Generalized Matrix Learning Vector Quantizer for the Analysis of Spectral Data

Petra Schneider[2], **Frank-Michael Schleif**[1], **Thomas Villmann**[1] and **Michael Biehl**[2]

1- University Leipzig, Dept. of Medicine, Semmelweisstrasse 10, 04103 Leipzig, Germany
2 - University of Groningen, Dept. of Math. & C.S., P.O. Box 407, 9700 AK Groningen, NL

Abstract. The analysis of spectral data constitutes new challenges for machine learning algorithms due to the functional nature of the data. Special attention is given to the used metric in such analysis. Recently a prototype based algorithm has been proposed which allows the integration of a full adaptive matrix in the metric. In this contribution we analyse this approach with respect to band matrices and its usage for the analysis of functional spectral data. The approach is tested on satellite data and data taken from food chemistry.

Keywords: GMLVQ, spectral data, band matrices, adaptive metric, functional data

1 Introduction

The analysis of high dimensional functional data is a common task in different fields of natural sciences like medicine and chemistry. The initial results of an experimental setting are typically given as functional spectral data. Prominent examples are mass spectrometric data (MS) in the field of clinical proteomics, nuclear magnetic resonance spectra (NMR) in the field of chemistry and metabolomics or satellite remote sensing spectroscopy in astronomy to name just a few. Focusing on classification, prototype based classification approaches such as Learning Vector Quantization (LVQ) as proposed by Kohonen [4] or multiple extensions [2, 1] have already proven to be valuable for the analysis of high dimensional data (see [10, 11]). Due to the complexity of the data the use of an appropriate distance measure is of special importance [5] to get an adequate representation of the data. In Generalized Relevance LVQ (GRLVQ) [3], the euclidean metric is replaced by a more powerful alternative, which introduces weights for the different features (scaled euclidean metric). This allows to scale the axes of the coordinate system of the data space in order to obtain better adaptation towards clusters with axes-parallel ellipsoidal shapes. In the previously published approaches to analyse such data, correlative effects between different features are ignored in general. For functional data correlative effects between neighbored data points are frequent and the order of the features is not any longer arbitrary.

The recently introduced Generalized Matrix LVQ (GMLVQ) [6, 7] adapts a full matrix of relevance factors in the distance measure and allows to take correlations between different features into account. Yet, full adaptive GMLVQ may suffer from to many adjustable parameters which is quadratic with the number of input dimensions. This can lead to instabilities and overfitting. In spectral data usually local correlations occur, such that GMLVQ can be restricted to band-limited matrices without significant loss of information.

In this paper we analyze this modifications for two different data sets from satellite remote sensing and food chemistry studies.

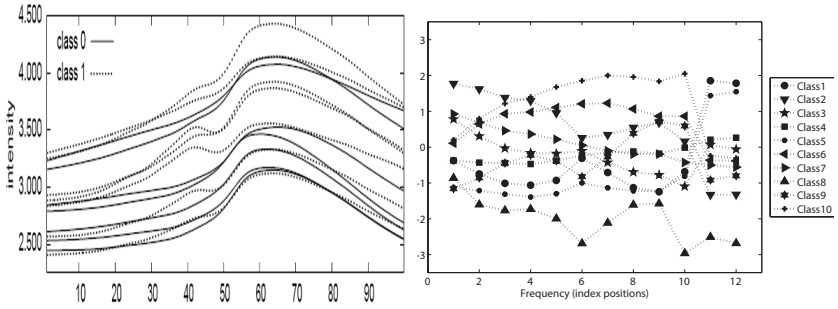

Fig. 1: Left: Plot of multiple spectra taken from the tecator data set. Spectra from class 0 (low fat) are plotted with straight lines and spectra of class 1 (high fat) are plotted with dashed lines. One clearly observes a visual separation of the spectra at a frequency around index 41, but it can also be seen that the single dimension 41 is not separating well between the two classes. Right: Average spectra for all classes of the satellite data set. The class labels are indicated by different point symbols.

2 Dataset

The first data set is a multispectral data set in 10 classes, obtained from [8] with 12 bands measured by a M-7 airborn-scanner. It contains a significant number of vegetative species or ground cover classes. The spectral bands cover from the visible to the near infrared: $0.40\mu m - 1.00\mu m$. The visible range mainly judges leaf pigments (chlorophyll) and the infrared range is most responsible for cell structures (spongy-mesophyll cells). Figure 1 (right) visualizes the mean spectra of the different classes. A detailed description is available in [8].

The second data set, publicly available at http://lib.stat.cmu.edu/datasets/tecator, contains 215 samples of 100-dimensional infrared absorbance spectra. The classification task consists in the predict of the binary fat content (low/high), of meat.

Figure 1 (left), shows example spectra of both classes. Apart from a tendency towards dints around channel 41 for high fat content, a substantial visual data overlap can be stated.

3 Generalized Matrix LVQ

LVQ aims at approximating a clustering by prototypes. Assume training data $(\xi_i, y_i) \in \mathbb{R}^N \times \{1, \ldots, C\}$ are given, N denoting the data dimensionality and C the number of different classes. A LVQ network consists of a number of prototypes which are characterized by their location in the weight space $\mathbf{w}_i \in \mathbb{R}^N$ and their class label $c(\mathbf{w}_i) \in \{1, \ldots, C\}$. Classification takes place by a winner takes all scheme. For this purpose, a (possibly parameterized) similarity measure d^λ is fixed for $\mathbb{R}^N$. Often, the standard euclidean metric is chosen. A data point $\xi \in \mathbb{R}^N$ is mapped to the class label $c(\xi) = c(\mathbf{w}_i)$ of the prototype i for which $d^\lambda(\mathbf{w}_i, \xi) \leq d^\lambda(\mathbf{w}_j, \xi)$ holds for every $j \neq i$ (breaking ties arbitrarily).

Learning aims at determining weight locations for the prototypes such that the given training data are mapped to their corresponding class labels. A very flexible learning approach has been introduced in [12]. It is derived as a minimization of the cost function

$$\sum_i \Phi\left(\frac{d_J^\lambda - d_K^\lambda}{d_J^\lambda + d_K^\lambda}\right) \tag{1}$$

where Φ is a monotonic function, e.g. the identity or the logistic function, $d_J^\lambda = d^\lambda(\mathbf{w}_J, \xi_i)$ is the distance of data point ξ_i from the closest prototype $\mathbf{w}_J$ with the same class label y_i, and $d_K^\lambda = d^\lambda(\mathbf{w}_K, \xi_i)$ is the distance from the closest prototype $\mathbf{w}_K$ with a different class label than y_i. Taking the derivatives with respect to the prototypes and metric parameters yields the adaptation rules.

The choice of the similarity measure as standard euclidean metric yields GLVQ. The squared *weighted* euclidean metric $d^\lambda(\mathbf{w}, \xi) = \sum_i \lambda_i (w_i - \xi_i)^2$ where $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$ constitutes a powerful alternative, GRLVQ, particularly suited for high dimensional data with a different (but priorly not known) relevance of the input dimensions.

In GMLVQ, a full matrix, which can account for arbitrary correlations of the dimensions, is used. The metric has the form

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

where Λ is a full matrix. Arbitrary euclidean metrics can be achieved by an appropriate choice of the parameters. The above similarity measure only leads to a squared distance if Λ is positive (semi-) definite. We can achieve this by substituting $\Lambda = \Omega\,\Omega^T$. As Λ is symmetric, we can assume that Ω itself is symmetric: $\Omega = \Omega^T$. To obtain the adaptation formulas we need to compute the derivatives of (1) with respect to $\mathbf{w}$ and Λ. We get the updates

$$\begin{aligned}
\Delta\mathbf{w}_J &= +\epsilon \cdot \phi'(\mu(\xi)) \cdot \mu^+(\xi) \cdot \Omega\Omega \cdot (\xi - \mathbf{w}_J) \\
\Delta\mathbf{w}_K &= -\epsilon \cdot \phi'(\mu(\xi)) \cdot \mu^-(\xi) \cdot \Omega\Omega \cdot (\xi - \mathbf{w}_K) \\
\Delta\Omega_{lm} &= -\epsilon \cdot \phi'(\mu(\xi)) \cdot \\
&\quad \Big(\mu^+(\xi) \cdot \Big([\Omega(\xi - \mathbf{w}_J)]_m (\xi_l - w_{J,l}) + [\Omega(\xi - \mathbf{w}_J)]_l (\xi_m - w_{J,m}) \Big) \\
&\quad -\mu^-(\xi) \cdot \Big([\Omega(\xi - \mathbf{w}_K)]_m (\xi_l - w_{K,l}) + [\Omega(\xi - \mathbf{w}_K)]_l (\xi_m - w_{K,m}) \Big) \Big)
\end{aligned}$$

for the prototypes and matrix elements Ω_{lm} with $\mu(\xi) = (d_J^\Lambda - d_K^\Lambda)/(d_J^\Lambda + d_K^\Lambda)$, $\mu^+(\xi) = 2 \cdot d_K^\Lambda/(d_J^\Lambda + d_K^\Lambda)^2$, and $\mu^-(\xi) = 2 \cdot d_J^\Lambda/(d_J^\Lambda + d_K^\Lambda)^2$. (See [6] for the derivation of these formulas.) Thereby, the learning rate for the metric can be chosen independently of the learning rate for the prototypes. Note that Ω is symmetric because these updates are symmetric. After each update, Λ is normalized to prevent the algorithm from degeneration. We set $\sum_i \Lambda_{ii} = \sum_{i,j} \Omega_{ij}^2 = 1$ which fixes the sum of diagonal elements and, here, the sum of eigenvalues. Band limited GMLVQ can be achieved by symmetric limiting the off-diagonals of Ω or Λ, respectively. This restriction leads to a focus on locally correlated frequency bands in spectral data. The width should be in correspondence to the correlation range in the spectra, which is problem specific. If k off-diagonals on both sides of the main diagonal are considered in Ω the respective band-width including the main diagonal is given as $n = 2 \cdot k + 1$ in Ω. We refer to this as GMLVQ-n.

4 Experiments and Results

We applied GMLVQ on both spectral data sets using different bandwidth settings and compared the classification performance with known results taken from [9, 8]. In the calculations for the tecator data, the spectra have been portioned randomly into $4/5$ for training and $1/5$ for test patterns averaged in a 5-fold crossvalidation. The satellite data where given in a predefined splitting into training and test.

For the satellite data set we determined the optimum bandwidth by optimizing the problem using 1 prototype for each class. An initial phase of 10 cycles of pure prototype training was mandatory before the adaptation of metric parameters. The learning was continued up to converges with an upper limit of 100 cycles.

We found a significantly increased accuracy for a bandwidth of 5 compared to a diagonal matrix (bandwidth 1). Whereas larger bandwidth does not achieve a significant improvement anymore. Hence, it is possible to reduce the number of free parameters without a degradation of classification performance. The relevant results are collected in Table 4 and Figure 2. These results are in good agreement with priorly findings published in [8]. This is based on incorporation of additional expert knowledge about relevant spectral frequencies for vegetation discrimination. In particular both visible and infrared frequencies contribute to the identification. A band width of 5 in our data set comprise at least parts of the visible and near-infrared spectrum. Therefore the respective correlations are taken into account. Smaller bandwidths lead to a loss of this correlation information, whereas larger ones, larger than 5, give no significant information gain.

For this optimum 5-band case the experiment was repeated using 5 prototypes per class. We achieved a prediction of 86.4% which is comparable to the result given in [8] with 91% using only 4 features. This selection was done such that the features were almost independent, but covering visible and infrared frequencies.

The main diagonal elements of matrix Λ (relevance profile), reflect that red and infra-red frequencies are especially relevant for the classification. This underlies the above mentioned features of chloro- and mesophyll level for vegetation discrimination.

For the tecator data set pretraining the prototypes using 1 prototype per class and with the euclidean metric was done for 20 cycles. The learning was continued up to converges with an upper limit of 200 cycles. In our experiments we found a bandwidth of 21 to be critical to achieve the classification performance of the full matrix compare Table 4 and Figure 2.

In Figure 3 the relevance profile for this case is depicted. We see that a region around index 41 is ranked most whereby a detailed analysis of the relevance matrix also shows that correlated dimensions are highly ranked as well. In comparison to the visual impression given in Figure 1 this is a plausible result which has not been found by simple GRLVQ using a scaled euclidean metric. This can be related to the fact that the discrimination power is not due to a single dimension but rather to a neighborhood effect.

5 Conclusion

In this article band-limited GMLVQ has been investigated for classification of spectral data. For both considered data sets one observed an overall improvement in prediction, compared to simple GRLVQ as it is also observed for full GMLVQ. However band-limitation can successfully be applied without significant information loss. The

Satellite		Tecator	
Algorithm	Prediction	Algorithm	Prediction
GMLVQ-1	78.5%	GMLVQ-1	59.5%
GMLVQ-3	82.9%	GMLVQ-3	59.5%
GMLVQ-5	86.2%	GMLVQ-11	78.6%
GMLVQ-7	86.3%	GMLVQ-21	92.9%
GMLVQ-9	86.6%	GMLVQ-31	92.9%
GMLVQ-11	86.4%	GMLVQ-41	90.5%
GMLVQ-F	86.6%	GMLVQF	95.2%
SVM-RBF	70.7%	SVM-RBF	68.9%
SVM-Lin	85.3%	SVM-Lin	73.3%
C-GRLVQ	n.a.	C-GRLVQ	97%

Table 1: Classification accuracies for the satellite and the tecator data set using different band width settings for GMLVQ (0 to F-full) in comparison to correlation based GRLVQ (C-GRLVQ) with 25 prototypes [9] and two types of a SVM (Lin-linear, RBF-radial basis function kernel) obtained using Yale (http://yale.cs.uni-dortmund.de)

obtained optimum bandwidths can be discussed in the light of spectra properties of the underlying problems. Thus band-limiting can be used to reduce the number of adjustable parameters of standard GMLVQ to improve the stability. These findings rise hope that this results may hold also for other kinds of spectral data such as mass spectra (MS) or Ion Mobility Spectroscopy (IMS) which is an important analysis technique in chemistry and the field of security.

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 423-9. 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):21-44, February 2005.

[3] B. Hammer and T. Villmann. Generalized relevance learning vector quantization. Neural Networks, 15(8-9):1059-1068, 2002.

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

[5] T. Villmann, F.-M. Schleif, and B. Hammer. Comparison of Relevance Learning Vector Quantization with other Metric Adaptive Classification Methods. Neural Networks, 19:610–622, 2006.

[6] P. Schneider, M. Biehl, and B. Hammer. Relevance Matrices in LVQ. In Proc. of ESANN 2007, pages 37–42, Bruges, Belgium, April 2007.

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

[8] D. Landgrebe. Signal Theory Methods in Multispectral remote sensing. Wiley, New Jersey, 2003.

[9] M. Strickert, N. Sreenivasulu, W. Weschke, T. Villmann and B. Hammer. eneralized Relevance LVQ (GRLVQ) with Correlation measures for Gene Expression analysis. Neurocomputing, 69(7-9):651-659, 2006.

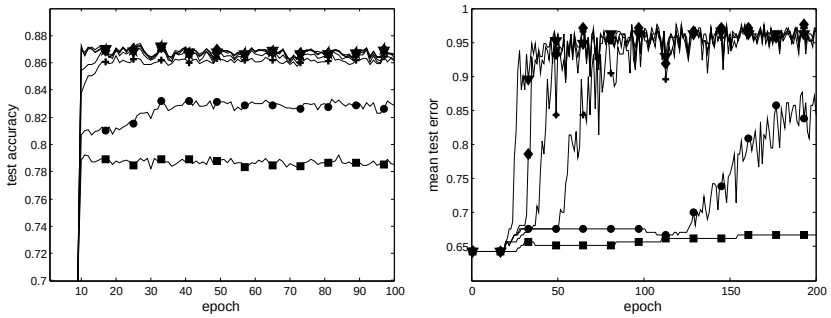

Fig. 2: The figure shows the process of the prediction accuracies for the different data sets using matrices of different bandwidth settings. Left: Satellite data set, Right: Tecator data set. Curves generated by GMLVQ with very small bandwidth show worse prediction accuracies. A clear gap between the prediction using no-bands (GRLVQ - curves with □) and few bands can be found. The final predictions depicted in the curves are collect in Table 4. Its also observable that a full matrix gives only a slight additional effect on the classifier performance.

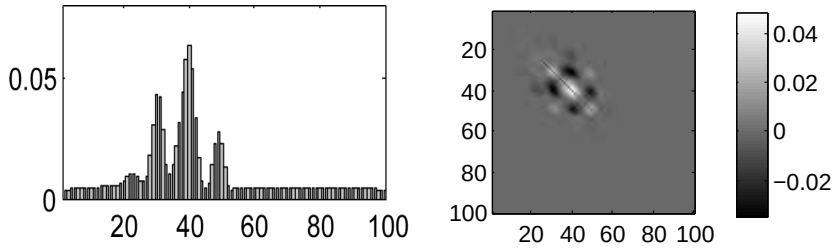

Fig. 3: Relevance profile for GMLVQ using 21 bands for the tecator data set. The x-axis shows the frequency index and the y-axis the relevance. For the tecator data the most relevant region is found around feature 41. In the right plot the off-diagonal elements of the corresponding Λ matrix are depicted with zero diagonal elements.

[10] 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 919-924. IEEE press, 2006.

[11] 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, 47(1):4-16,2008

[12] B. Hammer, M. Strickert, and T. Villmann. Supervised neural gas with general similarity measure. *Neural Processing Letters* 21(1): 21–44, 2005.