Advances in pre-processing and model generation for mass spectrometric data analysis

Frank-Michael Schleif[1]

University Leipzig
Department of Medicine
04107, Leipzig, Karl-Tauchnitz 25, Germany
schleif@informatik.uni-leipzig.de

Abstract. The analysis of complex signals as obtained by mass spectrometric measurements is complicated and needs an appropriate representation of the data. Thereby the kind of preprocessing, feature extraction as well as the used similarity measure are of particular importance. Focusing on biomarker analysis and taking the functional nature of the data into account this task is even more complicated. A new mass spectrometry tailored data preprocessing is shown, discussed and analyzed in a clinical proteom study compared to a standard setting.

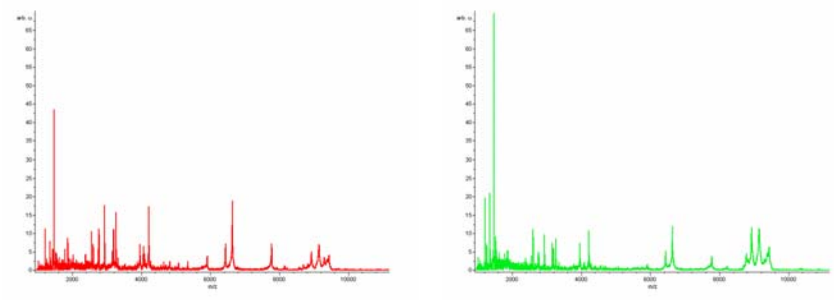
Keywords. similarity measures, functional data, proteomics, mass spectrometry, pre-processing, wavelet analysis, generalized peak list

1 Introduction

Analysis and visualization of clinical proteomic spectra obtained from mass spectrometric measurements is a complicated issue [1]. One major objective is the search for potential biomarkers in complex body fluids like serum, plasma, urine, saliva, or cerebral spinal fluid [2,3,4,5]. Typically the spectra are given as high-dimensional vectors. Thus, from a mathematical point of view, an efficient analysis and visualization of high-dimensional data sets is required. Moreover, the amount of available data is restricted: usually patient cohorts are small in comparison to data dimension. A further problem is that uncertainty in the data may occur. For example, the clinical diagnosis of a patient may be uncertain (fuzzy). Yet, most of machine learning classification models assume strict (crisp) decisions for training data. All these aspects show that classification learning is a crucial task.

The self-organizing map (SOM) constitutes one of the most popular unsupervised approaches for clustering, visualization and data mining of high-dimensional data [6]. SOMs belong to the prototype based methods of data representation. Due to its inherent regularization abilities SOMs are also applicable in case of sparse data sets. Basically, SOMs map the data nonlinearly onto a low-dimensional regular lattice of neurons in a topology-preserving fashion by

Dagstuhl Seminar Proceedings 07131
Similarity-based Clustering and its Application to Medicine and Biology
http://drops.dagstuhl.de/opus/volltexte/2007/1132



(a) Cancer Spectrum (b) Control Spectrum

Fig. 1. (a) MALDI-TOF spectrum of a colorectal cancer patient and (b) a healthy subject after peptide isolation with C8 magnetic beads. On the Y-axis the relative intensity is shown. The mass to charge ratio (m/z) is demonstrated on the X-axis in Dalton. The spectra are already preprocessed (baseline correction,recalibration) using ClinProTools 2.1

means of prototype matching, i.e. similar data points are mapped onto nearby or identical neurons under certain conditions [7]. Thereby, adaptation takes place as an unsupervised prototype learning. Recently, a semi-supervised counterpart is developed [8]. It allows the determination of a prototype based fuzzy classification model (FLSOM). In contrast to the widely applied multilayer perceptron [9], prototype based classification allows an easy interpretation, which is of particular interest for many (clinical) applications. FLSOM leads to a robust fuzzy classifier where efficient learning of fuzzy labeled or partially contradictory data is possible. Additionally, FLSOM gives the possibility to assess and to visualize *class similarity* by inspection of the generated class map, which represents the label distribution according to the FLSOM lattice structure and the learned class information. However, FLSOM differs from existing extensions of SOM for classification tasks like counterpropagation [10] or Fuzzy SOM [11] fundamentally: In contradiction to these models, for FLSOM the prototype adaptation is also influenced by the class information of the given data such that optimization according to class information is incorporated into the adaptation scheme.

In this contribution, after an introduction of the FLSOM approach and its theory, we apply the algorithm to the problem of classification of mass spectra in case of cancer disease. We show for a data set of colorectal cancer patients and controls, which was also used in a previous study, the successful application of our approach.

2 Data analysis by FLSOM

The fuzzy labeled self-organizing map (FLSOM) is a prototype based classification model, which is able to handle fuzzy labeled data (uncertain class decision) during training and which return fuzzy class decisions during recall. FLSOM is

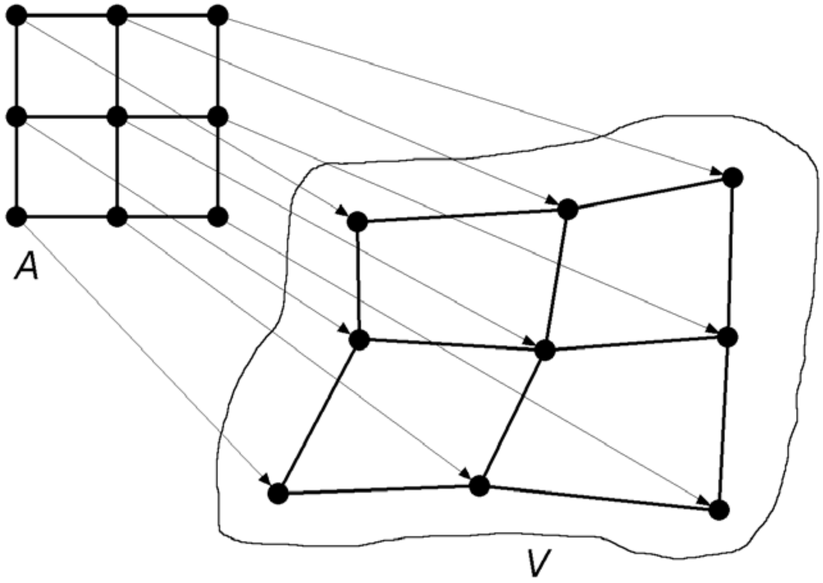


Fig. 2. The figure shows a mapping of a vector space (V) on a rectangular grid (A). Prototypes are associated with the rectangular grid by arrows. Multiple points in the vector space maybe represented by a single prototype which is associated with a grid position. In case of topological preservation interpretation of the mapping can be transfered to the potential high dimensional vector space.

an extension of the unsupervised self-organizing map (SOM). Therefore, we first shortly introduce the SOM and thereafter we develop the FLSOM scheme.

2.1 The Self-Organizing Map

As mentioned above, SOMs can be taken as unsupervised learning of topographic vector quantization with a topological structure (grid) within the set of prototypes (codebook vectors). Thereby, roughly speaking, topology preservation means that similar data points $\mathbf{v} \in V$ are mapped onto identical or neighbored grid locations which have pointers into the data space (weight vectors). The principle is depicted in Figure 2.

An exact mathematical definition is given in [7]. The weight vectors also are called prototypes, because they represent parts of the data space.

There exists a wide range of applications in pattern recognition ranging from spectral image processing to bioinformatics. The mathematics behind the original SOM model as proposed by Kohonen is rather complicated. In particular, the training process does not follow a gradient descent on any cost function for continuous data distributions [12]. However, Heskes proposed a variant of the original algorithm which, in practice, leads to at least very similar or identical results as the original SOM but for which a cost function can be established [13]. We will base our model on this formulation:

Assume that data $\mathbf{v} \in V \subseteq \mathbb{R}^d$ are given distributed according to an underlying distribution $P(V)$. A SOM is determined by a set A of neurons $\mathbf{r}$ equipped with weight vectors (prototypes) $\mathbf{w_r} \in \mathbb{R}^d$. The neurons are arranged on a lattice structure, which determines the neighborhood relation $N(\mathbf{r}, \mathbf{r}')$ between the neurons $\mathbf{r}$ and $\mathbf{r}'$. Denote the set of prototypes by $\mathbf{W} = \{\mathbf{w_r}\}_{\mathbf{r} \in A}$. The mapping description of a trained SOM defines a function

$$\Psi_{V \to A} : \mathbf{v} \mapsto s(\mathbf{v}) = \operatorname*{argmin}_{\mathbf{r} \in A} le(\mathbf{r}) \tag{1}$$

where

$$le(\mathbf{r}) = \sum_{\mathbf{r}' \in A} h_\sigma(\mathbf{r}, \mathbf{r}') \xi(\mathbf{v}, \mathbf{w}_{\mathbf{r}'}) \tag{2}$$

is the local neighborhood weighted error of distances $\xi(\mathbf{v}, \mathbf{w}_{\mathbf{r}'})$. $\xi(\mathbf{v}, \mathbf{w})$ is an appropriate distance measure, usually the quadratic Euclidean norm $\xi(\mathbf{v}, \mathbf{w_r}) = (\mathbf{v} - \mathbf{w_r})^2$. However, here we only suppose $\xi(\mathbf{v}, \mathbf{w})$ to be arbitrary assuming differentiability, symmetry and assessing some dissimilarity. The function

$$h_\sigma(\mathbf{r}, \mathbf{r}') = \exp\left(\frac{N(\mathbf{r}, \mathbf{r}')}{2\sigma^2}\right) \tag{3}$$

determines the neighborhood cooperation with range $\sigma > 0$. Large values of σ also correspond to high regularization whereas low values ignore this feature. In this formulation, an input stimulus $\mathbf{v}$ is mapped onto that position $\mathbf{r} \in A$ of the SOM, the local error $le(\mathbf{r})$ of which is minimum, whereby the average over all neurons according to the neighborhood is taken. We refer to this neuron $s(\mathbf{v})$ as the winner.

During the adaptation process a sequence of data points $\mathbf{v} \in V$ is presented to the map representative for the data distribution $P(\mathcal{V})$. Each time the currently most proximate neuron $s(\mathbf{v})$ according to (1) is determined. All prototypes are gradually adapted according to the neighborhood degree of the respective neuron to the winning one by

$$\triangle\, \mathbf{w_r} = -\epsilon h_\sigma(\mathbf{r}, s(\mathbf{v})) \frac{\partial \xi(\mathbf{v}, \mathbf{w_r})}{\partial \mathbf{w_r}} \tag{4}$$

with a small learning rate $\epsilon > 0$. This adaptation follows a stochastic gradient descent of the cost function introduced by Heskes [13]:

$$E_{\mathrm{SOM}} = \frac{1}{2C(\sigma)} \int P(\mathbf{v}) \sum_{\mathbf{r}} \delta_{\mathbf{r}}^{s(\mathbf{v})} \sum_{\mathbf{r}'} h_\sigma(\mathbf{r}, \mathbf{r}') \xi(\ \mathbf{v}, \mathbf{w}_{\mathbf{r}'}) d\mathbf{v} \tag{5}$$

were $C(\sigma)$ is a constant which we will drop in the following, and $\delta_{\mathbf{r}}^{\mathbf{r}'}$ is the usual Kronecker symbol checking the identity of $\mathbf{r}$ and $\mathbf{r}'$.

One main aspect of SOMs is the visualization ability of the resulting map due to its topological structure. Under certain conditions the resulting non-linear projection $\Psi_{V\rightarrow A}$ generates a continuous mapping from the data space V onto the grid structure on A. This mapping can mathematically be interpreted as an approximation of the principal curve or its higher-dimensional equivalents [14]. Thus, as pointed out above, similar data points are projected on prototypes which are neighbored in the grid space A. Further, prototypes neighbored in the lattice space should code similar data properties, i.e. their weight vectors should be close together in the data space according to the dissimilarity measure ξ. This property of SOMs is called topology preserving (or topographic) mapping realizing the mathematical concept of continuity. For a detailed and mathematical exact consideration of this topic we refer to [7]. Successful tools for assessing this map property are the topographic function and the topographic product [7],[15].

2.2 Fuzzy Labeled SOM (FLSOM)

SOM is a well-established model for nonlinear data visualization which, due to its above mentioned topology preserving properties, can also serve as an adequate preprocessing step for data completion, representation or interpolation. The formulation of the adaptation scheme in terms of a gradient descent of a cost function allows an extension to a semi-supervised learning scheme which leads to a classification model. Thereby, the resulting FLSOM is able to handle uncertainty in class assignments of training data as well as returns fuzzy classification decision in the recall phase. It differs from simple post labeling or separate post-learning of prototype labels as it takes place in counter propagation [10] or Fuzzy-SOM [11] in this way that in FLSOM the prototype adaptation is influenced by the class information. We now explain the model in detail.

Let $N(c)$ be the number of possible data classes. We assume that each training point $\mathbf{v}$ now is equipped with a label vector $\mathbf{x} \in \mathbb{R}^{N(c)}$ whereby each component $x_i \in [0,1]$ determines the soft assignment of $\mathbf{v}$ to class i for $i = 1, \ldots, N(c)$. Hence, we can interpret the label vector as probabilistic or possibilistic fuzzy class memberships. Accordingly, we enlarge each prototype vector $\mathbf{w_r}$ of the map by a label vector $\mathbf{y_r} \in [0,1]^{N(c)}$ which determines the portion of neuron $\mathbf{r}$ assigned to the respective classes. During training, prototype locations $\mathbf{w_r}$ and label vectors $\mathbf{y_r}$ are adapted according to the given labeled training data. For this purpose, we extend the cost function of the SOM as defined in (5) to a cost function for fuzzy-labeled SOM (FLSOM) by an term E_{FL} assessing classification accuracy. Thus the cost function becomes

$$E_{\mathrm{FLSOM}} = (1-\beta)\, E_{\mathrm{SOM}} + \beta E_{\mathrm{FL}} \tag{6}$$

where the factor $\beta \in [0,1]$ is a *balance factor*, which determines the influence both aspects data representation by usual SOM and classification accuracy. For the classification accuracy term we chose

$$E_{\mathrm{FL}} = \frac{1}{2} \sum_{\mathbf{r}} \int P\left(\mathbf{v}\right) \cdot ce\left(\mathbf{v}, \mathbf{r}\right) d\mathbf{v} \tag{7}$$

with *local, weighted classification errors*

$$ce\left(\mathbf{v},\mathbf{r}\right)=g_{\gamma}\left(\mathbf{v},\mathbf{w_{r}}\right)\cdot\vartheta\left(\mathbf{x}\left(\mathbf{v}\right),\mathbf{y_{r}}\right). \tag{8}$$

$g_{\gamma}\left(\mathbf{v},\mathbf{w_{r}}\right)$ is a Gaussian kernel defining a neighborhood range in the data space:

$$g_{\gamma}\left(\mathbf{v},\mathbf{w_{r}}\right)=\exp\left(-\frac{\xi\left(\mathbf{v},\mathbf{w_{r}}\right)}{2\gamma^{2}}\right). \tag{9}$$

The value $\vartheta\left(\mathbf{x}\left(\mathbf{v}\right),\mathbf{y_{r}}\right)$ describes the dissimilarity of the label vectors $\mathbf{x}$ and $\mathbf{y_{r}}$. Usually, the squared Euclidean distance $\vartheta\left(\mathbf{x}\left(\mathbf{v}\right),\mathbf{y_{r}}\right)=\left(\mathbf{x}-\mathbf{y_{r}}\right)^{2}$ is chosen. However, as in the case for the dissimilarity in the data space, other definitions are possible.

This choice of the classification accuracy term E_{FL} as a sum of weighted data space distances is based on the assumption that data points, close to a prototype $\mathbf{w_{r}}$, determine the corresponding label, if the underlying class distribution is sufficiently smooth. Note that the kernel $g_{\gamma}\left(\mathbf{v},\mathbf{w_{r}}\right)$ depends on the prototype locations, such that the classification term E_{FL} is influenced by both $\mathbf{w_{r}}$ and $\mathbf{y_{r}}$. Hence, the gradient of E_{FL} with respect to $\mathbf{w_{r}}$ is non-vanishing and yields

$$\frac{\partial E_{\mathrm{FL}}}{\partial\mathbf{w_{r}}}=-\frac{1}{4\gamma^{2}}\int P\left(\mathbf{v}\right)\cdot g_{\gamma}\left(\mathbf{v},\mathbf{w_{r}}\right)\cdot\frac{\partial\xi\left(\mathbf{v},\mathbf{w_{r}}\right)}{\partial\mathbf{w_{r}}}\cdot\vartheta\left(\mathbf{x}\left(\mathbf{v}\right),\mathbf{y_{r}}\right)d\mathbf{v} \tag{10}$$

which contribute to the overall gradient by

$$\frac{\partial E_{\mathrm{FLSOM}}}{\partial\mathbf{w_{r}}}=(1-\beta)\cdot\frac{\partial E_{\mathrm{SOM}}}{\partial\mathbf{w_{r}}}+\beta\cdot\frac{\partial E_{\mathrm{FL}}}{\partial\mathbf{w_{r}}} \tag{11}$$

Thus the complete prototype update becomes

$$\begin{aligned}\triangle\,\mathbf{w_{r}}=&-\epsilon(1-\beta)\cdot h_{\sigma}\left(\mathbf{r},s(\mathbf{v})\right)\frac{\partial\xi\left(\mathbf{v},\mathbf{w_{r}}\right)}{\partial\mathbf{w_{r}}}\\&+\epsilon\beta\frac{1}{4\gamma^{2}}\cdot g_{\gamma}\left(\mathbf{v},\mathbf{w_{r}}\right)\cdot\frac{\partial\xi\left(\mathbf{v},\mathbf{w_{r}}\right)}{\partial\mathbf{w_{r}}}\cdot\vartheta\left(\mathbf{x}\left(\mathbf{v}\right),\mathbf{y_{r}}\right).\end{aligned} \tag{12}$$

The gradient of E_{FLSOM} with respect to the label determines the adaptation rule for the prototype labels. Because E_{SOM} is independent on the prototype labels the respective derivative vanishes and we get

We obtain the update rules by taking the derivatives: Labels are only influenced by the second part E_{FL}, which yields

$$\frac{\partial E_{\mathrm{FLSOM}}}{\partial\mathbf{y_{r}}}=\frac{\partial E_{\mathrm{FL}}}{\partial\mathbf{y_{r}}} \tag{13}$$

and the corresponding learning rule therefore is

$$\triangle\,\mathbf{y_{r}}=\epsilon_{l}\beta\cdot g_{\gamma}\left(\mathbf{v},\mathbf{w_{r}}\right)\left(\mathbf{x}-\mathbf{y_{r}}\right) \tag{14}$$

with learning rate $\epsilon_{l}>0$. This learning scheme can be seen as a weighted average of the data fuzzy labels of those data $\mathbf{v}$ close to the associated prototype $\mathbf{w_{r}}$.

2.3 Topography and label distribution in FLSOM

As mentioned above, unsupervised SOMs generate a topographic mapping from the data space onto the prototype grid under specific conditions. If the classes are consistently determined with respect to the varying data, one can expect for supervised topographic FLSOMs that the labels become ordered within the grid structure of the prototype lattice. In this case the topological order of the prototypes should be transferred to the topological order of prototype labels such that we have a smooth change of the fuzzy class label vectors between neighbored grid positions $\mathbf{r}$. This is the consequence of following fact: the neighborhood function $h_{\sigma}(\mathbf{r}, \mathbf{s})$ of the usual SOM learning (4) forces the topological ordering of the prototypes. In FLSOM, this ordering is further influenced by the weighted classification error $ce(\mathbf{v}, \mathbf{r})$ (8). This classification error term contains the kernel $g_{\gamma}(\mathbf{v}, \mathbf{w}_{\mathbf{r}})$, eq. (9). Hence, the prototype learning and ordering (12) receives information of both data and class distribution. Thereby, for high value of the balancing parameter β the latter term becomes dominant. Otherwise, the kernel $g_{\gamma}(\mathbf{v}, \mathbf{w}_{\mathbf{r}})$ also triggers the label learning (14), which is, of course, also dependent on the underlying learned prototype distribution and ordering. Thus, a consistent ordering of the labels is obtained in FLSOM.

Hence, the evaluation of the similarities between the prototype label vectors yields suggestions for the similarity of classes, i.e. similar classes are represented by prototypes in a local spatial area of the SOM lattice. In case of overlapping class distributions this topographic class processing leads to prototypes with unclear decision, located between prototypes with clear vote. Further, if classes are not distinguish-able, there will exist prototypes responsive to those data which have class label vectors containing approximately the same degree of class membership for the respective classes. In this way FLSOM may be used for class similarity detection.

3 Data preprocessing by Wavelet Analysis

The analysis of functional data, is a common task in bioinformatics. Spectral data as obtained from mass spectrometric measurements in clinical proteomics are such functional data leading to new challenges for an appropriate analysis. Here we focus on the determination of classification models for such data. In general the available approaches for this task initially transform the spectra into a vector space followed by training a classifier. Hereby the functional nature of the data is typically lost, which may lead to suboptimal classifier models. Taking this into account a wavelet encoding is applied onto the spectral data leading to a compact *functional* representation. Thus, a functional representation of the data with respect to the used metric and a weighting or pruning of especially (priory not known) irrelevant function parts of the inputs, would be desirable. Further feature selection is applied based on a statistical pre-analysis of the data. Hereby a discriminative data representation is necessary. The extraction of such discriminant features is crucial for spectral data and typically done by a parametric peak picking procedure. This peak picking is often focus of criticism because peaks may be insufficiently detected and the functional nature of the data is partially lost. To avoid this difficulties we focus on the approach as given in [16] and apply a wavelet encoding to the spectral data to get discriminative features. Thereby the obtained wavelet coefficients are sufficient to reconstruct the signal, still containing all relevant information of the spectra. However this better discriminating set of features is typically more complex and hence a robust approach to determine the desired classification model is needed.

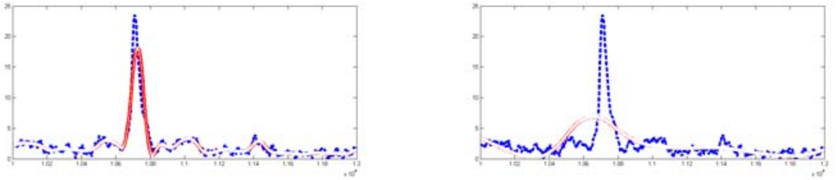
The classification of mass spectra involves in general the two steps peak picking to locate and quantify positions of peaks within the spectrum and feature extraction from the obtained peak list. In the first step a number of procedures as baseline correction, optional denoising, noise estimation and normalization must be applied[17,18]. Upon these prepared spectra the peaks have to be identified by scanning all local maxima and the associated peak endpoints followed by a S/N thresholding such that one obtains the desired peak list.

The procedure of baseline correction and recalibration (alignment) of multiple spectra is standard, and has been done using ClinProTools in this paper (details in [17])[1]. Here we propose an alternative feature extraction procedure preserving all (potentially small) peaks containing relevant information by use of the discrete wavelet transformation (DWT). The feature extraction has been done by Wavelet analysis using the Matlab Wavelet-Toolbox[2], due to the local analysis property of wavelet analysis the features can still be related back to original mass position in the spectral data which is essential for further biomarker analysis. In a first step a feature selection procedure using the Kolmogorov-Smirnoff test (KS-test) was applied. Thereby the test was used to identify features which show a significant ($p < 0.01$) discrimination between the two groups (cancer,control). This is done in accordance to [19] were also a generation to a multiclass experiment is given.

3.1 Feature Extraction and Denoising by Bi-orthogonal Discrete Wavelet Transform

Wavelets have been developed as powerful tools [20,21] used for noise removal and data compression. The discrete version of the continuous wavelet transform leads to the concept of a multiresolution analysis (MRA). This allows a fast and stable wavelet analysis and synthesis. The analysis becomes more precise if the wavelet shape is adapted to the signal to be analyzed. For this reason one can apply the so called bi-orthogonal wavelet transform[22] which uses two pairs of scaling and wavelet functions. One is for the decomposition/analysis and the other one for reconstruction/synthesis. The advantage of the bi-orthogonal wavelet transform is the higher degree of freedom for the shape of the scaling and wavelet function. In our analysis such a smooth synthesis pair was chosen to avoid artifacts. It can be expected that a signal in the time domain can be represented by a small number of a relatively large set of coefficients from the wavelet domain. The

[1] Biomarker software available at http://www.bdal.de
[2] The Matlab Wavelet-Toolbox can be obtained from www.mathworks.com



(a) Wavelet reconstruction $L = 6$ (b) Wavelet reconstruction$L = 8$

Fig. 3. Wavelet reconstruction of the spectra with $L = 6, 8$, x measurement positions, y-arbitrary unit. The original signal is plotted with the interrupted line (blue) and the reconstruction with the solid with a white band inside. One observes that a wavelet analysis with $L = 8$ (and 7 as well) is to rough to approximate the sharp peaks.

spectra are reconstructed in dependence of a certain approximation level L of the MRA which can be considered as a hard-thresholding. The denoised spectrum looks similar to the reconstruction as depicted in Figure 3. The starting point for an argumentation is the simplest example of a MRA which can be defined by the characteristic function $\chi_{[0,1)}$. The corresponding wavelet is the so-called *Haar* wavelet. Assume that the denoised spectrum $f \in L_2(\mathbb{R})$ has a peak with endpoints $2^j k$ and $2^j(k+1)$, the integral of the peak can be written as

$$\int_{2^j k}^{2^j(k+1)} f(t)dt = \int_{\mathbb{R}} f(t)\chi_{[2^j k, 2^j(k+1))}(t)dt$$

Obviously the right hand side is the Haar DWT scaling coefficient $c_{j,k} = \langle f, \psi_{j,k} \rangle$ at scale $a = 2^j$ and translation $b = 2^j k$. One obtains approximation- and detail-coefficients [22]. The approximation coefficients describe a generalized peak list of the denoised spectrum encoding primal spectral information and depend on the level L which is determined with respect to the measurement procedure. For linear MALDI-TOF spectra a device resolution of $500 - 800Da$ can be expected. This implies limits to the minimal peak width in the spectrum and hence, the reconstruction level of the Wavelet-Analysis should be able to model corresponding peaks. A level $L = 4$ is typically sufficient for a linear measured spectrum with ≈ 20000 measurement points, a level of $L = 6$ has been used for the data with ≈ 65000 measurement points. (see Figure 3). The level L can be automatically determined by considering expected peak width in Da and the reconstruction capabilities of wavelet analysis at a given level. Alternatively multiple levels can be tried and a standard peak picking approach can be applied on both, the original and the reconstructed spectrum. If the obtained peak lists are sufficiently similar, by means, that at least peaks with good S/N values in the original spectrum are sufficiently recovered in the reconstruction the taken level can be considered as acceptable for the experiment. Applying this procedure including the KS-test on the spectra with an initial number of ≈ 65000 measurement points per spectrum one obtains 1036 wavelet coefficients used as representative features per spectrum, still allowing a reliable functional representation of the data. An application of the KS-Test still keeps 199 coefficients for the final analysis. The

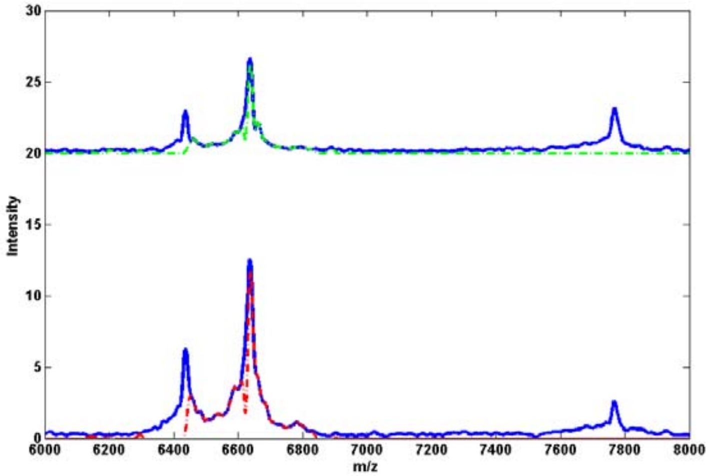


Fig. 4. Reconstructed region of some spectra of the two classes top control, bottom cancer. The straight lines indicate the reconstruction of the spectra by use of the chosen Wavlet approximation level upon approximation coefficients. The dotted line indicates the same reconstruction but with pruned coefficients which did not pass the statistical test. One observes that regions which are clearly non informative (near to the noise spectrum) are removed but also non-discriminating peaks (by means of the statistical test) are pruned.

effect of the KS-Test selection on the wavelet encoded spectra is shown in Figure 4.

4 Classification dependent metric adaptation - relevance learning

As mentioned above, the general dissimilarity measure $\xi\left(\mathbf{v},\mathbf{w}_{\mathbf{r}}\right)$ for the data space V is often chosen as squared Euclidean metric such that the derivative $\frac{\partial\xi(\mathbf{v},\mathbf{w}_{\mathbf{r}})}{\partial\mathbf{w}_{\mathbf{r}}}$ simply becomes $-2(\mathbf{v}-\mathbf{w}_{\mathbf{r}})$. Yet, other measures also can be applied, for example correlation measures [23]. However, more flexibility is obtained if $\xi\left(\mathbf{v},\mathbf{w}_{\mathbf{r}}\right)$ is parametrized and the parameters are also subject of optimization according to the given classification task [24],[25].

Generally, we consider a parametrized distance measure $\xi^{\lambda}(\mathbf{v},\mathbf{w})$ with a parameter vector $\lambda=(\lambda_1,\ldots,\lambda_M)$ with $\lambda_i\geq 0$ and normalization $\sum_i\lambda_i=1$. Then classification task depending parameter optimization is achieved by gradient descent, i.e. by consideration of $\frac{\partial E_{\mathrm{FLSOM}}}{\partial\lambda_l}$. Formal derivation yields

$$\frac{\partial E_{\mathrm{FLSOM}}}{\partial\lambda_l}=(1-\beta)\frac{\partial E_{\mathrm{SOM}}}{\partial\lambda_l}+\beta\frac{\partial E_{\mathrm{FL}}}{\partial\lambda_l}\tag{15}$$

with

$$\frac{\partial E_{\mathrm{SOM}}}{\partial \lambda_l} = \frac{1}{2}\sum_{\mathbf{r}} \int P(\mathbf{v}) \cdot \delta_{\mathbf{r}}^{s(\mathbf{v})} \sum_{\mathbf{r}'} h_\sigma(\mathbf{r},\mathbf{r}') \cdot \frac{\partial \xi^\lambda(\mathbf{v},\mathbf{w}_{\mathbf{r}})}{\partial \lambda_l} d\mathbf{v} \tag{16}$$

and

$$\frac{\partial E_{\mathrm{FL}}}{\partial \lambda_l} = -\frac{1}{4\gamma^2}\sum_{\mathbf{r}} \int P(\mathbf{v}) \cdot g_\gamma(\mathbf{v},\mathbf{w}_{\mathbf{r}}) \cdot \frac{\partial \xi^\lambda(\mathbf{v},\mathbf{w}_{\mathbf{r}})}{\partial \lambda_l} \cdot \vartheta\left(\mathbf{x}\left(\mathbf{v}\right),\mathbf{y}_{\mathbf{r}}\right) d\mathbf{v} \tag{17}$$

for the respective parameter adaptation.

4.1 Scaled Euclidean Metric

In case of $\xi^\lambda(\mathbf{v},\mathbf{w})$ being the *scaled* squared Euclidean metric

$$\xi^\lambda(\mathbf{v},\mathbf{w}) = \sum_i \lambda_i (v_i - w_i)^2 \tag{18}$$

(with $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$) the derivative becomes $\frac{\partial \xi(\mathbf{v},\mathbf{w}_i)}{\partial \mathbf{w}_i} = -2 \cdot \mathbf{\Lambda} \cdot (\mathbf{v} - \mathbf{w}_i)$ with $\mathbf{\Lambda}$ is a diagonal matrix and its i-th diagonal entry is λ_i. The corresponding learning rule for the metric parameter λ_l has the form

$$\triangle\, \lambda_l = -\epsilon_\lambda \frac{1-\beta}{2} \sum_{\mathbf{r}} h_\sigma(s(\mathbf{v}),\mathbf{r}) \cdot (v_l - (w_{\mathbf{r}})_l)^2 \tag{19}$$

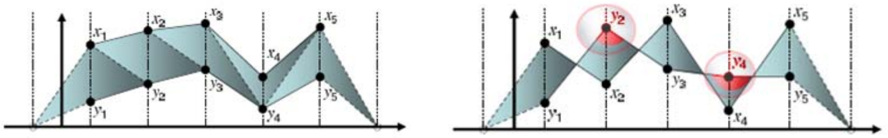
$$+\epsilon_\lambda \frac{\beta}{4\gamma^2} \sum_{\mathbf{r}} g_\gamma(\mathbf{v},\mathbf{w}_{\mathbf{r}}) \cdot (v_l - (w_{\mathbf{r}})_l)^2 \cdot \vartheta\left(\mathbf{x}\left(\mathbf{v}\right),\mathbf{y}_{\mathbf{r}}\right) \tag{20}$$

(subscript l denoting the component l of a vector) with learning rate $\epsilon_\lambda > 0$. This update is followed by normalization to ensure $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$.

The parameter optimization of the *scaled* squared Euclidean metric allows a useful interpretation. The parameter λ_i weight the dimensions of the data space. Hence, optimization of these parameters in dependence on the classification problem leads to a ranking of the input dimensions according to their classification decision relevance. Therefore, metric parameter adaptation of the scaled Euclidean metric is called *relevance learning*. In case of zero-valued λ_i this can also be seen as feature selection.

4.2 Generalized L^p-Norm

As pointed out before, the similarity measure $d^\lambda\left(\mathbf{v},\mathbf{w}\right)$ is only required to be differentiable with respect to λ and $\mathbf{w}$. The triangle inequality has not to be fulfilled necessarily. This leads to a great freedom in the choice of suitable measures and allows the usage of non-standard metrics in a natural way. We now review the functional metric as given in [26], the obtained derivations can be plugged into the above equations leading to FLSOM with a functional metric, whereby the data are functions represented by vectors and, hence, the vector dimensions are spatially correlated.



(a) Two functions: Euclidean = L^p-norm (b) Two functions: Euclidean $\neq$ L^p-norm

Fig. 5. Schematical ilustration of the L^p-norm. The first plot (a) indicates the case where the distance between two functions is equal considering Euclidean or L^p-norm. In the plot (b) parts of the functions are interchanging (crossings) thereby the distances using Euclidean distance is still the same as within plot (a) but for the L^p-norm another distance is obtained which indeed gives a more realistic measure of the distances of the two functions.

Common vector processing does not take the spatial order of the coordinates into account. As a consequence, the functional aspect of spectral data is lost. For proteom spectra the order of signal features (peaks) is due to the nature of the underlying biological samples and the measurement procedure. The masses of measured chemical compounds are given ascending and peaks encoding chemical structures with a higher mass follows chemical structures with lower masses. In addition multiple peaks with different masses may encode parts of the same chemical structure and hence are correlated.

Lee proposed a distance measure taking the functional structure into account by involving the previous and next values of x_i in te i-th term of the sum, instead of x_i alone. Assuming a constant sampling period τ, the proposed norm is:

$$\mathcal{L}_p^{fc}\left(\mathbf{v}\right) = \left(\sum_{k=1}^{D} \left(A_k\left(\mathbf{v}\right) + B_k\left(\mathbf{v}\right)\right)^p \right)^{\frac{1}{p}} \tag{21}$$

with

$$A_k\left(\mathbf{v}\right) = \begin{cases} \frac{\tau}{2}|v_k| & \text{if } 0 \leq v_k v_{k-1} \\ \frac{\tau}{2}\frac{v_k^2}{|v_k|+|v_{k-1}|} & \text{if } 0 > v_k v_{k-1} \end{cases} \quad B_k\left(\mathbf{v}\right) = \begin{cases} \frac{\tau}{2}|v_k| & \text{if } 0 \leq v_k v_{k+1} \\ \frac{\tau}{2}\frac{v_k^2}{|v_k|+|v_{k+1}|} & \text{if } 0 > v_k v_{k+1} \end{cases} \tag{22}$$

are respectively of the triangles on the left and right sides of x_i. Just as for L_p, the value of p is assumed to be a positive integer. At the left and right ends of the sequence, x_0 and x_D are assumed to be equal to zero. The concept of the L^p-norm is shown in Figure 5.

The derivatives for the functional metric taking $p = 2$ are given in [26]. Now we consider the scaled functional norm where each dimension v_i is scaled by a parameter $\lambda_i > 0$ $\lambda_i \in (0, 1]$ and $\sum_i \lambda_i = 1$. Then the scaled functional norm is:

$$\mathcal{L}_p^{fc}\left(\lambda\mathbf{v}\right) = \left(\sum_{k=1}^{D} \left(A_k\left(\lambda\mathbf{v}\right) + B_k\left(\lambda\mathbf{v}\right)\right)^p \right)^{\frac{1}{p}} \tag{23}$$

with

$$A_k\left(\lambda\mathbf{v}\right)=\begin{cases}\frac{\tau}{2}\lambda_k\left|v_k\right| \text{ if } 0\le v_kv_{k-1}\\ \frac{\tau}{2}\frac{\lambda_k^2v_k^2}{\lambda_k|v_k|+\lambda_{k-1}|v_{k-1}|} \text{ else}\end{cases}\qquad B_k\left(\lambda\mathbf{v}\right)=\begin{cases}\frac{\tau}{2}\lambda_k\left|v_k\right| \text{ if } 0\le v_kv_{k+1}\\ \frac{\tau}{2}\frac{\lambda_k^2v_k^2}{\lambda_k|v_k|+\lambda_{k+1}|v_{k+1}|}\text{else}\end{cases}\tag{24}$$

The prototype update for $p=2$ changes to:

$$\frac{\partial\delta_2^2\left(\mathbf{x},\mathbf{y},\lambda\right)}{\partial x_k}=\frac{\tau^2}{2}\left(2-U_{k-1}-U_{k+1}\right)\left(V_{k-1}+V_{k+1}\right)\triangle_k\tag{25}$$

with

$$U_{k-1}=\begin{cases}0 \text{ if } 0\le\triangle_k\triangle_{k-1}\\ \left(\frac{\lambda_{k-1}\triangle_{k-1}}{\lambda_k|\triangle_k|+\lambda_{k-1}|\triangle_{k-1}|}\right)^2 \text{ else}\end{cases}\quad ,\ U_{k+1}=\begin{cases}0 \text{ if } 0\le\triangle_k\triangle_{k+1}\\ \left(\frac{\lambda_{k+1}\triangle_{k+1}}{\lambda_k|\triangle_k|+\lambda_{k+1}|\triangle_{k+1}|}\right)^2 \text{ else}\end{cases}$$

$$V_{k-1}=\begin{cases}\lambda_k \text{ if } 0\le\triangle_k\triangle_{k-1}\\ \frac{\lambda_k|\triangle_k|}{\lambda_k|\triangle_k|+\lambda_{k-1}|\triangle_{k-1}|} \text{ else}\end{cases}\quad ,\ V_{k+1}=\begin{cases}1\lambda_k \text{ if } 0\le\triangle_k\triangle_{k+1}\\ \frac{\lambda_k|\triangle_k|}{\lambda_k|\triangle_k|+\lambda_{k+1}|\triangle_{k+1}|} \text{ else}\end{cases}$$

and $\triangle_k=x_k-y_k$ For the λ-update one observes:

$$\begin{aligned}\frac{\partial\mathcal{L}_p^{fc}\left(\lambda\mathbf{v}\right)}{\partial\lambda_k} &= \frac{\partial\left(\sum_{k=1}^{D}\left(A_k\left(\lambda\mathbf{v}\right)+B_k\left(\lambda\mathbf{v}\right)\right)^p\right)^{\frac{1}{p}}}{\partial\lambda_k}\\ &= p\left(\sum_{k=1}^{D}\left(A_{k-1}\left(\lambda\mathbf{v}\right)+A_{k+1}\left(\lambda\mathbf{v}\right)\right)^p\right)^{\frac{1-p}{p}}\frac{\partial\left[\sum_{k=1}^{D}\left(A_k\left(\lambda\mathbf{v}\right)+B_k\left(\lambda\mathbf{v}\right)\right)^p\right]}{\partial\lambda_k}\\ &= C_p\frac{\partial\left[\sum_{k=1}^{D}\left(A_k\left(\lambda\mathbf{v}\right)+B_k\left(\lambda\mathbf{v}\right)\right)^p\right]}{\partial\lambda_k}\\ &= C_p\frac{\sum_{k=1}^{D}\partial\left[\left(A_k\left(\lambda\mathbf{v}\right)+B_k\left(\lambda\mathbf{v}\right)\right)^p\right]}{\partial\lambda_k}\\ &= C_p\frac{\partial\left[\left(A_{k-1}\left(\lambda\mathbf{v}\right)+B_{k-1}\left(\lambda\mathbf{v}\right)\right)^p+\left(A_k\left(\lambda\mathbf{v}\right)+B_k\left(\lambda\mathbf{v}\right)\right)^p+\left(A_{k+1}\left(\lambda\mathbf{v}\right)+B_{k+1}\left(\lambda\mathbf{v}\right)\right)^p\right]}{\partial\lambda_k}\\ &= C_p\left(c_p^{k-1}\frac{\partial\left[A_{k-1}\left(\lambda\mathbf{v}\right)+B_{k-1}\left(\lambda\mathbf{v}\right)\right]}{\partial\lambda_k}+c_p^{k}\frac{\partial\left[A_k\left(\lambda\mathbf{v}\right)+B_k\left(\lambda\mathbf{v}\right)\right]}{\partial\lambda_k}+*\right)\\ * = c_p^{k+1}&\frac{\partial\left[A_{k+1}\left(\lambda\mathbf{v}\right)+B_{k+1}\left(\lambda\mathbf{v}\right)\right]}{\partial\lambda_k}\end{aligned}$$

with the following expressions

$$\begin{aligned}c_p^j &= p\cdot\left(A_j\left(\lambda\mathbf{v}\right)+B_j\left(\lambda\mathbf{v}\right)\right)^{p-1}\\ &= p\cdot\left(\begin{array}{c}\begin{cases}\frac{\tau}{2}\lambda_j\left|v_j\right| & \text{if } 0\le v_jv_{j-1}\\ \frac{\tau}{2}\frac{\lambda_j^2v_j^2}{\lambda_j|v_j|+\lambda_{j-1}|v_{j-1}|} & \text{if } 0> v_jv_{j-1}\end{cases}\\ +\begin{cases}\frac{\tau}{2}\lambda_j\left|v_j\right| & \text{if } 0\le v_jv_{j+1}\\ \frac{\tau}{2}\frac{\lambda_j^2v_j^2}{\lambda_j|v_j|+\lambda_{j+1}|v_{j+1}|} & \text{if } 0> v_jv_{j+1}\end{cases}\end{array}\right)^{p-1}\end{aligned}$$

putting all together and with some minor mathematical transformations one obtains:

$$\frac{\partial \mathcal{L}_p^{fc}(\lambda \mathbf{v})}{\partial \lambda_k} = C_p \begin{cases} 0 + c_p^k \left(\frac{\tau}{2} \left| v_k \right| \right) & \text{if } 0 \leq v_{k-1} v_k \\ \frac{1}{2}\tau \frac{\lambda_k^2 c_p^k v_k^2 |v_k| - c_p^{k-1} |v_k| v_{k-1}^2 \lambda_{k-1}^2 + 2\lambda_k c_p^k v_k^2 |v_{k-1}| \lambda_{k-1}}{(\lambda_k |v_k| + |v_{k-1}| \lambda_{k-1})^2} & \text{if } 0 > v_{k-1} v_k \end{cases}$$
$$+ C_p \begin{cases} c_p^k \left(\frac{\tau}{2} \left| v_k \right| \right) + 0 & \text{if } 0 \leq v_{k+1} v_k \\ \frac{1}{2}\tau \frac{\lambda_k^2 c_p^k v_k^2 |v_k| - c_p^{k+1} |v_k| v_{k+1}^2 \lambda_{k+1}^2 + 2\lambda_k c_p^k v_k^2 |v_{k+1}| \lambda_{k+1}}{(\lambda_k |v_k| + |v_{k+1}| \lambda_{k+1})^2} & \text{if } 0 > v_{k+1} v_k \end{cases}$$

Using this parametrization one can emphasize/neglect different parts of the function for classification. This distance measure can be put into FLSOM as shown above and has been applied subsequently in the analysis of clinical proteom spectra.

5 Generic determination of model confidence

Classical statistical methods which are used to determine classification models are often inappropriate or not applicable in case of modern high-dimensional and high-throughput data sets. Recent achievements in machine learning such as multiple extensions on support vector machines [27,28,29] or modern prototype networks [**?**,25,24,30] directly aim on the processing of such data in a computational efficient way. A typical drawback of these methods is the lack of useful measurements of confidence in the obtained predictions. Thereby a judment of the model generation in general by means of generalization ability as well as with respect to the classification decision would be desirable. In [31] the method of hedging predictions or so called conformal predictors is introduced which calculates confidence and so called credibility values for arbitrary types of classifier as long as they fit to a specific framework. Gammerman et al. are focusing on Support Vector Machines in their paper [31], here we will use this method in the context of prototype based classifiers.

5.1 Conformal prediction

The concept of conformal prediction was introduced in [31,32,33] and is a generic method to determine confidence and so call credibility values for outputs of e.g. classification or regression models. Here we will focus on the classification part in conformal prediction to obtain measures, which are similar to p-values indicating the reliability of an obtained classification result. This is very important in the clinical domain because a plain classification decision is in general inappropriate if one can not judge the safety of the result [18]. Thereby in principle any kind of classifier can be improved by conformity measures as long as the model calculates beside of class labels a further measure indicating some kind of strangeness of the result with respect to the obtained model. One possible measure is the distance to the decision boundary as mentioned in [31]. Taking this fact into account the sample margin for prototype based networks can be used in general. Conformal prediction tries to estimate a distribution of these *non-conformity* (strangeness) measures. Upon this distribution the obtained non-conformity value for a classification result can be statistically evaluated leading to two measure the confidence of the result, which is similar to a p-value and an credibilty value which gives some measure of the validity of the model with respect to the current sample e.g. some kind of outlier test. Thereby the results of conformal prediction remain automatically valid under the randomness assumption [31,32,33]. Thereby its assumed, that the objects and their labels are generated independently from the same probability distribution. This appears to be a strong assumption but in fact it is a much weaker assumption than assuming a parametric statistical model. Thereby conformal predictors never overrate the accuracy and reliability of their predictions [31,32,33]. As conformal predictors are provably valid, efficiency with respect to computational performance as well as with respect to the effort to extend a classifier to an conformal predictor, are the only things which we need to worry about. When the stochastic mechanism significantly deviates from the model, conformal predictors remain valid but their efficieny inevitably suffers [31,32,33]. Details on the derivation including proofs are given in [31,32,33], we will not go into details here but only sketch the specific non-conformity measure used to determine conformal prediction by FLSOM. Thereby we use the transductive variant of conformal prediction as presented in [31]. To obtain a effective non-conformity score for each point matched against a determined model and to estimate the distribution of these values one has to consider a smooth measure which gives information about the conformity of a classification decision. For prototype based networks one natural measure of non-conformity ($C(\mathbf{v_i}, c_i)$ for a given sample $\mathbf{v_i}$ and a given (crisp) labeling c_i is the sample margin as the distance of the data point to the closest prototype with the same label (+) normalized by the distance of this item to the closest prototype with an alternative labeling (-):

$$C(\mathbf{v_i}, c_i) = d^{+}_{\min,\lambda}(\mathbf{w_r}, \mathbf{v_i}) / d^{-}_{\min,\lambda}(\mathbf{w_r}, \mathbf{v_i}) \tag{26}$$

Thereby the classifier decision is considered to be safe if the obtained non-conformity score is small - by means of a small distance of the datapoint to its closest prototype with the same labeling. For FLSOM we do not necessary have a crisp labeling of the prototypes or data points and hence the non-conformity score has to be adapted slightly. For FLSOM we have two terms come into play. One the one hand side the quantization error and on the other side the label error. If both values are small the item can be considered to be represented correct by the FLSOM model. Hence we suggest to use the following conformity measure for sample point $\mathbf{v_i}$ with a given vector labeling $\mathbf{c_i}$

$$C(\mathbf{v_i}, c_i) = 0.5 * (\textstyle\sum_r^k d_\lambda(\mathbf{w_r}, \mathbf{v_i}) + \sum_r^k d(\mathbf{y_r}, \mathbf{c_i})) \tag{27}$$

Thereby the summation could be also limited to the winner neuron or by considering only the k nearest prototypes within the FLSOM grid structure. To make quantization and label error more comparable distances and label errors $e(y)$ have been normalized to $\sum_r d_\lambda(\mathbf{w_r}, \mathbf{v_i}) = 1$ and $\sum_r d(\mathbf{y_r}, \mathbf{c_i}) = 1$.

6 Clinical data

Serum protein profiling is a promising approach for classification of cancer versus non-cancer samples. The data used in this paper are taken from a colorectal cancer (CRC) study and patients from healthy individuals [34].

The standardized circumstances for sample collection and the data set are described in detail in [34]. Here it should be mentioned only that for each profile a mass spectrum is obtained within a mass-to-charge-ratio of 1000 to 11000Da. Two sample spectra are depicted in 1. The data have been preprocessed as explained before using the approach published in [16]. Thereby the spectra are encoded by $\approx$ 200 wavelet-coefficients which leads to a data reduction of $\approx$ 99.9% using the rawdata and is twice the number of peaks as obtained by the standard peak picking approach as proposed in [17]. Thereby the preprocessing stage has to be included in the crossvalidation procedure to avoid overfitting, for the considered data set it could be observed that the discriminating wavelet coefficients (with respect to the ks-test) at $p \leq 0.01$ remain the same in a $5-$fold cross validation. The wavelet method was used as mentioned in the previous section with $L = 6$.

The data set consist of 123 samples whereby 73 are taken from patients suffering from colorectal cancer and the remaining 50 samples are taken from a matched healthy control group[3]. Colorectal cancer is among the most common malignancies and remains a leading cause of cancer-related morbidity and mortality. It is well recognized that CRC arises from a multistep sequence of genetic alerations that result in the transformation of normal mucosa to aprecursor adenoma and ultimately to carcinoma. Given the natural history of CRC, early diagnosis appears to be the most appropriate tool to reduce disease-related morality. Currently, there is no early diagnostic test with sufficient diagnostic quality, which can be used as a routine screening tool. Therefore, there is a need for new biomarkers for colorectal cancer that can improve early diagnosis, monitoring of disease progression and therapeutic response and detect disease recurrence. Furthermore, these markers may give indications for targets for novel therapeutic strategies. In addition to potential markers validated by further post analysis on identified masses, generic classification models with high validity maybe of value as well.

7 Experiments

The available data set for investigation consists of overall 123 proteomic expression profiles generated by MALDI-TOF mass spectrometry (MS). In [34] an experimental setting was shown focusing on Fishers Linear Discriminant Analysis (LDA) combined with a principal component (PCA) approach to reduce the dimensionality of the underlying problem with promising results. Thereby the peak picking was avoided by a simple binning approach and the PCA was used

[3] In the article of [34] some additional selections with respect to the cancer group has been done - here work work on the whole data set

to get a sufficient reduction of the dimensionality of the feature space. PCA is focusing on maximal explained variance in the data [35], this however is typically no good criteria in the analysis of clinical proteom spectra because the group separations are in general not indicated by large variations in the intensities [18]. Hence a PCA approach will in general fail to give sufficient results. Although the PCA got sufficient results in [34] a more generic approach for the analysis of clinical proteom spectra taken form MALDI-MS is desirable.

Such an approach is to determine the decision plane with respect to the known class label information which is pointed out by multiple authors e.g. [36,37]. Taking these into account we focus on a supervised data analysis and reduce the dimensionality of the data by use of a problem specific wavelet analysis combined with a statistical selection criterium. Thereby we avoid statistical assumptions with respect to the underlying data sets, but take only measurement specific knowledge into account.

Hence we have a 199-dimensional space of wavelet coefficients and we use multiple algorithms and metrics to determine classification models. Thereby we focus of the presented FLSOM algorithm which beside a classification model leads to a (under some constraints) topological preserving visualization of these high dimensional data.

To be comparable with the study in [34], we trained in a first investigation a FLSOM with data only of the groups A and B. Thereby we used a 7×3 FLSOM lattice, the size of which is determined by a growing SOM (GSOM) [38]. The balancing parameter was declared as $\beta = 0.85$, which emphasizes the classification term in (12) but prevents instabilities for higher values [30]. To be generalizing and regularizing we used the inherent regularization abilities of SOMs by non-vanishing neighborhood range σ in the neighborhood function h_σ in (3). To do so and to prevent violations in topology preservation the remaining value was chosen as $\sigma = 0.5$ [39].

8 Results

A typical FLSOM obtained from multiple runs is depicted in Figure 8. One observes a clear separation between cancer and control data. The overlapping region between the classes is rather small which is also supported by the relative good crossvalidation results for the linear classification models. For this data set the obtained FLSOM using different metrics are topological preserving. The FLSOM approaches obtained $\approx 86\%$ cross validation accuracy in a 5-fold crossvalidation, using scaled Euclidean metric which is a similar good accuracy as in [34]. Thereby in addition to the good classification accuracy a ranking of individual features as well as a planar visualization of the high dimensional data is obtained. The latter one allows for interpretation of similarities between sub groups of patients (see Figure 8)

The relevance parameters λ_i of the scaled Euclidean metric are adapted parallely. This leads to a ranking of the input dimensions according to their importance for classification. A typical relevance profile using scaled Euclidean

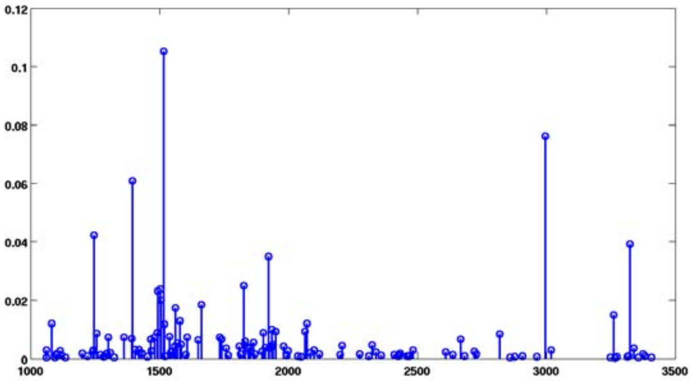


Fig. 6. Visualization of a typical relevance profile obtained by FLSOM using scaled Euclidean metric. Peaks with larger values indicate higher relevance with respect to the classification task. The x-axis indicates the relative mass position of the corresponding wavelet coefficient in the original spectrum. The y-axis is a relevance measure $\in [0, 1]$.

metric is depicted in Figure 6. The most important frequencies are indicated by straight arrows in Fig. 7, dashed arrows refer to further highly relevant frequencies. The depicted frequencies contribute substantially to classification accuracy and, therefore, are important for distinction of the classes.

A comparison of the FLSOM results using the different metrics and alternative algorithms is given in Table 1. Thereby it should be noted, that in [34] a part of the cancer class spectra has been removed from the model generation due to quality constraints, while in our analysis all spectra have been used. The lower three rows of the table contain results obtained on alternative data preparations, namely peaklists (CPT results) and the preparation as given in [34]. In [34] a leave one out (LOO) cross validation has been used to determine the generalization ability of the approach, LOO is a restriction which is typical for small data sets. LOO however has some drawbacks as pointed out in [40,41,42]. We used a 5-fold CV in accordance to the suggestions in [40] because the number of sample is not so small and they are reliable homogeneous per group as depicted in Figure 7.

One observes that the results are competitive with respect to other classifiers but it should be mentioned again that FLSOM is not focusing on classification but equally on visualization and interpretability of the given high dimensional data sets. In that way classification accuracy as well as a modeling of the data distribution is optimized. In average the different methods obtain a cross validation accuracy of $\approx 89\%$ using the presented generic preprocessing approach. Thereby the wavelet prepared data perform similar than a standardized peak picking approach with other parameters fixed. The approach in [34] obtained slightly better results in the LOO cross validation but is to much focused on

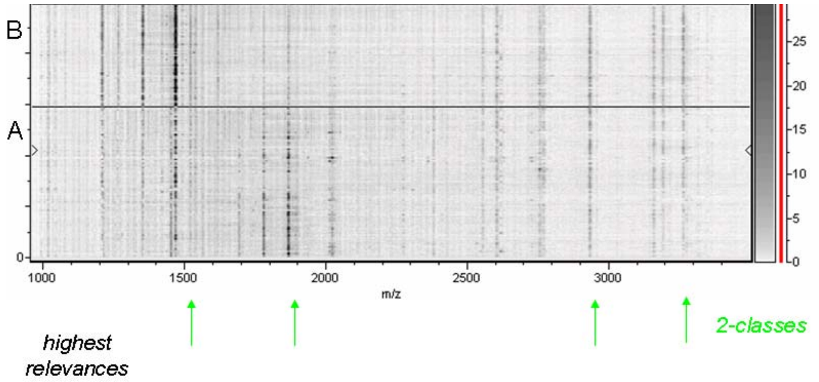


Fig. 7. A gel view of the two classes with the cancer class (region A) and control class (region B). The relevant mass positions are indicated by arrows (bottom) using the relevance profile of FLSOM with scaled Euclidean metric.

explained variance which can not be generalized to other clinical proteomics problems in general. Considering the cross validation results in Table 1 it can be observed, that similar results were obtained using the different metrics. However the metrics itself show different properties. The relevance profile of the scaled Euclidean metric indicates most important data features in a univariate interpretation whereas the generalized L^p norm takes local neighborhoods or correlations in the data space into account while keeping the functional nature of the MS spectra. Therefore also descents in the function and not just peaks as well as correlative effects can be interpreted as relevant features. In 1 results are shown using the functional metric has indicated alternative regions with similar separation capability. Thereby relatively small peaks are identified which, combined with the neighborhood are indeed informative. Characteristic for those regions identified in the considered data is, that not a single peak has been identified but a trace of a local biochemical pattern. Here the pattern typically consists of a peak with moderate intensities and small but not perfect differences between the two classes and a valley close to the peak with a quite clear (but also not perfect) missing of mass information for one class. This valley could not be identified as a peak by a peak picking procedure because the region has no peak characteristic, nevertheless it could be observed that for one class at this valley mass intensities has been measured whereas for the other class the intensities are zero or very low. Thereby this trace of information can be further analysed by e.g. LC/MS techniques to test if a potential useful pattern can be observed which in the current linear measurement has not been sufficiently resolved so far.

The respective learned data distribution using FLSOM with the L^p-norm is depicted in Fig.8 Each square represents a label vector $\mathbf{y_r}$ of a prototype $\mathbf{w_r}$. The position is according localization $\mathbf{r}$ in the grid. The hight of the bars reflects the fuzzy amount for the respective class as indicated above. These findings are in agreement with clinical expectations. We observe the fine conformity of the

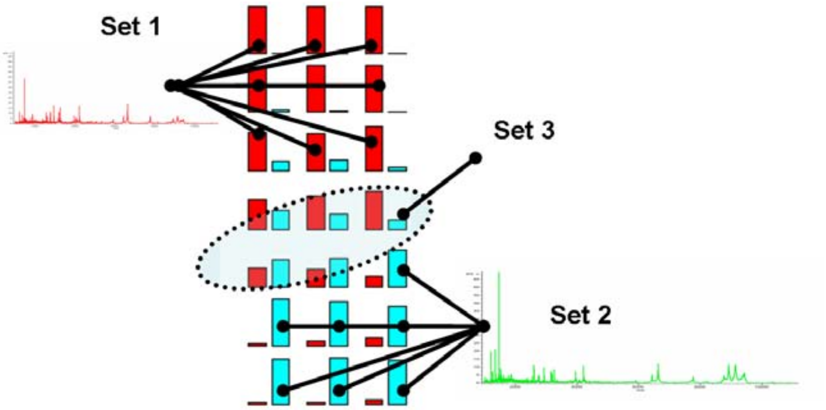


Fig. 8. Visualization of the FLSOM using the L^p metric. The FLSOM consists of 7×3 cells with each cell containing two bars indicating a fuzzy labeling. The first bar is responsible for the cancer class, while the second for control. A high bar for cancer indicates that spectra which are mapped to the considered cell are more likely to belong to cancer than to control. A clear separation of the two classes with a small overlap region can be observed. For each spectrum in the data set an associated cell on the grid by the SOM mapping can be identified. A raw analysis shows three sets of spectra. Set $1, 2$ contains quit homogenous spectra of the corresponding classes while the spectra in set 3 show multiple inhomogenities e.g. some of the cancer spectra show a bad S/N ratio for peaks and are in overall more noisy. There are also some spectra which show strong fluctuation in the intensities. Considering the mapping as well as the fuzzy label of the corresponding map a specific clustering of the high dimensional data is obtained. In case of multiple classes this further leads to a similarity highlighting of the different classes.

detected class similarities with the clinical expectations. Hence, FLSOM successfully discovered the underlying class structure. Initial results using the conformal prediction approach are promising the conformal prediction on the test data sets give similar accuracy than with the standard classifiers but in addition for each datapoint a confidence and credibility measure becomes available which allows a judgment of the classification decision for each single patient in a statistical manner.

9 Conclusions

We presented a specific pre-processing for mass spectrometric data analysis combined with an extension of the SOM for supervised classification tasks, which takes classification task explicitly into account for prototype adaptation during the gradient descent based adaptation process. The presented processing of the spectra aims on a natural compact encoding of the signals by means of a

Method	Rec.	CV - 5 fold	CP/Confidence
FLSOM-EUC	89.62%	86.12%	89.4%/0.88
FLSOM-L^p	89.23%	86.17%	n.a.
FLSOM-M	83.74%	87.94%	n.a.
SRNG-EUC	100%	90.24%	86.9%/0.87
SVM-Linear	96.75%	89.43%	n.a.
SVM-kNN (CPT)-LOO	96.58%	92.52%	n.a.
SVM-kNN (CPT)-5CV	96.58%	87.84%	n.a.
LDA+PCA -LOO	92.9%	92.6%	n.a.

Table 1. Recognition and cross validation accuracies for FLSOM using different similarity measures in comparison to alternative standard approaches. The results for LDA/PCA are taken from the article [34]. Thereby it should be noted, that in [34] a part of the cancer class spectra has been removed from the model generation due to quality constraints, while in our analysis all spectra have been used. The lower three rows of the table contain results obtained on alternative data preparations, namely peaklists (CPT results) and the preparation as given in [34]. The approach available in CPT with SVM+kNN first determines a ranking of the peaks by interpretation of the weight vector of a linear SVM. In a second step a kNN classifier is trained on the best peaks. The last column gives some results for the conformal prediction approach.

functional representation, while the classification model is especially suited to deal with high dimensional sparse data and allows strong regularizations to reduce overfitting effects. Thereby, each prototype dynamically adapt its assigned class label depending on the balancing between clustering and classification in the FLSOM model. In this way the statistical as well as label properties of the data influence prototype positions and fuzzy label learning. The visualization abilities of SOMs based on the topology preservation property of unsupervised SOMs then can be used for visual inspection of the class labels of the prototypes which may allow a better understanding of the underlying classification decision scheme. Further, the FLSOM is able to detect class similarities. In an initial setup the presented scenario has been embedded into a conformal prediction approach which allows the determination of clinical relevant confidence measures. Thereby the extension of conformal prediction for multiple types of prototype based classifiers has been presented. The FLSOM has been applied to classification of mass spectrometric data (profiles) of cancer disease and controls. Beside a comparable classification accuracy the model automatically discovered the class similarities in good agreement to clinical expectation. This allows a more specific interpretation of the classification models. Thus, FLSOM can be used not only for classification and visualization but also for detection of class dependencies. ACKNOWLEDGMENT: The author is grateful to M. Kostrzewa and T. Elssner for useful discussions and support in interpretation of the results (both Bruker Daltonik Leipzig, Germany). Further I would like to thank Alexander Gammerman and Luo Zhiyuan for helpful discussions on Hedging predictions

(both Computer Learning Research Center (CLRC), Royal Holloway, University of Londong, UK).

References

1. Pusch, W., Flocco, M., Leung, S., Thiele, H., Kostrzewa, M.: Mass spectrometry-based clinical proteomics. Pharmacogenomics **4** (2003) 463–476
2. Fiedler, G., Baumann, S., Leichtle, A., Oltmann, A., Kase, J., Thiery, J., Ceglarek, U.: Standardized peptidome profiling of human urine by magnetic bead separation and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Clinical Chemistry **53** (2007) 421–428
3. Schäffeler, E., Zanger, U., Schwab, M., et al., M.E.: Magnetic bead based human plasma profiling discriminate acute lymphatic leukaemia from non-diseased samples. In: 52st ASMS Conference 2004. (2004) TPV 420
4. Schipper, R., loof, A., de Groot, J., Harthoorn, L., van Heerde, W., Dransfield, E.: Salivary protein/peptide profiling with seldi-tof-ms. Annals of the New York Academy of Science **1098** (2007) 498–503
5. Guerreiro, N., Gomez-Mancilla, B., Charmont, S.: Optimization and evaluation of seldi-tof mass spectrometry for protein profiling of cerebrospinal fluid. Proteome science **4** (2006) 7
6. Kohonen, T.: Self-Organizing Maps. Volume 30 of Springer Series in Information Sciences. Springer, Berlin, Heidelberg (1995) (2nd Ext. Ed. 1997).
7. Villmann, T., Der, R., Herrmann, M., Martinetz, T.: Topology Preservation in Self–Organizing Feature Maps: Exact Definition and Measurement. IEEE Transactions on Neural Networks **8** (1997) 256–266
8. Schleif, F.M., Elssner, T., Kostrzewa, M., Villmann, T., Hammer, B.: Analysis and visualization of proteomic data by fuzzy labeled self organizing maps. In: Proc. of CBMS 2006. (2006) 919–924
9. Bishop, C.: Pattern Recognition and Machine Learning. Springer Science+Business Media, LLC, New York, NY (2006)
10. Hecht-Nielsen, R.: Counterprogagation networks. Appl. Opt. **26** (1987) 4979–4984
11. Vuorimaa, P.: Fuzzy self-organizing map. Fuzzy Sets and Systems **66** (1994) 223–231
12. Erwin, E., Obermayer, K., Schulten, K.: Self-organizing maps: Ordering, convergence properties and energy functions. Biol. Cyb. **67** (1992) 47–55
13. Heskes, T.: Energy functions for self-organizing maps. In Oja, E., Kaski, S., eds.: Kohonen Maps. Elsevier, Amsterdam (1999) 303–316
14. Hastie, T., Stuetzle, W.: Principal curves. J. Am. Stat. Assn. **84** (1989) 502–516
15. Bauer, H.U., Pawelzik, K.R.: Quantifying the neighborhood preservation of Self-Organizing Feature Maps. IEEE Trans. on Neural Networks **3** (1992) 570–579
16. Schleif, F.M., Hammer, B., Villmann, T.: Supervised neural gas for functional data and its application to the analysis of clinical proteom spectra. In: Proc. of IWANN 2007. (2007) 1036–1044
17. Ketterlinus, R., Hsieh, S.Y., Teng, S.H., Lee, H., Pusch, W.: Fishing for biomarkers: analyzing mass spectrometry data with the new clinprotools software. Bio techniques **38** (2005) 37–40
18. Schleif, F.M.: Prototype based Machine Learning for Clinical Proteomics. PhD thesis, Technical University Clausthal, Technical University Clausthal, Clausthal-Zellerfeld, Germany (2006)

19. Waagen, D., Cassabaum, M., Scott, C., Schmitt, H.: Exploring alternative wavelet base selection techniques with application to high resolution radar classification. In: Proc. of the 6th Int. Conf. on Inf. Fusion (ISIF'03), IEEE Press (2003) 1078–1085
20. A.K. Louis, P. Maaß, A.R.: Wavelets: Theory and Applications. Wiley (1998)
21. Leung, A., Chau, F., Gao, J.: A review on applications of wavelet transform techniques in chemical analysis: 1989-1997. Chem. and Int. Lab. Sys. **43** (1998) 165–184(20)
22. Cohen, A., Daubechies, I., Feauveau, J.C.: Biorthogonal bases of compactly supported wavelets. Comm. Pure Appl. Math. **45** (1992) 485–560
23. Villmann, T., Strickert, M., Brüß, C., Schleif, F.M., Seiffert, U.: Visualization of fuzzy information in in fuzzy-classification for image sagmentation using MDS. In: Proc. of ESANN 2007. (2007) 103–108
24. Hammer, B., Villmann, T.: Generalized relevance learning vector quantization. Neural Netw. **15** (2002) 1059–1068
25. Hammer, B., Strickert, M., Villmann, T.: Supervised neural gas with general similarity measure. Neural Proc. Letters **21** (2005) 21–44
26. Lee, J., Verleysen, M.: Generalizations of the lp norm for time series and its application to self-organizing maps. In Cottrell, M., ed.: 5th Workshop on Self-Organizing Maps. Volume 1. (2005) 733–740
27. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
28. Platt, J.C.: Fast training of support vector machines using sequential minimal optimization. MIT Press, Cambridge, MA, USA (1999)
29. Keerthi, S.S., Lin, C.J.: Asymptotic behaviors of support vector machines with gaussian kernel. Neural Comp. **15** (2003) 1667–1689
30. Sato, A., Yamada, K.: Generalized learning vector quantization. In Touretzky, D.S., Mozer, M.C., Hasselmo, M.E., eds.: Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference. MIT Press, Cambridge, MA, USA (1996) 423–9
31. Villmann, T., Hammer, B., Schleif, F.M., Geweniger, T.: Fuzzy classification by fuzzy labeled neural gas. Neural Networks **19** (2006) 772–779
32. Gammerman, A., Vovk, V.: Hedging predictions in machine learning. The Computer Journal **50** (2007) 151–163
33. Gammerman, A., Vovk, V.: Discussions on hedging predictions in machine learning. The Computer Journal **50** (2007) 164–172
34. Gammerman, A., Vovk, V.: Rejoinder hedging predictions in machine learning. The Computer Journal **50** (2007) 173–177
35. de Noo, M., Deelder, A., van der Werff, M., zalp, A., Martens., B.: MALDI-TOF serum protein profiling for detection of breast cancer. Onkologie **29** (2006) 501–506
36. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, New York (2001)
37. Zhang, Z., Page, G., Zhang, H.: Fishing expedition - a supervised approach to extract patterns from a compendium of expression profiles. In Lin, S.M., Johnson, K.F., eds.: Methods of Microarray Data Analysis II, Kluwer Academic Publishers (2002) Papers from CAMDA 01.
38. Lee, Y., Lee, C.K.: Classification of multiple cancer types by multicategory support vector machines using gene expression data. Bioinformatics **19** (2003) 1132–1139
39. Villmann, T., Bauer, H.U., Villmann, T.: The GSOM-algorithm for growing hypercubical output spaces in self-organizing maps. In: Proceedings of WSOM'97, Workshop on Self-Organizing Maps, Espoo, Finland, June 4–6. Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland (1997) 286–291

40. Der, R., Herrmann, M.: Instabilities in self-organized feature maps with short neighborhood range. In Verleysen, M., ed.: Proc. ESANN'94, European Symp. on Artificial Neural Networks, Brussels, Belgium, D facto conference services (1994) 271–276
41. Molinaro, A., Simon, R., Pfeiffer, R.: Prediction error estimation: A comparison of resampling methods. Bioinformatics **21** (2005) 3301–3307
42. Kearns, M.J., Mansur, Y., Ng, A., Ron, D.: An experimental and theoretical comparison of model selection methods. Machine Learning **27** (1997) 7–50
43. Bartlett, P.L., Boucheron, S., Lugosi, G.: Model selection and error estimation. Machine Learning **48** (2002) 85–113