Classification of mass-spectrometric data in clinical proteomics using learning vector quantization methods

Thomas Villmann, Frank-Michael Schleif, Markus Kostrzewa, Axel Walch and Barbara Hammer

Submitted: 15th August 2007; Received (in revised form): 25th January 2008

Abstract

In the present contribution we propose two recently developed classification algorithms for the analysis of mass-spectrometric data—the supervised neural gas and the fuzzy-labeled self-organizing map. The algorithms are inherently regularizing, which is recommended, for these spectral data because of its high dimensionality and the sparseness for specific problems. The algorithms are both prototype-based such that the principle of characteristic representants is realized. This leads to an easy interpretation of the generated classification model. Further, the fuzzy-labeled self-organizing map is able to process uncertainty in data, and classification results can be obtained as fuzzy decisions. Moreover, this fuzzy classification together with the property of topographic mapping offers the possibility of class similarity detection, which can be used for class visualization. We demonstrate the power of both methods for two exemplary examples: the classification of bacteria (listeria types) and neoplastic and non-neoplastic cell populations in breast cancer tissue sections.

Keywords: classification; vector quantization; class visualization; machine learning; fuzzy-labeled self-organizing map; mass spectrometry

Corresponding author. Thomas Villmann, Medical Department, University Leipzig Germany. Tel: +49 (0)341 9718868; E-mail: thomas.villmann@medizin.uni-leipzig.de

Thomas Villmann is a senior researcher in the Medical Department, University of Leipzig, Germany and leads the Computational Intelligence group (http://www.uni-leipzig.de/~compint/). He holds a PhD and the venia legendi in Computer Science. His research areas comprise the theory of prototype-based vector quantization, neural networks and machine learning as well as respective applications in medical data analysis, bioinformatics and satellite remote sensing. Several research stays have taken him to Belgium, France, Netherlands, and the USA. He is founding member of the German chapter of the European Neural Network Society (GNNS).

Frank-Michael Schleif studied Computer Science at Leipzig University, graduating in 2002. He then became a PhD student at Leipzig University. In 2003, he joined Bruker Biosciences and continued his PhD studies at the Clausthal University of Technology. In 2006, he received a PhD in Computer Science and was awarded as the best C.S. PhD thesis at TUC. He is currently a research scientist in the MetaSTEM project team at IZKF (Interdisciplinary Center for Clinical Research) and a member of the Computational Intelligence Group at the Medical Department of the Leipzig University. His research activities focus on machine learning methods, statistical data analysis and algorithm development.

Markus Kostrzewa is a biologist and holds a PhD in Biology from the University Gießen. After a postdoctoral position at the University Gießen (genetic and physical mapping of a part of human chromosome 5 research in the field of hereditary neurodegenerative disorders molecular diagnostics of hereditary human diseases), he joined Bruker Daltonik GmbH Leipzig. Currently, he is the vice R&D Manager for the areas “Clinical Proteomics” and “Microorganism identification” and Director of Molecular Biology R&D. Further, he is member of the German Society of Human Genetics and the German Society of Neurogenetics.

Axel Walch is a pathologist and group leader of the Biomedical Microscopy Laboratory in the Institute of Pathology (Helmholtz Zentrum München, German Research Center for Environmental Health). His areas of expertise comprise quantitative pathology, cancer tissue biomarkers, esophageal cancer, gastric cancer and translation of molecular techniques to diagnostic pathology. The primary research focus is the identification of predictive and prognostic biomarkers in cancer.

Barbara Hammer received her PhD in Computer Science in 1995 and her venia legendi in Computer Science in 2003, both from the University of Osnabrueck, Germany. From 2000 to 2004, she was leader of the junior research group ‘Learning with Neural Methods on Structured Data’ at University of Osnabrueck. In 2004, she became professor for Theoretical Computer Science at Clausthal University of Technology, Germany. Several research stays have taken her to Italy, UK, India, France, and the USA. Her areas of expertise include hybrid systems, self-organizing maps, clustering, recurrent networks and their in bioinformatics, industrial process monitoring, or cognitive science.
INTRODUCTION

Exploration and analysis of mass-spectrometric data in the field of clinical proteomics have become one of the key problems in computational proteomics. Thereby, the complexity of the mass-spectrometric data is one difficult problem. Frequently, the data are given as huge-dimensional functional vectors with several thousands dimensions according to the resolution on the mass axis. Further, usually the number of samples is limited to a few data sets due to clinical restrictions. From a mathematical point of view, the data space to be explored is sparsely filled. Thereby, the spectra may be overlaid by noise such that the contained signal is difficult to extract. A further problem arises for data analysis methods as consequence of the high dimensionality: the data can always be separated more or less independent of the separation criteria [1]. Thus, any method is confronted with the problem of the detection of the underlying regularities. Usually, this problem is overcome by cross validation. Yet, the certainty of such an evaluation is diminished here, because of the humble number of data. Therefore, advanced methods for data analysis in mass spectrometry are required to be regularizing inherently, robust and to be able to deal with high-dimensional, sparse, noisy data.

In the following we will restrict us to classification problems. Thereby, we will concentrate to the following aspects: How we can achieve a good classification accuracy and how we can visualize classification results in an adequate manner. The latter problem is related to the problem of class similarity detection. Moreover, each classification result depends on the underlying similarity/dissimilarity measure for data.

Classification in classical statistics is frequently realized by Fisher's discriminant analysis (FDA) or linear/quadratic discriminant analysis (LDA/QDA) [2]. FDA optimizes the inter-intra-class correlation ratio by weighting the data dimensions to obtain a good separation plane, i.e. it is based on a weighted Euclidean distance for data similarity. LDA/QDA tries to optimize the Bayes error of the classification by utilization of the (class dependent, QDA) covariance, i.e. the Mahalanobis distance between data is used inherently [3]. These classical statistic approaches are more and more supplemented by machine-learning tools, which provide adaptive and robust methods for pattern recognition in complex data [4, 5, 6]. Thereby, machine-learning algorithms comprise approaches like artificial neural networks (ANNs), evolutionary algorithms (EAs), decision trees (DTs), clustering and other [7].

Beside the pure classification accuracy of a generated classification model, its interpretability plays an important role. Standard methods use (linear) principal component analysis (PCA) and FDA or classical hierarchical clustering [8, 9]. Additionally, advanced preprocessing procedures, like denoising using wavelets or ‘intelligent’ peak picking heuristics including problem specific expert knowledge, are applied to improve the accuracy. Here, the flexibility of machine-learning methods offers new ways which may result in better results [10]. For example, multilayer perceptron neural networks (MLPs) as universal function approximators offer, on the one hand side, greatest flexibility in learning and adaptation to achieve good classification results [11]. On the other hand, however, their decision scheme is more or less a ‘black box’, because all the information for the decision is distributed over the whole network. In contrast, prototype-based classifiers realize the principle of characteristic representatives for data subsets or decision regions between them. Thus, the interpretation becomes easy. Examples for such tools are Support Vector Machines (SVM) [12], Kohonen’s Learning Vector Quantization (LVQ), Self-Organizing Maps (SOMs) [13] and respective variants. New developments include the utilization of nonstandard metrics (functional norms, scaled Euclidean metric) and task-dependent automatic metric adaptation (feature selection), fuzzy classification and similarity-based visualization of data. These properties offer new possibilities for analysis also of mass-spectrometric data.

In the present article we will give insights to two recently developed prototype-based classifiers which fulfill the above requirements. The supervised neural gas (SRNG) and the fuzzy-labeled SOM (FLSOM) are robust prototype-based neural classifiers, which are inherently regularizing by neighborhood cooperation between prototypes and which are easy to interpret. Moreover, as we will explain FLSOM is able to detect class similarities and offers the possibility of fuzzy classification. Both algorithms share the flexibility of utilization of arbitrary data metrics, which may be adapted during the training process as well in dependence on the classification task to be learned [14].
The article is structured as follows. First, we shortly review both methods, classification based on LVQ and FLSOM, pointing out their different properties and abilities. Thereby, we will emphasize the ability of the usage of general, task adequate, similarity measures in both methods. Further, we will highlight the class similarity detection probability of the semi-supervised FLSOM, which can be used for adequate class visualization or clinical interpretation. The theoretic part is followed by two clinical example investigations. The first one is an investigation of mass-spectrometric bacteria data to find an adequate classification. In the second application a classification of neoplastic and non-neoplastic cell populations in histological sections of breast cancer tissue is considered. For both applications we demonstrate the advanced abilities of the methods. Concluding remarks complete the article.

PROTOTYPE-BASED CLASSIFIERS

Usually, spectrometric data in proteomic analysis are given as vectors \( v \in V \subseteq \mathbb{R}^D \). \( D \) is the data dimension which may be huge in this field. It depends on the spectral range and sampling resolution. Because the data represent spectra, or more general functions, they are called sometimes functional data. We remark that for functional data, the sequence of data dimensions is not independent.

In our consideration we assume, that there exist an underlying (unknown) data probability density \( P \) in \( V \). Further, we assume for the training data that for each data vector \( v \) a unique class label \( c(v) \) exists. Prototype-based classifiers distribute prototypes \( w_r \in \mathbb{R}^D, \; r \in A \), as representations for classes in the data space \( V \), whereby \( A \) is a given index set. The prototypes should represent class distributions in the data space and borders between different classes. For this purpose, each prototype has a class label \( y_r \).

Several approaches exist: the well-known LVQ family introduced by Kohonen tries to minimize the Bayes classification error, but the adaptation dynamic is only a heuristic Hebbian like and does not perform a gradient descent on the classification error [13]. SVMs are based on structural risk optimization using a separation margin maximization approach [15]. Both methods are very powerful. In particular, SVMs frequently show superior results [12]. However, if new data become available for training a complete new learning has to be applied for SVM. Both methods have in common that they are not able to handle uncertainty in classification for training data (fuzzy class memberships). Further, the obtained classification model is crisp by means of binary class decisions.

We now review two recently developed classification schemes, which both are inherently regularizing to address the above-mentioned problem of noisy and sparse data in huge-dimensional data spaces. The first one is a generalization of Kohonen’s LVQ scheme providing a gradient descent on a cost function. The second one extends the unsupervised SOM, such that a semi-supervised fuzzy classifier is obtained with excellent visualization abilities and the feature of class similarity detection. Both methods share the ability to proceed arbitrary (differentiable) may be parametrized data similarity measures, which itself can be in parallel subject of optimization with respect to the classification task.

Classification by supervised neural gas

In this article we introduce the supervised neural gas (SNG) algorithm. For this purpose we firstly present the generalized learning vector quantization approach and subsequently we derive from it the SNG by adding neighborhood cooperativeness.

The generalized learning vector quantization approach

As mentioned above, LVQ does not minimize the classification error by gradient descent prototype adaptation. Therefore Sato and Yamada introduced a cost function based on a classification function \( \mu \) such that the respective gradient descent is similar to the heuristic LVQ learning scheme preserving the Hebbian characteristic [16]. For a given data point \( v \) with class label \( c(v) \) the two best matching prototypes with respect to the data metric \( d \), usually the quadratic Euclidian, are determined: \( w_{r^+} \) has minimum distance \( d^+(v, w_{r^+}) \) and the class labels are identically: \( y_{r^+} = c(v) \). The other best prototype \( w_{r^-} \) has minimum distance \( d^-(v, w_{r^-}) \) but the class labels are different: \( y_{r^-} = c(v) \). Then the classification function \( \mu(v) \) is defined as

\[
\mu(v) = \frac{d^+ - d^-}{d^+ + d^-}
\]  

The value \( d^+ - d^- \) yields the hypothesis margin of the classifier [17]. Then the generalized LVQ (GLVQ) is derived as gradient descent of the cost function

\[
C_{GLVQ} = \sum_v f(\mu(v))
\]
with respect to the prototypes. \( f \) is the sigmoid function

\[
    f(x) = \frac{1}{1 + \exp(-x)}
\]

(3)

In one learning step for a given data point, both \( \mathbf{w}_r \) and \( \mathbf{w}_r^- \) are adapted in parallel. Taking the derivative yields the updates

\[
    \Delta \mathbf{w}_r = -\varepsilon^+ \cdot f'_{\mu(\mathbf{v})} \cdot \xi^+ \cdot \frac{\partial d(\mathbf{v}, \mathbf{w}_r)}{\partial \mathbf{w}_r} 
\]

and

\[
    \Delta \mathbf{w}_r^- = \varepsilon^- \cdot f'_{\mu(\mathbf{v})} \cdot \xi^- \cdot \frac{\partial d(\mathbf{v}, \mathbf{w}_r^-)}{\partial \mathbf{w}_r^-} 
\]

where \( \varepsilon^+ \) and \( \varepsilon^- \in (0, 1) \) are the learning rates. The logistic function \( f(x) \) is evaluated at position \( \mu(\mathbf{v}) \), and we get

\[
    \xi^+ = \frac{2 \cdot d^-}{(d^+ + d^-)^2}
\]

and

\[
    \xi^- = \frac{2 \cdot d^+}{(d^+ + d^-)^2}
\]

### Supervised neural gas

Yet, so far no inherently regularization is involved in the classification model. This feature can be included by combination of GLVQ with an unsupervised neural prototype vector quantizer—the neural gas (NG) [18].

The unsupervised Neural Gas The NG realizes a Hebbian learning of prototypes together with regularization by neighborhood cooperativeness between prototypes. The level of cooperativeness is determined in dependence on the similarity of the prototypes to a given data vector: Let \( \mathbf{W} \) be the set of prototypes and \( L(\mathbf{v}, \mathbf{W}) \) be the ordered list of prototype indices such that for each pair \( r_i, r_k \in L \) with \( i < k \) the relation \( d(\mathbf{v}, \mathbf{w}_{r_i}) \leq d(\mathbf{v}, \mathbf{w}_{r_k}) \) holds. Then the position \( i(\mathbf{r}) \) denotes the rank of the competition of the prototypes to be the best matching for \( \mathbf{v} \). The degree of cooperativeness is defined by

\[
    h_{\sigma}^{NG}(\mathbf{r}, \mathbf{v}, \mathbf{W}) = \exp \left( \frac{-i(\mathbf{r})}{2\sigma^2} \right)
\]

with neighborhood range \( \sigma \in \mathbb{R}^+ \). Then, the update of the prototypes is

\[
    \Delta \mathbf{w}_r = h_{\sigma}^{NG}(\mathbf{r}, \mathbf{v}, \mathbf{W})(\mathbf{v} - \mathbf{w}_r)
\]

for a given input vector \( \mathbf{v} \). This adaptation scheme realizes a stochastic gradient descent on a cost function and is mathematically equivalent to the dynamic of a diffusing gas into the space spanned by the data samples according to the data density [18, 19]. Thereby, the parameter \( \sigma \) describes the viscosity of the gas. Hence, high viscosity (high \( \sigma \)-values) corresponds to regularization and, hence, diminishes the possibility of overfitting of the gas particles (prototypes) to the data space whereas low values relax this requirement but lead to smaller description error [18]. Thus, one has to balance both aspects. Typical final \( \sigma \)-values are in the range \( <0.3 \) for good description errors but non-vanishing regularization whereas starting values are in the range of the number of prototypes.

Combination of GLVQ and NG Including the unsupervised NG into the GLVQ the SNG is obtained with inherent regularization ability according to the unsupervised NG [20]. For this purpose, we modify the GLVQ cost function (2) to

\[
    C_{SNG} = \frac{1}{C(\sigma, N_{\mathbf{W}(\mathbf{c})})} \sum_{\mathbf{v}} h_{\sigma}^{NG}(\mathbf{r}, \mathbf{v}, \mathbf{W}(\mathbf{c})) \cdot f(\mu'(\mathbf{v}))
\]

whereby, \( \mathbf{W}(\mathbf{c}) \) is the subset of all prototypes \( \mathbf{w}_r \) the class labels \( \mathbf{y}_r \) of which are equal to the class label \( \mathbf{c}(\mathbf{v}) \) of the data point. \( C(\sigma, N_{\mathbf{W}(\mathbf{c})}) \) is a constant depending on the cardinality \( N_{\mathbf{W}(\mathbf{c})} \) of the subset \( \mathbf{W}(\mathbf{c}) \) and the regularization level \( \sigma \). Derivation of this cost function leads to a similar adaptation scheme as for GLVQ. However, now a neighborhood cooperativeness is included between all prototypes of the correct class.

The update formulas for the prototypes can be obtained taking the derivative. For each \( \mathbf{v} \), all prototypes \( \mathbf{w}_r \in \mathbf{W}(\mathbf{c}) \) are adapted by

\[
    \Delta \mathbf{w}_r = -\varepsilon^+ \cdot \text{sgd}
\]

and the closest wrong prototype is adapted by

\[
    \Delta \mathbf{w}_{r^-} = \varepsilon^- \cdot \sum_{\mathbf{w}_r \in \mathbf{W}(\mathbf{c})} \text{sgd}
\]

whereby \( \varepsilon^+ \) and \( \varepsilon^- \in [0, 1] \) are learning rates and the logistic function is evaluated at position

\[
    \mu'(\mathbf{v}) = \frac{d_{r^+} - d_{r^-}}{d_{r^+} + d_{r^-}}
\]
The terms \( \xi^+ \) and \( \xi^- \) are obtained as
\[
\xi^+ = \frac{2 \cdot d_r}{(d_r + d_r^-)^2}
\]
and
\[
\xi^- = \frac{2 \cdot d_r^-}{(d_r + d_r^-)^2}
\]
Note that the updates of GLVQ are recovered for vanishing regularization \( \sigma \rightarrow 0 \). We remark that SNG also optimizes the hypothesis margin because the cost function contains the term \( d_r - d_r^- \).

The final classification of unknown data points \( v \) is then realized by a winner take all mapping for both GLVQ and SNG:
\[
v \mapsto c(v) = y_r \quad \text{such that } d(v, w_r) \text{ is minimum.} \tag{4}
\]
It has been demonstrated that SNG/GLVQ achieve excellent classification results [20, 21, 22, 23].

**Semi-supervised fuzzy classification by fuzzy-labeled SOM and class similarity detection**

**The fuzzy-labeled SOM–FLSOM**

We now turn to the more general task of fuzzy classification and classification visualization. For this purpose we assume that the number \( N_c \) of potential classes is known in advance. Then the class label \( c(v) \) is taken as a class membership vector \( c(v) \in \mathbb{R}^{N_c} \). The indices \( c_i(v) \in [0,1] \) of which describe the fuzzy degree of class membership of the data vector \( v \) and sum up to \( \sum_i c_i = 1 \). Analogously, the prototype labels are taken as vectors \( y_r \in \mathbb{R}^{N_c} \). In the following we will extend the unsupervised SOM model to deal with classification tasks.

SOMs are powerful models for unsupervised vector quantization [13]. In SOMs the index set \( A \) is a regular grid, usually a rectangular or hexagonal two-dimensional lattice. The indices \( r \) of the prototypes now indicate a location in the grid and, therefore, a natural metric \( || \cdot ||_A \) between them is induced. The mapping is like in SNG/GLVQ again a winner take all rule, which reads in the Heskes’ variant of SOMs as
\[
v \mapsto s(v) = \arg \min_{r \in A} \sum_{r' \in A} h_{c}(r, r') \cdot d(v, w_{r'}) \tag{5}
\]
with
\[
h_{c}(r, r') = \exp\left(-\frac{||r - r'||_A}{2\sigma^2}\right) \tag{6}
\]
as neighborhood function [24]. The SOM lattice can be considered as an elastic grid with viscosity in terms of the neighborhood parameter \( \sigma \) [13]. Again, as for NG, high values of \( \sigma \) are related to strong regularization whereas low values give only weak regularization. To prevent overfitting \( \sigma \)-values greater than 0.35 are recommended [25].

SOMs perform a topographic mapping of data under certain conditions [26], i.e. similar data points are mapped onto the same or onto neighbored grid locations. [For a detailed discussion of topographic mapping and more general lattice structures we refer to [26, 28].] The degree of topology preservation can be estimated by the topographic product TP [27]. TP values nearby zero indicate adequate topographic mapping. For optimum results the lattice size and dimension can be dynamically adapted during learning based on the evaluation of the receptive field of the prototypes [28]. Thereby, the receptive field of a prototype is all those data vectors which are mapped onto that prototype by Equation (5). This growing SOM (GSOM) generates a non-linear PCA of the data [29].

The learning in the Heskes–SOM follows a gradient descent on a cost function:
\[
E_{\text{SOM}} = \frac{1}{2C(\sigma)} \int P(v) \sum_{r} \delta_t(v) \sum_{r'} h_{c}(r, r') \xi(v, w_r) d v \tag{7}
\]
where \( C(\sigma) \) is a constant, which we will drop in the following, and \( \delta_t \) is the usual Kronecker symbol checking the identity of \( r \) and \( r' \). All prototypes are adapted according to
\[
\Delta w_r = -\varepsilon h_{c}(r, s(v)) \frac{\partial \xi(v, w_r)}{\partial w_r} \tag{8}
\]
with learning rate \( \varepsilon > 0 \).

Now we extend the cost function of the SOM as defined in Equation (7) to a cost function for semi-supervised fuzzy classification by
\[
E_{\text{FLSOM}} = (1 - \beta)E_{\text{SOM}} + \beta E_{\text{FL}} \tag{9}
\]
where \( E_{\text{FL}} \) measures the classification accuracy. The factor \( \beta \in [0,1] \) is a factor balancing unsupervised and supervised learning. One can simply choose \( \beta = 0.5 \), for example. However, instabilities in the adaptation dynamic occur if \( \beta > 0.85 \), whereas a great stability also in sense of accuracy is obtained if \( 0.5 < \beta \leq 0.85 \) [30].

We specify the classification term \( E_{\text{FL}} \) as
\[
E_{\text{FL}} = \frac{1}{2} \int P(v) \sum_{r} g_r(v, w_r) (x - y_r)^2 d v \tag{10}
\]
where \( g_f(v, w_r) \) is a Gaussian kernel describing a neighborhood range in the data space:

\[
g_f(v, w_r) = \exp \left( -\frac{d(v, w_r)}{2\gamma^2} \right) \quad (11)
\]

This choice is based on the assumption that data points close to the prototype determine the corresponding label if the underlying classification is sufficiently smooth. Note that \( g_f(v, w_r) \) depends on the prototype locations, such that \( E_{\text{FL}} \) is influenced by both \( w_r \) and \( y_r \). Hence, prototype adaptation is now influenced by the classification task via the labels:

\[
\frac{\partial E_{\text{FLSOM}}}{\partial w_r} = \frac{\partial E_{\text{SOM}}}{\partial w_r} + \frac{\partial E_{\text{FL}}}{\partial w_r} \quad (12)
\]

which yields

\[
\Delta w_r = -\epsilon(1 - \beta) \cdot h_0(r, s(v)) \cdot \frac{\partial d(v, w_r)}{\partial w_r} + \epsilon \beta \frac{1}{4\gamma^2} \cdot g_f(v, w_r) \cdot \frac{\partial d(v, w_r)}{\partial w_r} \cdot (x - y_r)^2 \quad (13)
\]

The label adaptation is only influenced by the second part \( E_{\text{FL}} \). The derivative \( \frac{\partial E_{\text{FL}}}{\partial y_r} \) yields

\[
\Delta y_r = \epsilon \beta \cdot g_f(v, w_r) \cdot (x - y_r) \quad (14)
\]

with learning rate \( \epsilon > 0 \). This label learning performs to a weighted average of the data fuzzy labels of those data close to the associated prototypes.

Classification of unknown data is again obtained by the mapping rule Equation (4), but now giving a fuzzy class membership vector response. Usually, the classification accuracy of FLSOM is slightly less then the accuracy of a pure (good) classifier, because the balancing parameter \( \beta \) cannot be set to the unit due to numerical stability reasons [31]. Hence, a remaining unsupervised amount of data information may lead to reduced accuracy. However, this disadvantage is compensated by the feature of inherent class similarity detection and the visualization abilities of FLSOM [32].

**Class visualization and class similarity detection**

As mentioned above, unsupervised SOMs generate a topographic mapping from the data space onto the prototype grid \( A \) under specific conditions. If the classes are consistently determined with respect to the varying data in a classification problem, one can expect for the semi-supervised topographic FLSOM that the class labels \( y_r \) become ordered within the distribution over the grid structure of the lattice \( A \). 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 between the fuzzy class label vectors within the neighbored of the considered grid locations. This is the consequence of the following fact: the neighborhood function \( h_\alpha(r, s) \) of the usual SOM learning Equation (8) forces the topological ordering of the prototypes. In FLSOM, this ordering is further influenced by the weighted classification error

\[
\epsilon_r(v, r) = g_f(v, w_r) \cdot (x - y_r)^2 \quad (15)
\]

which contains the data space neighborhood \( g_f(v, w_r) \), Equation (11). Hence, the prototype ordering contains information of both data density and class distribution, whereby for high balancing value \( \beta \) the latter term becomes dominant. Otherwise, the data space neighborhood \( g_f(v, w_r) \) also triggers the label learning Equation (14), which, of course, also depends on the underlying learned prototype distribution and ordering. Thus, a consistent ordering of the labels is obtained in FLSOM [33].

As a consequence, 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 FLSOM lattice \( A \). In case of overlapping class distributions the topographic processing leads to prototypes with unclear decision (labels), located between prototypes with clear vote. Further, if classes are not distinguishable, there will exist prototypes responsive to those data, which have class label vectors containing, approximately, the same degree of fuzzy class membership for the respective classes.

The fuzzy class membership vectors allow an easy visualization of the classification using their similarity property. For this purpose, all label vectors \( y_r, r \in A \) are embedded into a color space preserving their similarities. This can be realized by multi-dimensional scaling (MDS), for example. Doing so, similar classes are coded by similar colors, which may be used for class visualization [32].

**Classification task-dependent metric adaptation**

For the above-introduced prototype-based classification algorithms, the concept of classification-dependent metric adaptation can be applied.

Usually, the dissimilarity measure \( d(v, w_r) \) for the data space \( V \) is chosen as squared Euclidean metric in GLVQ, SNG and FLSOM. Thus the derivative \( \frac{\partial d(v, w_r)}{\partial w_r} \) simply becomes \(-2(v - w)\).
Depending on the classification task, this choice could be not optimum. Therefore, more appropriate (differentiable) similarity measures can be plugged into these algorithms instead, reflecting the nature of data or structure of classification. For example, Lee and Verleysen proposed a functional metric derived from the general Minkowski-metric for functional data paying attention to the spatial correlation between the components of functional vectors [34]. Other example, frequently used in biological and biochemical problems, are the Pearson correlation [35] and the Tanimoto kernel [36]. Due to the general formulation of the methods above, these metrics can easily plugged into the algorithms.

Yet, more flexibility can be obtained if $d(v, w)$ is a parametrized similarity measure. Then, the respective parameters may be also subject of optimization according to the given classification task [37, 20].

Generally, we consider a parametrized distance measure $d^\lambda(v, w)$ with a parameter vector $\lambda = (\lambda_1, \ldots, \lambda_M)$ with $\lambda_i \geq 0$ and normalization $\sum_{i=1}^{M} \lambda_i = 1$. Then, a classification task depending parameter optimization is achieved again by a gradient descent of the above cost functions but here with respect to these metric parameters. However, one has to pay attention to the fact that the prototype dynamic assumes a quasi-constant data metric. Hence, the adaptation of the metric has to be done with lower magnitude. This also prevents an overfitting of the metric.

One important example of a parametrized metric is the scaled squared Euclidean metric

$$d^\lambda(v, w) = \sum_i \lambda_i (v_i - w_i)^2$$

(with $\lambda_i \geq 0$ and $\sum \lambda_i = 1$). The derivative $\partial d^\lambda(v, w)/\partial w$ becomes $-2 \cdot \Lambda \cdot (v - w)$ with $\Lambda$ is a diagonal matrix and its $i$-th diagonal entry is $\lambda_i$, and $\partial d^\lambda(v, w)/\partial \lambda_i = (v_i - w_i)^2$.

The parameter optimization of the scaled squared Euclidean metric allows a useful interpretation. The parameters $\lambda_i$ 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 [37]. This weighting is structurally similar to the weighting in FDA. In case of zero valued $\lambda$, relevance learning can also be seen as feature selection. The vector $\lambda$ is called relevance profile. It can be used for advanced data investigation as it is shown in the applications.

We denote the algorithms GLVQ, SNG and FLSOM with incorporated metric adaptation (often called relevance learning) by GRLVQ, SRNG and RFLSOM, respectively. A block diagram showing the dependencies is provided in Figure 1.

**APPLICATION OF SNG/GRLVQ AND FLSOM FOR CLINICAL DATA SETS**

In this section we demonstrate the application of both, SNG/GRLVQ and FLSOM, to classification of two spectrometric data sets in bioinformatics. The problems are characteristic for bioinformatic tasks and therefore exemplary: (i) identification of bacteria. (ii) breast cancer tissue slice classification.

For both problems, the data metric for FLSOM was chosen as the squared scaled Euclidean metric Equation (16).

**Description of data and preprocessing**

The data for both problems are based on mass-spectrometric profiles measured by linear MALDI-TOF MS devices from Bruker Daltonik, Bremen, Germany.

**The bacteria data set**

The bacteria samples are obtained from extracts from listeria cell cultures (originial culture stems from the German Resource Center for Biological Material—DSMZ). The extracts, covered by an HCCA matrix, have been applied onto the MSP 96 target ground steel [38]. Profiling spectra were generated on a linear
Autoflex MALDI-TOF MS. Details can be found in [39].

Listeria is a bacterial genus containing six species. This species consist of *Listeria monocytogenes*, *Listeria innocua*, *Listeria ivanovii*, *Listeria seeligeri*, *Listeria welshimeri* and *Listeria grayi*. Listeria occur very common in nature environments and are also present in water, plants, food and the bowel of humans. The identification of listeria is therefore an important problem in biology. Listeria are known to be the bacteria responsible for listeriosis, a rare but lethal food-borne infection that has a devastating mortality rate of 25% [40]. Listeria, also have a particularly high occurrence rate in newborns because of its ability to infect the fetus by penetrating the endothelial layer of the placenta [38]. Thereby *L. monocytogenes* is considered to be pathogenic for humans. Listeria in food are relatively rare but due to the increasing industrial production of food with many processing steps, the risk of a listerial contamination is increasing, which rises the needs for improved product safety and quality control. The diagnosis of listeria at an early stage is important for therapeutic approaches on humans. The expression of a infection caused by listeria may delay upto 8 weeks. To identify whether a listeria infection is present, the blood or matter is taken from the patient and a cultivation is tried. This, however, fails in part and, hence, the disease can not be diagnosed in time.

In the available data set all six listeria types are present. Thereby, for the *L. grayi* a subgroup of *Listeria grayi murrei* can be identified and for *L. ivanovii* a distinction into the subgroups *L. ivanovii ssp ivanovii* and *L. ivanovii ssp londoniensis* can be made. Thus, the data set consists of 109 profile spectra in eight classes with at least six samples for each class. The spectral range is between 2 kDa and 20 kDa. The obtained spectra have been smoothed, baseline corrected, and aligned following the standardized preprocessing tool BIOTYPER™ 1.1 from Bruker Daltonik, Bremen, Germany. The involved peak picking generates a peak list vector for each spectrum. All peak list vectors are aggregated such that finally a data matrix with 937 intensity components and 109 rows is achieved as data base. Thereby, a peak shift tolerance of 300 ppm was used. A classification tree obtained by the BIOTYPER™ 1.1 software, which inherently applies a single-linkage clustering, yields a class separation tree as depicted in Figure 2. This tree can serve for comparison for FLSOM class similarity detection.

**The breast cancer tissue data set**

The breast cancer tissues are collected at the Institute of Pathology in Neuherberg, Helmholtz-Zentrum-National Research Center for Environment and Health, Germany. The generic measurement procedure can be summarized as follows. The frozen tissue is cut using a cryomicrotome in sections of 12 μm and transferred to a conductive slide, washed in ethanol and coated by matrix. Reference sections are used for histological staining (Hematoxylin and Eosin, immunohistochemical staining for HER2) and histomorphological classification. The slices are subsequently measured in a Ultiflex II MALDI-TOF and subsequently visualized using the FlexImaging™ tool provided by Bruker Daltonik, Bremen, Germany [41].

The breast cancer tissue slices are manually labeled by a clinical expert (pathologist). In this exemplary study the slice of one patient is used. Four different spatial regions of the slice are marked according to histomorphologically classified cell types: connective tissue, inflammation, and two morphologically distinct tumor cell populations (tumor-type-1, tumor-type-2). From the whole slice 687 spectral record are generated, 438 of them are labeled with at least 51 records per region.

These profiles were preprocessed (baseline correction, alignment, peak picking, and peak feature extraction by means of maximum intensities) according to the CLINPROTOOLS™ 2.1 from Bruker.
Daltonik, Bremen, Germany [42]. Finally, each preprocessed data record is a 70-dimensional data vector.

Application of the methods and interpretation of the results

Both data sets are analyzed by SNG and FLSOM using the scaled quadratic Euclidean metric as data similarity measure. For comparison we also applied SVM with different kernels, and exemplary for traditional statistic classification analysis, LDA based on linear PCA. The SVM-approaches use a linear kernel for SVM1, the Tanimoto-kernel for SVM3 and a radial basis function (RBF) kernel for SVM2. The Tanimoto-kernel is without any additional parameter. The parameter C of the linear kernel was inherently determined according to the estimation procedure proposed in [43]. The SVM parameters for the RBF-kernel have been automatically determined by a grid search [44].

Additionally, for the bacteria data set a class dependence tree provided as standard solution of the BIOTYPERTM 1.1 tool based on hierarchical clustering is also available for comparison [38].

In case of FLSOM application we further investigate the detected class similarities and provide visualization results. The optimum FLSOM lattice size was estimated by a GSOM using standard Euclidean metric for data.

Results for the bacteria data set

First, we applied a growing variant of SNG and obtained three prototypes per class, which is sufficient to get excellent results. The growing scheme is based on the unsupervised growing neural gas (GNG) [45]. Further increasing of the number does not lead to any improvements. The neighborhood range σ for regularization is slowly decreased to 0.3 during the adaptation process such that a weak regularization remains. For FLSOM a two-dimensional 12 × 4 lattice structure is suggested by the GSOM. Due to the large number of prototypes for FLSOM in comparison to SNG and paying attention to the sparse data the final regularization parameter for FLSOM is set to σ = 0.4, which yields non-vanishing regularization. The balancing parameter was set to β = 0.05 in the beginning and increasing up to the final value of β = 0.85. The topology preservation of the FLSOM is preserved, as the topographic product value TP = −0.0066 indicates. The 5-fold cross-validated classification accuracy results are collected in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>LDA</th>
<th>SVM1</th>
<th>SVM2</th>
<th>SVM3</th>
<th>SNG</th>
<th>FLSOM</th>
</tr>
</thead>
<tbody>
<tr>
<td>61.5%</td>
<td>34.9%</td>
<td>&lt;10%</td>
<td>96.3%</td>
<td>97.8%</td>
<td>73.4%</td>
<td></td>
</tr>
</tbody>
</table>

SVM1 is a linear SVM, SVM2 uses a radial basis function kernel and SVM3 uses a Tanimoto-distance-kernel. LDA is based on linear PCA suggesting five principal components to be sufficient. For FLSOM majority vote is taken to obtain the crisp classification.

Thereby, we observed equal accuracies for each class, approximately.

Both, SNG and FLSOM show very good performance in comparison to the other algorithms. In particular, we remark that SVM heavily depends on the used kernel, as it is also known from other applications [46]. The slightly decreased accuracy of the FLSOM is the consequence of the balancing parameter β < 1, which is necessary for stability reasons of the algorithm as described before. However, this disadvantage is compensated by the class similarity detection feature provided by FLSOM [33]. These results are now under deeper consideration. The fuzzy class label vectors yr of the prototypes are depicted in Figure 3a) according to their arrangement on the FLSOM lattice. An exemplary interpretation of the barplots in Figure 3 is provided by Figure 4.

The distribution of the prototype labels suggests the following interpretation of FLSOM-detected class similarities, which should also be compared to the above given separation tree obtained by the BIOTYPERTM 1.1 software depicted in Figure 2: The Listeria of grayi types (class 1 and 2) should not be distinguished according to their proteomic fingerprint. The class 8 (L. welshimeri) is clearly isolated from each other. Classes 4, 5 and 7 (L. ivanovii ssp ivanovii, L. sedligeri and L. ivanovii ssp londoniensis) are very similar. Further there is a class similarity between the classes 3 and 4 (L. innocua and L. ivanovii ssp ivanovii). Although this similarity in the tree classification can not be ruled out, the tree visualization suggests a stronger separation. This ‘separation’ would be disappear, if the respective branch would be rotated. Thus, the FLSOM label distribution is more adequate. Further, FLSOM detected a similarity between the classes 6 and 4 (L. monocytogenes and L. ivanovii ssp ivanovii), which is also not easily detectable in the tree visualization. The class 3 (L. innocua) shows multiple similarities to several other species based on the proteomic fingerprint. This sharing property can not
Figure 3: Distribution of the class responsibilities within the FLSOM-lattices for (a) the listeria classification problem ($12 \times 4$-FLSOM) and (b) the breast cancer problem ($15 \times 5$-FLSOM). The label vectors $y_r$ are depicted as barplots arranged according to the FLSOM-grid structure. Each barplot refers to a label vector, whereby the height of the bars within is according to the probability that the prototype is responsible for the respective class. Flat lines show ‘dead’ prototypes, i.e. which did not win the competing process for the available data. For an exemplary interpretation of a barplot, see Figure 4.

Figure 4: Exemplary interpretation of the barplots arranged on the grids: Here are depicted five barplots arranged on a chain (taken from Figure 3). The left barplot suggest high-class responsibility for class 4 and minor probability for class 3. The next barplot does not make a clear decision between the classes 2, 3 and 4 whereas in the third barplot the classes 1 and 2 dominate with no significant distinction between them. The last two plots show a unique decision for class 1. Thus we can observe a smooth transition from class 4 decision to class 1 going from left to right. Because the underlying prototypes are topologically ordered, we may conclude that the respective class similarities are ordered, too.
be reflected adequately in the tree classification. However, an expert biologist independently suggested a similarity between both types. [Personal communication with Dr Thomas Maier, Bruker Daltonik Leipzig, Germany.]

Thus summarizing, the FLSOM provides detailed class similarity description together with comparable classification accuracy, whereas SNG achieves best accuracy.

**Results for the breast cancer tissue data set**

Again, we applied SNG with three prototypes per class and the neighborhood range $\sigma$ for regularization also slowly decreasing to 0.3 during the adaptation process. For FLSOM a two-dimensional $15 \times 5$ lattice structure is suggested here by the GSOM. As for the listeria data set the final regularization parameter for FLSOM is set to $\sigma = 0.4$, which yields non-vanishing regularization. The balancing parameter setting was the same as for the bacteria data set. The topographic product value $TP = 0.0001$ ensures the topographic mapping of the FLSOM. The 5-fold cross-validated classification accuracy results are collected in Table 2 with approximately same accuracy for each class for the several algorithms (variance 0.147).

The obtained accuracies for SNG and FLSOM show high levels with a slightly decreased value for FLSOM, whereas SNG achieves the overall best result.

For class similarity investigation we consider the distribution of the label vectors within the FLSOM-grid, which is depicted in Figure 3. The connective-tissue class is well separated from the others. Further, we have in the distribution plane an overlapping region between the both tumor classes, which indicates a similarity between them. The FLSOM detects a clear distinction between tumor-1-class and the connective tissue class according to the spatial distribution of the respective labels in the FLSOM.

---

**Table 2:** Classification accuracies for the different classifiers for the breast cancer data set

<table>
<thead>
<tr>
<th></th>
<th>LDA</th>
<th>SVM$_1$</th>
<th>SVM$_2$</th>
<th>SVM$_3$</th>
<th>SNG</th>
<th>FLSOM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>59.8%</td>
<td>62.8%</td>
<td>42.7%</td>
<td>84.2%</td>
<td>80.4%</td>
<td>72.4%</td>
</tr>
</tbody>
</table>

SVM$_1$ is a linear SVM, SVM$_2$ uses a radial basis function kernel and SVM$_3$ uses a Tanimoto-distance-kernel. LDA is based on linear PCA suggesting eight principal components to be sufficient. For FLSOM majority vote is taken to obtain the crisp classification.

---

**Figure 5:** Breast cancer tissue section with manually labeled areas used for classification training is depicted left. Right hand, the classification obtained by the FLSOM classifier is plotted using an MDS RGB-color embedding of the FLSOM-label vectors $y_r$. Thereby, similar colors represent similar class properties as detected by FLSOM (black — not used for classification). One clearly see the fine agreement with the manual labeling. (A color version of the figure can be obtained from the authors on request.)
grid, whereas small overlapping between tumor-
2-class and the connective-tissue class occurs. The
inflammation class shows similarity to type-2-tumor.
Using an MDS color embedding of all label vectors $y_r$
of the FLSOM into the RGB-color space, the
FLSOM classification of the tissue can be easily
visualized, Figure 5. A high agreement between
original manually labeled tissue and obtained coloring
based on the classification and class similarity detec-
tion generated by the FLSOM can be observed.

Figure 6: Relevance profile of the scaled Euclidian metric for the breast cancer problem. The highest relevance peak can be assigned to the data dimension 26 which is assigned to the 4971 Da-peak in the original proteomic spectrum.

Figure 7: Recoloring of the original tissue according to the 4971 Da-intensities. High intensities are colored red, low values are coded by blue colors. (A color version of the figure can be obtained from the authors on request.)
Further information can be obtained by consideration of the learned relevance profile of the problem specific scaled Euclidian metric. It is depicted in Figure 6. The highest relevance for class separation can be assigned to the 4971 Da-peak in the original proteomic spectra. Recoloring of the original tissue according to the 4971 Da-intensities shows that this peak mainly separates the connective tissue class from the other classes, Figure 7. Analogously, the other relevance peaks could be evaluated. In this way, a detailed analysis of the information contained in the FLSOM model can be obtained.

CONCLUDING REMARKS
In this article two recently developed prototype-based methods for classification, SNG and FLSOM, are reviewed in the light of the analysis of mass-spectrometric data in bioinformatics. Both approaches are adaptive machine learning approaches and allow easy retraining, if new data become available. They are both inherently regularizing by non-vanishing neighborhood cooperativeness, such that they are able to handle sparse, high-dimensional and noisy data. As demonstrated for two exemplary problems in classification of proteomic spectra (bacteria and breast cancer tissue), the generated classification models show good performance compared to other machine learning and statistical methods.

Additionally, FLSOM provides the possibility of processing uncertain class information for training data (fuzzy) and returns a fuzzy classification scheme. Moreover, FLSOM provides a class similarity detection based on the fuzzy labels, which give the possibility of deeper class analysis offering more information than simple classification trees. The fuzzy classification can further be used for class dependent data visualization whereby similar class information is encoded by similar colors such that an easy interpretation can be made.

Compared to traditional statistical methods for classification like FDA/LDA/QDA, the algorithms SNG and FLSOM offer more variability due to their local view on data by prototypes. Thus, each prototype can respond for different data regions specifically. This ability can be further emphasized using prototype specific metric adaptation as it was suggested in [47].

Otherwise, both SNG/FLSOM and SVM belong to the class of prototype based classifiers. The basic conceptual difference between them is that SVM takes data points on the class borders (or its representations in the kernel space) as prototypes, whereas SNG/FLSOM prototypes are local averages of data points, which are nearby or on the class borders. This reference principle has makes easy interpretations possible. Further, if new data become available SNG/FLSOM can be adapted by fast retraining whereas for SVM models retraining is difficult or impossible if the underlying theory of convex optimization for structural risk minimization should be kept.

Future developments of SNG/FLSOM could incorporate the knowledge on data structures. For example, there exist several methods in unsupervised data analysis like principal component analysis or clustering for functional data [48]. For prototype based classification schemes Sobolev-metrics could be an interesting alternative to standard (scaled) Euclidean metric [14, 49, 50].

Key Points

- Prototype-based fuzzy classification of mass spectra by supervised topographic mapping.
- Class visualization based on the detected class similarities by similarity preserving color coding.

References


40. Todar K. Todar’s online textbook of bacteriology – listeria monocytogenes and listeriosis. Department of Biology,


