Fishing for bio-marker candidates in ms-data by use of Learning Vector Quantization

Frank Michael Schleif
University of Leipzig, Dept. of Mathematics and Computer Science, Leipzig, Germany
Bruker Daltonik GmbH, R. & D., Numerical-Toolbox-Group, Leipzig/Bremen, Germany

During last years, proteomic profiling based on mass spectrometry became an important tool for studying diseases at the protein and peptide level in a higher throughput manner. Thereby the identification of characteristic masses for a specific kind of disease is a complicated task. In addition the classification accuracy and safety is especially important for medical applications. The talk gives an overview about current research activities in the field of Learning Vector Quantization with a special focus on classifier development for the search of proteomic patterns. The first part of the talk shows basic data pre-processing steps which are necessary to obtain discriminant features for a subsequently classification procedure. In the second part a new approach for fuzzy labeled learning vector quantization is introduced. In this method one allows dynamically adapted fuzzy labels to indicate the responsibility of prototypes for a class. The new learning algorithm performs a gradient descent on a cost function derived from the Neural GAS algorithm. The proposed algorithm generates a prototype based classifier. Thereby, to each prototype a fuzzy class information is assigned which indicates the probabilities for class responsibility. Further the algorithm is capable for automatically metric adaptation which aims on feature selection and hence the search for bio-marker candidates. In addition this algorithm can handle fuzzy labeled data which allows the clinicians to keep a (small) fuzziness in their diagnosis instead of making it unnecessary strict. The new approach is described and its performances on real life data taken from clinical studies is demonstrated. Its shown how this new method can be successfully applied to the analysis of proteomic data and used for bio marker research.