EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

QBiC – TechTalk

Quantitative Biology Center (QBiC)

Tuesday, June 30, 2015, 4.15 pm,
Lecture hall N11, Hörsaalzentrum, Auf der Morgenstelle 3

"Machine Learning for Personalized Medicine: How to Integrate and Interpret Data from Different Molecular Measurements"

Speaker:
Dr. Nico Pfeiffer, Max Planck Institute for Informatics,
Computational Biology & Applied Algorithmics,
Saarbrücken

DFG Deutsche Forschungsgemeinschaft

Machine Learning for Personalized Medicine: How to Integrate and Interpret Data from Different Molecular Measurements

Dr. Nico Pfeiffer, Max Planck Institute for Informatics, Computational Biology & Applied Algorithmics, Saarbrücken

Personalized or stratified medicine promises better, more effective and possibly even cheaper treatments for the majority of patients. More and more different high-throughput techniques are available to enable the discovery of potential biomarkers that can help predict how certain patient subgroups behave compared to other subgroups (e.g., will a certain cancer patient benefit from a specific cancer treatment or not). Methods to really make sense of the masses of data are slightly lacking behind both on the level of prediction performance and in the means provided for interpreting the results. In the following I describe two projects addressing these problems, which will be presented during my talk:

In standard biomedical analyses principal component analysis is used as a dimensionality reduction technique to get an exploratory view of the sample similarities. Having several different measurements per patient, this is not straight-forward anymore.
Therefore, we applied and extended current multiple kernel learning for dimensionality reduction approaches. On the one hand, we added a regularization term to avoid over-fitting during the optimization procedure, and on the other hand, we showed that one can even use several kernels per data type and thereby alleviate the user from having to choose the best kernel functions and kernel parameters for each data type beforehand.
Applied to molecular measurements from cancer patients, we identified biologically meaningful subgroups for five different cancer types.
Survival analysis revealed significant differences between the survival times of the identified subtypes, with p-values comparable or even better than state-of-the-art methods.
Moreover, our resulting subtypes reflect combined patterns from the different data sources, and we demonstrated that input kernel matrices with only little information have less impact on the integrated kernel matrix. Our subtypes show different responses to specific therapies, which could eventually assist in treatment decision-making.
The method could be used for many other patient groups and many other diseases.

The second project deals with personalized prediction of phenotypes based on molecular measurements like DNA methylation from whole genome bisulfite sequencing.
Molecular measurements from cancer patients can be influenced by several external factors. This makes it harder to reproduce the values of the measurements. Furthermore, cancer types can be very heterogeneous meaning that there might be different underlying causes for the same type of cancer among different individuals. When trying to predict the stage of a certain cancer type, this can lead to problems if the model is not taking into account those potential biases in the data, especially if they differ between training and test set.
We introduced a method that can estimate these biases on a per-feature level and incorporate calculated feature confidences into a weighted combination of classifiers with disjoint feature sets. In this way, the method makes a prediction available that is adjusted for the potential biases on a per-patient basis, providing a personalized prediction for each test patient. Moreover, we showed how to visualize the learned classifiers to display interesting associations with the target label. Applied to a leukemia data set we found several ribosomal proteins associated with the risk group, which might be interesting targets for follow-up studies.