements with statistically identical special distribution are merged. The dimensionality reduction process ends with the detection of isotopic envelopes, which are also reduced to a single feature. Isotopic envelopes are detected by examining the distance between successive peaks, their shape, and their spatial distribution. The results show a significant reduction in the dimensionality of the data, from 9454 elements of the spectrum model to 888 features in the final set. These results confirm the second hypothesis of the paper.

The following sections describe the training of the classifiers on the processed data. Two groups of classifiers were trained. The first group was trained with an algorithm that uses multinomial logistic regression. The model is trained by iteratively performing logistic regression to find the best feature from the remaining set and adding it to the predictor list of the final model. The second set of classifiers are fully connected neural networks with two hidden layers, where the number of nodes is equal to the number of features. The performance of the classifiers was evaluated using metrics such as accuracy, precision, negative predictive value, sensitivity, specificity, f1 score, and ROC curves, precision-sensitivity curves, and their areas under the curve.

The process of feature importance evaluation was described next. Feature importance is assessed by assigning a score to each feature in unit models and averaging

the results to determine the total feature importance score. For logistic regression models, scores are based on the feature's place in the predictor list. For neural networks, black-box model interpretation methods were used, LIME and Shapley values.

The last chapter of the thesis is the discussion about the experiments, the results, the conclusions drawn, and the goals for the future.