Advanced metric adaptation in Generalized LVQ for classification of mass spectrometry data

P. Schneider, M. Biehl
Mathematics and Computing Science, University of Groningen, The Netherlands
email: {petra, biehl}@cs.rug.nl

F.-M. Schleif
AG Computational Intelligence, University of Leipzig, Germany
email: schleif@informatik.uni-leipzig.de

B. Hammer
Institute of Computing Science, Clausthal University of Technology, Germany
email: hammer@in.tu-clausthal.de

Keywords: GLVQ, adaptive metric, proteomics, mass spectrometry

***Abstract*— Metric adaptation constitutes a powerful approach to improve the performance of prototype based classication schemes. We apply extensions of Generalized LVQ based on different adaptive distance measures in the domain of clinical proteomics. The Euclidean distance in GLVQ is extended by adaptive relevance vectors and matrices of global or local influence where training follows a stochastic gradient descent on an appropriate error function. We compare the performance of the resulting learning algorithms for the classification of high dimensional mass spectrometry data from cancer research. High prediction accuracies can be obtained by adapting full matrices of relevance factors in the distance measure in order to adjust the metric to the underlying data structure. The easy interpretability of the resulting models after training of relevance vectors allows to identify discriminative features in the original spectra.**

1 Introduction

The search for disease marker is a prominent task in a clinical study. Recent achievements in clinical proteomics are especially promising to get predictive models or markers for different kinds of diseases [6, 4]. Clinical proteomics refers to the analysis of the proteome, the whole set of proteins of an organism, in the clinical domain. The data are typically taken from blood or urine samples. One prominent technique to analyse proteomic data is available by mass spectrometry. The obtained spectral data are high dimensional measurements, with more than 10000 measurement points and need an appropriate preprocessing as well as sophisticated high level data analysis approaches, to attain validated signal patterns. Focusing on classification between different clinical states the generalization ability as well as the interpretability of the models is especially important. Prototype based classification approaches such as Learning Vector Quantization (LVQ) as proposed by Kohonen [5] or multiple extensions [2, 1] have already proven to be valuable in such experiments (see [7, 8]). Due to the complexity of the data and the desired task of biomarker discovery the used metric becomes very important as shown in [10]. In the previously published approaches correlative effects between different features were ignored in general. For LVQ type algorithms with generalized relevance learning [3], a more powerful alternative, which includes adaptive relevance factors in the diagonal matrix of the metric, is available. This allows to scale the axes in order to obtain better adaptation towards clusters with axes-parallel ellipsoidal shapes.

For mass spectrometry data of proteome studies local and global correlations are very likely to provide additional information for the classification method which have not been captured so far. In this context, the recently introduced Generalized Matrix LVQ (GMLVQ) [11, 14] as an example of a prototype based classifier is applied to two different data sets of proteome studies. We allow for a full adaptive matrix, i.e. the possibility to adapt to arbitrary (local) ellipsoids which correspond to locally correlated input dimensions. We show that this general method leads to efficient and powerful classifiers with excellent generalization ability on this kind of data. The set of discriminative features can still be related back to the original spectral masses, allowing for the identification of potential biomarker patterns.

2 Clinical data and preprocessing

The proposed method is evaluated on two clinical data sets obtained from blood plasma samples. The first data set has been taken from 45 patients suffering from colorectal cancer and appropriate 50 non-diseased controls. The second more complex data set was taken from 50 patients with

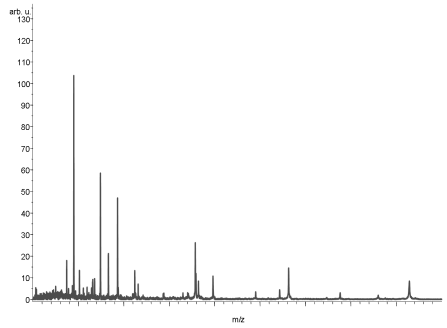


Figure 1: Sample from a linear MALDI-TOF mass spectrum taken from the lung cancer data set.

lung cancer and 50 healthy controls. All samples have been measured using a linear MALDI-TOF MS [1]. This leads to spectra with approximately 20000 measurement points ($1kDa - 10kDa$). Each spectrum has been baseline corrected and aligned as recommended in the standard procedure by use of ClinProTools [2]. The details on the sample collections and the overall measurement procedure can be found in [9]. A sample spectrum is depicted in Figure 1.

To obtain discriminant features for our analysis a multiresolution wavelet analysis by a discrete biorthogonal wavelet transform has been applied. Here, bior3.7 wavelets at scale $L = 4$ have been taken as the final feature set (details in [9]). By application of this procedure we finally analyse sets of 95 or 100 spectra with 1408 wavelet coefficients. In one of the subsequently shown experiments we used the data in a measurement range of $1500kDa - 3500kDa$ where most of the signal information was observed, encoded by 417 coefficients. By use of that approach the features can be still related back to the original mass positions in the spectra which is important for potential biomarker analysis by additional measurements where the mass positions are relevant. In addition the chosen level of abstraction in the wavelet analysis allows to preserve also small peak widths.

3 Generalized Matrix LVQ

Learning Vector Quantization belongs to the class of distance based classification schemes. The training data is approximated by a set of prototype vectors in the same space, which can be used for nearest prototype classification. Consider a C-class classification problem in a n-dimensional space with the given training data $X = \{(\xi^j, c(\xi^j)) \in \mathbb{R}^n \times \{1, \ldots, C\}\}_{j=1}^m$. An LVQ-classifier is defined by a set of prototype vectors $\mathbf{w}^i \in \mathbb{R}^n, i = 1, \ldots, c$, which represent the different classes and are masked with labels $c(\mathbf{w}^i) \in \{1, \ldots, C\}$. Classification is based on a distance measure d, which evaluates the similarity between given data and the prototypes. A data point ξ is assigned to class label $c(\xi) = c(\mathbf{w}^i)$ of prototype $\mathbf{w}^i$ for which $d(\mathbf{w}^i, \xi) \leq d(\mathbf{w}^j, \xi)$ holds for all $j \neq i$.
Learning aims at finding a set of prototypes such that the training samples are mapped on their respective class labels. Generalized LVQ [1] does so by minimizing the cost function S in equation (1) with a stochastic gradient descent procedure.

$$S = \sum_{i=1}^{m} f(\mu(\xi^i)) \text{ with } \mu(\xi^i) = \frac{d^J - d^K}{d^J + d^K} \qquad (1)$$

where f is a monotonically increasing function, $d^J = d(\mathbf{w}^J, \xi^i)$ is the distance of feature vector ξ^i to the closest prototype $\mathbf{w}^J$ with $c(\xi^i) = c(\mathbf{w}^J)$ and $d^K = d(\mathbf{w}^K, \xi^i)$ is the distance to the closest prototype $\mathbf{w}^K$ with $c(\xi^i) \neq c(\mathbf{w}^K)$. The derivatives of $f(\mu(\xi))$ with respect to $\mathbf{w}^J$ and $\mathbf{w}^K$ for a randomly selected example ξ yield the update rules for the GLVQ-algorithm [1].
The underlying similarity measure d is of special importance for the performance of LVQ-classifiers. GLVQ is based on the standard Euclidean metric and thus possibly fails for high-dimensional or heterogeneous data [13]. An extension of GLVQ, Generalized Relevance LVQ [3], is based on the more powerful *weighted* Euclidean metric $d_\lambda(\mathbf{w}, \xi) = \sum_i \lambda_i (\xi_i - w_i)^2$ with $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$. The vector $\lambda \in \mathbb{R}^n$ weights the input dimensions according to their relevance to solve the classification task. It is adapted to the data during training via a stochastic gradient descent as well. This strategy helps to prune out irrelevant or noisy dimensions and allows to identify the features which provide the most discriminative power. The metric becomes even more powerful by assigning an individual weight vector λ^j to each prototype $\mathbf{w}^j$. This method (Localized GRLVQ) has been investigated in [12] and takes into account that relevances might differ between different classes.
Recently, we have extended the weighted Euclidean metric d_λ by introducing a full matrix $\Lambda \in \mathbb{R}^{n \times n}$ of relevance factors in the distance measure [11, 14]. The metric has the form

$$d_\Lambda(\mathbf{w}, \xi) = (\xi - \mathbf{w})^T \Lambda (\xi - \mathbf{w})$$

This approach allows to account for correlations between different input features. A set of points equidistant from a prototype can have the shape of a rotated ellipsoidal, whereas the relevance vector λ in GRLVQ only results in a scaling parallel to the coordinate axis.
For the distance measure d_Λ to be well defined, the matrix Λ has to be positive (semi-) definite. For this reason, Λ is substituted by $\Lambda = \Omega\Omega^T$ with $\Omega \in \mathbb{R}^{n \times n}$. To obtain the adaptation formulas, the derivatives of (1) with respect to

[1] Bruker Daltonik GmbH, Bremen, Germany
[2] http://clinprot.bdal.de

$\mathbf{w}^J$, $\mathbf{w}^K$ and Ω have to be computed.

$$\Delta \mathbf{w}^J = +\epsilon_1 \cdot \frac{4\, f'(\mu_\Lambda(\xi))\, d_\Lambda^K}{(d_\Lambda^J + d_\Lambda^K)^2} \cdot \Omega\Omega \cdot (\xi - \mathbf{w}^J)$$

$$\Delta \mathbf{w}^K = -\epsilon_1 \cdot \frac{4\, f'(\mu_\Lambda(\xi))\, d_\Lambda^J}{(d_\Lambda^J + d_\Lambda^K)^2} \cdot \Omega\Omega \cdot (\xi - \mathbf{w}^K)$$

$$\Delta \Omega_{lm} = -\epsilon_2 \cdot \frac{2\, f'(\mu_\Lambda(\xi))}{(d_\Lambda^J + d_\Lambda^K)^2} \cdot \Big(d^K \left([\Omega^J]_m (\xi_l - w_l^J) + [\Omega^J]_l (\xi_m - w_m^J) \right) - d^J \left([\Omega^K]_m (\xi_l - w_l^K) + [\Omega^K]_l (\xi_m - w_m^K) \right) \Big)$$

where $[\Omega^J] = \Omega(\xi - \mathbf{w}^J)$ and $[\Omega^K] = \Omega(\xi - \mathbf{w}^K)$.
Note that we can assume $\Omega^\top = \Omega$ without loss of generality and that the symmetry is preserved under the above update. After each update step Λ has to be normalized to prevent the algorithm from degeneration. It is enforced that $\sum_i \Lambda_{ii} = 1$ by dividing all elements of Λ by the raw value of $\sum_i \Lambda_{ii}$. In this way the sum of diagonal elements is fixed which coincides with the sum of eigenvalues here. This extension of GRLVQ is named Generalized Matrix LVQ (GMLVQ) [11, 14].
By attaching local matrices Λ^j to the individual prototypes $\mathbf{w}^j$, ellipsoidal isodistances with different widths and orientations can be obtained. The algorithm based on this more general distance measure is called Localized GMLVQ (LGMLVQ) [11, 14].

4 Results

As a preprocessing step, the features were normalized to zero mean and unit variance. In a first experiment, all 1408 wavelet coefficients were used to perform classification with GRLVQ and LGRLVQ. Each class was approximated by one prototype vector. The initial learning rate for the prototypes was set to $\epsilon_1 = 0.01$, for the relevance vectors it was chosen slightly smaller and set to $\epsilon_2 = 0.005$. The learning rates were continuously reduced in the course of learning. We implemented a learning rate schedule of the form

$$\epsilon_{1,2}(t) = \frac{\epsilon_{1,2}}{1 + c(t - \tau_{1,2})}$$

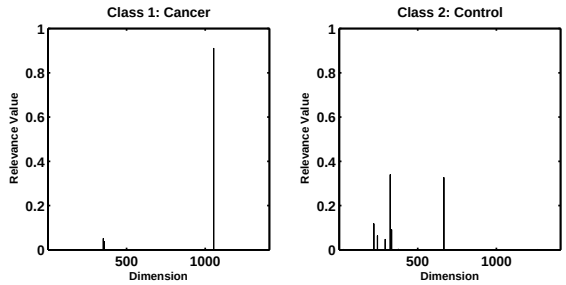
with $\tau_1 = 1$, $\tau_2 = 50$ and $c = 0.001$. Here, t counts the number of randomly shuffled sweeps through the training set, τ_2 gives the start of relevance learning after an initial phase of pure prototype training. Because of the limited number of training samples, we performed a 5-fold cross validation to evaluate the quality of the classification (see Table 1 for the mean test errors and standard deviations over the five training sets). The resulting relevance profiles reflect that the algorithms select only a very small subset of coefficients to distinguish between the two classes. The majority of relevance values goes down to zero in all runs. Pruning of less relevant dimensions clearly improves the algorithms performance compared to GLVQ which shows divergent behaviour in all experiments. After GRLVQ training, the classification is even based on only two features, but still achieving a mean accuracy above 80%. Figure 2 depicts typical relevance profiles which are achieved when training with the local weighted Euclidean metric. Depending on the training set, the number of relevant features per class varies from one to five (colorectal cancer data) and from two to six (lung cancer data). One identified region in the colorectal cancer data is depicted in Figure 3. It provides a clear separation between the two classes. Furthermore, we could observe that most misclassifications are caused by samples belonging to the diseased-class. But increasing the number of prototypes for this class indicates over-fitting effects and does not improve the prediction accuracies on the test sets. By inspection of peaks with a relevance of > 0.05 we observed that some features contain signal information with very low intensities. Such weak signals are however probably not usable in case of later biomarker identification and should be removed, therefore the dimensionality of the feature space may be further reduced taking such additional expert knowledge into account.

colorectal cancer		
Algorithm	Mean(Test)	σ
GRLVQ	84.21	5.3
LGRLVQ	88.42	6.9

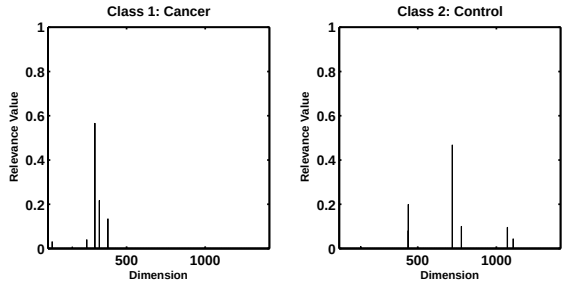
lung cancer		
Algorithm	Mean(Test)	σ
GRLVQ	84	4.2
LGRLVQ	82	0.08

Table 1: Mean values and standard deviations of the classification accuracies (in %) over the five different test sets, based on 1408 wavelet coefficients.

In a second experiment, (L)GRLVQ and (L)GMLVQ were applied to a reduced subset of only 417 coefficients. The selection of the coefficients ranges from $1500kDa$ to $3500kDa$ and was provided by a biological expert. The parameter settings were chosen as before and we used the same learning rates for the relevance vectors and the matrices, respectively. Table 2 summarizes the results obtained by the different algorithms. By using the new distance measure d_Λ, the classification performance could be improved in comparison to GRLVQ and LGRLVQ for both data sets. The observed improvement is more than 3% for the colorectal cancer classification and more than 6% for lung cancer classification. Despite the large number of parameters we observed a very stable behaviour and fast convergence of GMLVQ and LGMLVQ in less then 150 cycles. Training of relevance vectors results in a strong pruning of less relevant dimensions again, similar to the previ-



(a) Colorectal cancer data set



(b) Lung cancer data set

Figure 2: Local Relevance Profiles after LGRLVQ training with 1408 wavelet coefficients.

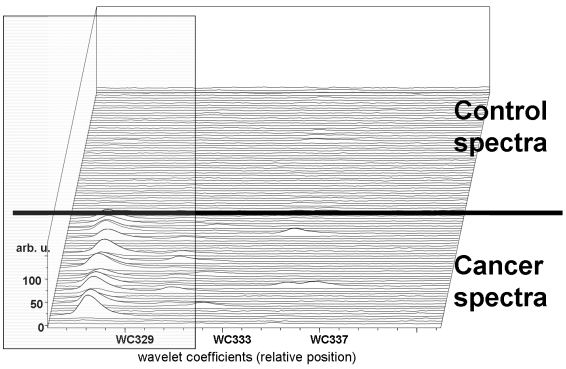


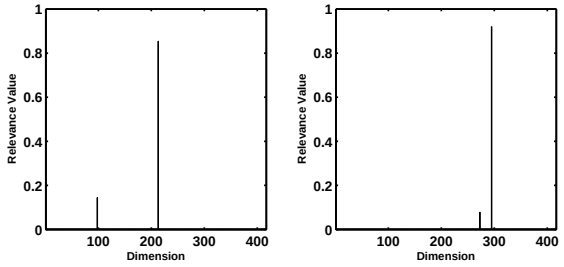
Figure 3: Relevant region for the colorectal data set around wavelet coefficient 325 (area in the box on the left). One observes a quite good separation between the two classes. The wavelet coefficient encodes the left shoulder of the peak which is more discriminative than the peak maximum with respect to the same region in the control group.

ous experiments (see Figure 4). But it is interesting to note that after matrix learning, classification is not based on a reduced subset of the original features any more. Figure 5 visualizes the diagonal elements of the global relevance matrices which came out of GMLVQ training with the same training sets used for Figure 4. By comparing the profiles it becomes clear that there is no special focus on particular coefficients any more. This indicates that other directions in data space provide more information to distinguish between the two classes than the original input features. The new features detected by GMLVQ and LGMLVQ provide more discriminative power, whereas they allow no semantic interpretation.

colorectal cancer		
Algorithm	Mean (Test)	σ
GRLVQ	84.21	5.3
LGRLVQ	88.42	5.8
GMLVQ	91.58	4.7
LGMLVQ	92.63	6.0

lung cancer		
Algorithm	Mean (Test)	σ
GRLVQ	83	2.7
LGRLVQ	80	1.0
GMLVQ	83	1.2
LGMLVQ	89	6.5

Table 2: Mean values and standard deviations of the classification accuracies (in %) over the five different test sets, based on a subset of 417 wavelet coefficients.

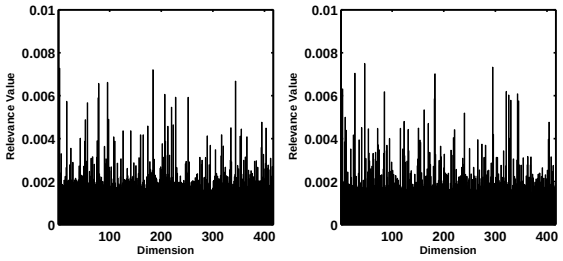


(a) (b)

Figure 4: Global relevance vectors after GRLVQ training with 417 wavelet coefficients. (a) Colorectal cancer data set, (b) Lung cancer data set.

5 Conclusion

In this article different variants of relevance learning in the context of learning vector quantization have been successfully applied in the classification of mass spectrometric data. The obtained classification results show for each data set an overall improvement in the prediction accuracy using the Generalized Matrix LVQ approach. Local relevance learning has been found to be valuable with respect to improved prediction accuracy as well as with respect to the identification of class specific relevant features. The



(a) (b)

Figure 5: Diagonal elements of the global relevance matrices after GMLVQ training with 417 wavelet coefficients. (a) Colorectal cancer data set, (b) Lung cancer data set.

relevant features indicated by Generalized Relevance LVQ could be verified to be relevant with respect to the underlying masses in the original spectral data through visual interpretation by experts. The wavelet encoding facilitates also the identification of features which belong to small peaks and not fully resolved peak regions like peak-shoulders.
Interestingly, matrix adaptation turned out to be quite robust despite from a large number of free parameters (which is quadratic with respect to the input dimensionality). It can be assigned to the fact that matrix learning constitutes a large margin optimization method with excellent gerenalization ability such as GRLVQ itself where generalization bounds which are independent on the number of free parameters but which depend on the hypothesis margin of the classifier can be derived [11, 14].

References

[1] A. Sato and K. Yamada. Generalized learning vector quantization. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference, pages 4239. MIT Press, Cambridge, MA, USA, 1996.

[2] B. Hammer, M. Strickert, and T. Villmann. Supervised neural gas with general similarity measure. Neural Processing Letters, 21(1):2144, February 2005.

[3] B. Hammer and T. Villmann. Generalized relevance learning vector quantization. Neural Networks, 15(8-9):10591068, 2002.

[4] R. Ketterlinus, S-Y. Hsieh, S-H. Teng, H. Lee, and W. Pusch. Fishing for biomarkers: analyzing mass spectrometry data with the new clinprotools software. Bio techniques, 38(6):3740, 2005.

[5] T. Kohonen. Self-Organizing Maps, volume 30 of Springer Series in Information Sciences. Springer, Berlin, Heidelberg, 1995. (2nd Ext. Ed. 1997).

[6] J. Villanueva, J. Philip, D. Entenberg, and C.A. Chaparro et al. Serum peptide profiling by magnetic particle-assisted, automated sample processing and maldi-tof mass spectrometry. Anal. Chem., 76:15601570, 2004.

[7] F.-M. Schleif, T. Elssner, M. Kostrzewa, T. Villmann, and B. Hammer. Analysis and visualization of proteomic data by fuzzy labeled self organizing maps. In Proceedings of CBMS 2006, pages 919924. IEEE press, 2006.

[8] F.-M. Schleif, T. Villmann, and B. Hammer. Prototype based fuzzy classification in clinical proteomics. Special issue of International Journal of Approximate Reasoning on Approximate reasoning and Machine learning for Bioinformatics, page to appear, 2006

[9] F.-M. Schleif, T. Villmann, and B. Hammer. Supervised Neural Gas for Classification of Functional Data and its Application to the Analysis of Clinical Proteom Spectra. IWANN 2007, in press.

[10] T. Villmann, F.-M. Schleif, and B. Hammer. Comparison of Relevance Learning Vector Quantization with other Metric Adaptive Classification Methods. Neural Networks, 19:610622, 2006.

[11] P. Schneider, M. Biehl, and B. Hammer. Relevance Matrices in LVQ. To appear in Proc. ESANN'2007, in press (2007)

[12] B. Hammer, F.-M. Schleif, and T. Villmann. On the Generalization Ability of Prototype-Based Classifiers with Local Relevance Determination, Technical Report, Clausthal University of Technology, Ifi-05-14, 2005.

[13] M. Verleysen, D. Francoisand G. Simon, and V. Wertz. On the effects of dimensionality on data analysis with neural networks. Springer-Verlag. Artificial Neural Nets Problem solving methods, Lecture Notes in Computer Science 2687,pagesII105–II112, 2003.

[14] M. Biehl, B. Hammer, and P. Schneider. Matrix Learning in Learning Vector Quantization, Technical Report, Insitute of Informatics, Clausthal University of Technology, 2006.