Statistical methods for mass spectrometry-based clinical proteomics

Kakourou, A.A.

Summary

This dissertation is concerned with new statistical methods for the construction of diagnostic rules based on clinical mass spectrometry proteomic data. Our main interest focuses on the comparison of mass spectral proteomic profiles collected from healthy individuals and cancer patients in the context of distinct case-control studies. A key objective in such studies is the construction of discriminant rules for distinguishing between individuals as to the presence or absence of the disease as well as for predicting the health status of future patients.

Throughout this thesis, we consider two different types of mass spectrometry-based proteomic data, obtained with two distinct technologies. The first data we consider contain mass spectral measurements of affected and unaffected individuals, the samples of which were collected in a breast cancer case-control study. The proteomic profile of each of those individuals was generated using MALDI-TOF mass spectrometry. The second data set we analyse consists of proteomic expression measuements of healthy individuals and pancreatic cancer patients. These profiles were obtained using a MALDI-FTICR technology. In what follows, we give detailed description of the novel statical methods presented in this thesis for the construction of diagnostic allocation rules which can deal at the same time with the additional statistical challenges which accompany these particular types of data.

Chapter 1 provides a general introduction to some of the key concepts in clinical proteomic research. We start by introducing the process by which the data are generated as well as the pre-processing steps that need to be taken in order to translate the raw proteomic expression into a summarized data set which can be used in downstream statical analyses. Next, we describe the special characteristics of the MS proteomic data and discuss the statistical challenges which may arise when analyzing such data. We proceed with introducing the basic steps involved in proteomic diagnosis. These include: a) the choice of model, b) the construction of the diagnostic rule and c) the evaluation of the diagnostic rule. We give a detailed description of each of these steps.

Chapters 2 and 3 describe methods for combining different types of information in order to enhance calibration ability and predictive performance of allocation rules in clinical diagnostic studies. Chapter 2 presents an approach for combining a collection of distinct classification methods. Chapter 3 reviews various methods, among which an adaptation of the combination approach introduced in chapter 2, for combining a collection of distinct omic data sets. The proposed combination methods are illustrated through the analysis of

the MALDI-TOF breast cancer case-control data.

Chapter 2 presents an approach to construct a classification rule through the combination of distinct classifiers in order to improve predictions. In a first stage this is achieved by building individual classification models separately, each one using the entire proteomic data set. A double leave-one-out cross-validatory approach is used in order to calibrate the independent classifiers and obtain unbiased estimates of the class predicted probabilities which will later be used as new input variables for the construction of the "combined" classification rule. We present two different approaches to combine the estimated class probabilities in order to derive the combination-based rule. The first approach is based on weighted sums and it is more straightforward in application since it does not require optimization of the weights. Unbiased evaluation of the predictive performance of such weighted sum discriminant combination rule is then relatively easy to derive, as the cross-validatory nature of the predicted probabilities of each separately calibrated classifier is preserved entirely. The second approach is to regard the estimates of each constituent classifier as a feature for the calibration of a model-based combination which can be more flexible and can lead to a more elaborate classification rule. Given that the estimated class probabilities are fitted themselves, the model-based combination requires to be embedded in a cross-validatory scheme itself in order to obtain unbiased estimates of the final "combined" classification rule.

The combination ideas presented in this chapter were motivated by the results, suggestions and discussion generated in the mass spectrometry proteomic competition (Mertens, 2008; Hand, 2008). A key result of the competition was that the majority of the participants chose simple linear methods to analyze the proteomic data while they all reported similar classification results. The idea of investigating an alternative approach to using one single method to calibrate the diagnostic rule, that is to use a collection of distinct classification procedures to build a combined classifier, was initially recommended by Hand (2008) as "an effective way to improve classification predictions". Indeed, results from the real data analysis as well as a simulation study based on reusing the proteomic mass spectra data showed that in many situations gains in classification performance and predictive accuracy can be achieved through a combination-based approach. We restricted the constituent classifiers used in the construction of the combined classifier to linear classification methods, in the first instance because these produced favorable results in the proteomic competition. Moreover, these methods are regarded as reliable and stable generally for high-dimensional data problems with relatively small sample sizes in Omics research.

Chapter 3 addresses the problem of integrating distinct omic data sets in order to derive a "combined" diagnostic rule which outperforms single-omic-based prediction models. While the combination approaches presented in chapter 2 are based on combining predicted estimates obtained from different methods but based on a common source of predictors, the combination methods discussed in chapter 3 are applied in the context of combination of predictions of a common outcome based on different sources of predictors.

We consider two different types of procedures for combination of single-source predictions, both relying on regularized regression. The first type is a “parallel” approach, in which a new vector of predictions is obtained as a weighted sum of single-source predictions of the outcome of interest. The specific weight may be fixed beforehand, leading to a weighted sums combination approach (averaging the single-source predictions) or estimated, leading to a model-based combination approach. Both of these approaches were introduced in chapter 1 in the context of combining estimates derived based on distinct classifiers fitted to the same proteomic data set. On the other hand, the methods discussed in chapter 3 are applied to the estimates derived using the same classifier but fitted to distinct (omic) data sets. As in chapter 2, the individual predictions in the model-based combination are treated as new covariates and therefore a cross-validatory setting is required in order to calibrate the final combination rule and obtain unbiased final combined predictions. The second type of combination methods we consider is a “sequential” approach in which we choose *a priori* one of the omic sources as “primary”. In terms of model formulation this translates into introducing the vector of individual predictions based on the “primary” source as an extra covariate with fixed weight when fitting a prediction model based on the “secondary” source of omic predictors. In the case of regularized logistic regression, this is implemented by considering the vector of predictions based on the “primary” source as an offset term. As in the model-based parallel combination, the use as covariate of a vector of predictions (which are fitted themselves) requires an extra layer of cross-validation to account for the uncertainty of calibrating the first source of predictors in the procedure.

To evaluate the proposed combination methods we regard two independent omic applications. Firstly, we revisit the problem of the combination of different fractionations in proteomic spectrometry for breast cancer diagnosis. Secondly, we assess the possible combination of transcriptomics and metabolomics for the prediction of obesity. Results from both applications show that improved estimates can be obtained by combining predictions based on different omic sources, outperforming single-omic predictions.

Chapter 4 until 6 present new methods to deal with the additional statistical challenges researchers may be presented with when analyzing high-throughput MALDI-FTICR mass spectrometry data. All analyses presented in these chapters are based on the data from the pancreatic cancer case-control study and the proposed methods are specific to this particular type of MS-based proteomic data.

Chapter 4 considers the problem of summarizing and analyzing MALDI-FTICR mass spectrometry data in an appropriate and meaningful way. To this end, a fast and simple procedure to preprocess the individual spectra prior to summarizing and analyzing the acquired data is proposed. The objective of the preprocessing algorithm is to identify the isotopic clusters in the individual spectra and quantify their corresponding isotopic peaks by utilizing prior knowledge on isotopic clustering information. The procedure relies solely on the known statistical properties of MALDI-FTICR MS data, in particular that successive isotopic peaks of a peptide molecule are commonly separated by 1 Da. This renders the approach particularly relevant in the context of large-scale clinical studies as it can be applied fast across many spectra from different patients. The algorithm finds m/z positions of peaks which belong to isotopic clusters in a completely non-parametric fashion, thus avoiding any computationally intensive steps like model fitting to the observed spectra. Through this procedure, the complete expression in the individual spectra is reduced to clusters of isotopic expression on which summary measures can be defined.

We investigate various ways of summarizing the expression within the derived isotope clusters with the objective to evaluate the different choices with regards to their predictive ability. To derive the summary measures, information on either the overall intensity level or the shape of the observed isotopic cluster pattern is used. In this way we investigate whether there is additional information in the shape of the isotopic cluster patterns, associated with the heath status, which is independent of information in the intensity level. We propose several measures to carry out these investigations, applied either to the observed isotopic cluster pattern directly or to the residuals measuring the deviation of that observed pattern from the expected one. Our empirical results show that both intensity and shape carry information on the class outcome and can be used for prediction purposes. However, summary measures based on intensity were found to be more informative than summary measures based on shape.

The proposed methods and associated analyses presented in chapters 5 and 6 are based and dependent on the data generated by the pre-processing algorithm introduced in chapter 4. Each of the approaches presented in these chapters explores a distinct problem which may arise in proteomics research when working with FTICR proteomic data, which was not addressed in chapter 4.

Chapter 5 follows up on the previously reported results asserting the predictive power of the pancreatic cancer data, and more specifically the predictive ability of isotope clusters in high-resolution MS data, and its objective is to: a) to identify a collection of isotope clusters associated with the disease outcome and b) to optimally integrate the intensity and shape infirmation. These questions are addressed in a way which allows for the assessment of the added-value of shape beyond intensity in predicting the health status of an individual while maintaining predictive performance. We proposed an approach which utilizes our prior knowledge about the relative predictive power of each individual source as well as the distinctive structure of the proteomic data, namely that intensity and shape measures are tied together in pairs of isotope expression. To explore the problem of isotope selection and at the same time address the problem of assessing the additional predictive value of shape, we used a Bayesian model formulation by which we can introduce multiple layers of selection. In doing so, we made the explicit assumption that the shape source is complementary. In terms of model fitting, this is translated by assuming that a shape measure can be included in the set of predictors on the condition that it is accompanied by/coupled with its corresponding intensity while the reverse does not need to hold. This assumption allows us to make simultaneous inference on which isotope clusters are the most informative with respect to the class outcome and for which isotopes - and to what extent - shape has a complementary value in separating the two groups. We presented a post-hoc analysis which allows researchers to focus on a restricted set of potentially interesting isotope clusters for further investigation and gives insight into the relative predictive capacity of shape when integrated with intensity. We finally presented results from a simulation study to demonstrate how the method behaves under a controlled situation. Results from this simulation study indicated that the proposed Bayesian approach is able to successfully select the isotope clusters which are truly predictive of the class outcome as well as accurately estimate the effects of both intensity and shape measures although, in some cases, the model tends to overestimate the relative importance of shape after accounting for intensity.

Chapter 6 describes new methods for the calibration of diagnostic rules based on high-resolution mass spectrometry proteomic data which may be subject to the limit of detection (LOD). The LOD is related to the limitation of instruments in measuring low-concentration proteins. As a consequence, peak intensities below the LOD threshold are often reported as missing values during the quantification step of proteomic analysis. We describe an adaption of censored regression methodology in order to estimate the average individual expression within isotopic clusters, prior to calibration of prediction rules, assuming that any missing intensities in our data set are due to left-censoring mechanisms on low abundant proteins. Our objective is to investigate whether, starting from the incomplete proteomic data, we can recuperate information on the average proteomic expression which will allow as to construct prediction rules of comparable performance as if we had the complete information. The proposed methods are evaluated with respect to their predictive performance, by replacing the incomplete spectral measurements with the derived estimates of individual expression, accounted for the LOD, and using those as predictors for the construction of diagnostic rules. We combined censored regression with borrowing of information across data to account for potential lack of information and measurement uncertainty. In terms of model formulation, this is achieved by including individual-specific random effects in the censored regression model. We demonstrated different variants of using the random effect formulation for calibration and assessment of the final prediction rules and we additionally showed how this formulation may allow for a type of variable selection based on the random effect variance. Results from both internal and external validations indicated that using the estimates from the proposed methods as input variables results in comparable predictive accuracy to the one achieved using the complete intensity information. Ignoring the unobservable peak intensities, as an alternative to deal with the LOD, resulted in poor predictions as compared to the proposed methods, while substituting the unobservable peak intensities with the LOD value, a method commonly used in proteomic reserach, exhibited similar classification performance as the proposed methods.