Differentiable Kernels in Generalized Matrix Learning Vector Quantization

Kästner, M.; Nebel, D.; Riedel, M.; Biehl, M.; Villmann, T.

Published in:
Machine Learning and Applications (ICMLA), 2012 11th Conference on

DOI:
10.1109/ICMLA.2012.231

Differentiable Kernels in Generalized Matrix Learning Vector Quantization

M. Kästner, D. Nebel, M. Riedel, M. Biehl, and T. Villmann

***Abstract*—In the present paper we investigate the application of differentiable kernel for generalized matrix learning vector quantization as an alternative kernel-based classifier, which additionally provides classification dependent data visualization. We show that the concept of differentiable kernels allows a prototype description in the data space but equipped with the kernel metric. Moreover, using the visualization properties of the original matrix learning vector quantization we are able to optimize the class visualization by inherent visualization mapping learning also in this new kernel-metric data space.**

I. Introduction

Prototype based vector quantization provides a powerful concept for unsupervised and supervised data analysis and processing. Prominent examples for unsupervised models applied in data clustering or visualization are the self-organizing map (SOM,[16]), neural gas (NG, [17]) as a robust version of the k-means or fuzzy variants like fuzzy-c-means (FCM, [4], [5]) and alternatives thereof. Supervised prototype based approaches for classification tasks are mainly influenced by the learning vector quantization models (LVQ, [16]) and support vector machines (SVM,[27]). Whereas LVQ models generate class typical prototypes, SVMs determine prototypes (support vectors) defining the class borders. Both paradigms belong to the class of margin classifiers [10]. An important feature considered in the last years is the application of non-standard metrics for these models to improve the classifier performance for domain specific problems like processing of functional data, e.g. spectra, time series [15], [20], [32] or better interpretability of the adapted models (relevance and matrix learning, [13], [29]). In particular, matrix learning in the generalized LVQ model (GLVQ, [25]) provides a great flexibility, robustness and classification performance in many applications as well as excellent class visualization abilities [6], [9], [8].

One of the most powerful concepts in classification remains the idea of kernel mapping realized in SVMs. According to this idea, the data as well as the prototypes are formally mapped into a high-dimensional (infinite) feature mapping Hilbert space (FMHS), which offers frequently a great flexibility and good separation possibility. The mapping is uniquely determined by the kernel, delivering an inner product in the FMHS. Yet, this mapping is done only implicitly. It turns out that the data handling can be processed without application of the explicit mapping by use of the kernel properties. This advantage on the one hand side, however, makes it more difficult to interpret the model on the other hand, because the prototypes in these models are given as infinite-dimensional representations in the FMHS. Moreover, the SVM prototypes are not typical representatives of the classes, as mentioned before. During the last years, several variants of LVQ were developed to integrate the kernel mapping idea in those models but keeping the idea of class-typical prototypes (Kernel GLVQ, KGLVQ) [26], [24], [23]. Yet, these models also have to deal with the problem of the infinite representation of prototypes. Usually, the infinite representation is approximated by a finite one using the Nystrøm-approximation approach, which obviously leads to an information loss in general.

Recently, an interesting alternative was proposed: If the kernel function is required to be differentiable in addition to the usual kernel properties, then the mapping into the FMHS can be avoided. Instead, the data as well as the prototypes are treated in the original data space but equipped with the kernel induced metric [33]. It turns out that KGLVQ with differentiable kernels (DK-GLVQ) are principle alternatives to SVMs with comparable performance [34]. In this article we combine this DK-GLVQ with the idea of matrix learning in GLVQ. We show that a further improvement for classification accuracy can be achieved. Further, we demonstrate the visualization skills provided by the matrix learning concept in this framework.

In the following, first we will briefly review the matrix learning in GLVQ. After this we consider the basic principle of DK-GLVQ and discuss the combination with the matrix learning idea. We demonstrate the power of this approach for three data sets, two of them from the standard UCI database. The third one is a real world application in food industry. For the latter application we also show the visualization abilities of the model.

II. Learning Vector Quantization and Matrix Learning

For LVQ we suppose that the data are given as vectors $\mathbf{v} \in V \subseteq \mathbb{R}^n$, and the prototypes of the LVQ model are subsumed in the set $W = \{\mathbf{w}_k \in \mathbb{R}^n, k = 1 \ldots M\}$. Each data vector $\mathbf{v}$ of the training data belongs to a class $x_{\mathbf{v}} \in \mathcal{C} = \{1, \ldots, C\}$. The prototypes are also equipped with

M. Kästner, D. Nebel, M. Riedel, and T. Villmann are with the Computational Intelligence Group at the Department for Mathematics/Natural & Computer Sciences, University of Applied Sciences Mittweida, Mittweida, Germany (email: {kaestner,nebel,riedel,villmann}@hs-mittweida.de). M. Biehl is with the Intelligent Systems Group at the Bernoulli Institute for Computer Sciences and Mathematics, University Groningen, The Netherlands (email: m.biehl@rug.nl).

978-0-7695-4913-2/12 $26.00 © 2012 IEEE
DOI 10.1109/ICMLA.2012.231

labels $y_{\mathbf{w}_k} \in \mathcal{C}$ indicating their responsibility to the several classes. The relation between data and prototypes is judged by a dissimilarity $d(\mathbf{v}, \mathbf{w}) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$ given in the data space V and frequently chosen as the (quadratic) Euclidean metric. Yet, the dissimilarity measure has not necessarily to be a distance. At least, it has to fulfill the requirements of a dissimilarity measure [21] with the additional constraint of differentiability in the second argument.

Standard LVQ distributes the prototypes in such a way that the classification error is heuristically optimized [16]. The generalization thereof, the generalized LVQ (GLVQ, [25]) minimizes an approximated classification error based on a stochastic gradient descent scheme [25]. The cost function minimized by GLVQ is

$$E(W) = \frac{1}{2} \sum_{\mathbf{v} \in V} f(\mu(\mathbf{v})), \tag{1}$$

where the function

$$\mu(\mathbf{v}) = \frac{d^+(\mathbf{v}) - d^-(\mathbf{v})}{d^+(\mathbf{v}) + d^-(\mathbf{v})} \tag{2}$$

is the classifier function with $d^+(\mathbf{v}) = d(\mathbf{v}, \mathbf{w}^+)$ denotes the dissimilarity between the data vector $\mathbf{v}$ and the closest prototype $\mathbf{w}^+$ with the same class label $y_{\mathbf{w}^+} = x_{\mathbf{v}}$, and $d^-(\mathbf{v}) = d(\mathbf{v}, \mathbf{w}^-)$ is the dissimilarity degree for the best matching prototype $\mathbf{w}^-$ with a class label $y_{\mathbf{w}^-}$ different from $x_{\mathbf{v}}$. The transformation function f is a monotonically increasing function usually chosen as sigmoidal or the identity function. A typical sigmoidal choice is the Fermi function

$$f(x) = \frac{1}{1 + a \cdot \exp\left(-\frac{(x - x_0)^2}{2\varsigma^2}\right)} \tag{3}$$

with $x_0 = 0$ and $a = 1$ as standard parameter values. The ς-parameter controls the approximation of the classification error: For $\varsigma \to 0$ the function $E(W)$ converges to the negative classification error. Learning of $\mathbf{w}^+$ and $\mathbf{w}^-$ is performed in GLVQ by the stochastic gradient of the cost function $E(W)$ for a given data vector $\mathbf{v}$ according to

$$\frac{\partial E(W)}{\partial \mathbf{w}^+} = \xi^+ \cdot \frac{\partial d^+(\mathbf{v})}{\partial \mathbf{w}^+} \text{ and } \frac{\partial E(W)}{\partial \mathbf{w}^-} = \xi^- \cdot \frac{\partial d^-(\mathbf{v})}{\partial \mathbf{w}^-} \tag{4}$$

with

$$\xi^+ = f' \cdot \frac{2 \cdot d^-(\mathbf{v})}{\left(d^+(\mathbf{v}) + d^-(\mathbf{v})\right)^2} \tag{5}$$

and

$$\xi^- = -f' \cdot \frac{2 \cdot d^+(\mathbf{v})}{\left(d^+(\mathbf{v}) + d^-(\mathbf{v})\right)^2}. \tag{6}$$

For the quadratic Euclidean metric we simply get the derivatives

$$\frac{\partial d^\pm(\mathbf{v})}{\partial \mathbf{w}^\pm} = -2\left(\mathbf{v} - \mathbf{w}^\pm\right)$$

realizing a vector shift of the prototypes.

In matrix learning GLVQ (GMLVQ, [29]) the (quadratic) Euclidean distance is replaced by a quadratic form

$$d_\Lambda(\mathbf{v}, \mathbf{w}) = (\mathbf{v} - \mathbf{w})^\top \Lambda (\mathbf{v} - \mathbf{w}) \tag{7}$$

with $\Lambda = \Omega^\top \Omega$ to ensure the positive definiteness. Using its factorization, eq. (7) can be written in the form

$$d_\Omega(\mathbf{v}, \mathbf{w}) = (\Omega(\mathbf{v} - \mathbf{w}))^2 \tag{8}$$

with an arbitrary matrix $\Omega \in \mathbb{R}^{m \times n}$, i.e. the data and the prototypes are mapped into the $\mathbb{R}^m$ and afterward the quadratic Euclidean norm is calculated. The resulting derivative $\frac{\partial d_\Omega^\pm(\mathbf{v})}{\partial \mathbf{w}^\pm}$ for the prototype update in (4) is obtained as

$$\frac{\partial d_\Omega^\pm(\mathbf{v})}{\partial \mathbf{w}^\pm} = -2\Lambda\left(\mathbf{v} - \mathbf{w}^\pm\right).$$

Metric learning takes place in this GMLVQ by the Ω-update, again realized as a stochastic gradient descent:

$$\frac{\partial E(W)}{\partial \Omega_{r_1, r_2}} = \xi^+ \cdot \frac{\partial d_\Omega^+(\mathbf{v})}{\partial \Omega_{r_1, r_2}} + \xi^- \cdot \frac{\partial d_\Omega^-(\mathbf{v})}{\partial \Omega_{r_1, r_2}} \tag{9}$$

where $\frac{\partial d_\Omega^\pm(\mathbf{v})}{\partial \Omega_{r_1, r_2}}$ follows the relation

$$\frac{\partial d_\Omega(\mathbf{v}, \mathbf{w})}{\partial \Omega_{r_1, r_2}} = 2\left[\mathbf{\Omega}(\mathbf{v} - \mathbf{w})\right]_{r_1} \left[\mathbf{v} - \mathbf{w}\right]_{r_2} \tag{10}$$

and a subsequent renormalization has to take place to ensure $\sum_{i,j} \Omega_{i,j}^2 = \sum_i \Lambda_{i,i} = 1$ after completing the adjustment. We explicitly remark here that m does not need to be equal to n. In particular, $m < n$ is an option for inherent regularization and class visualization ($m = 2, 3$) [9], [28]. Further, in GMLVQ the matrix Λ can be interpreted as a correlation matrix determining the correlations between the data dimensions, which are useful for classification [29]. Even for $m = n$ the algorithm shows inherent regularization because the Ω-adjustment in GMLVQ via (9) can be related to class dependent principal component analysis, such that the learned matrices tend to be generated by the class eigenvectors [6], [7].

III. Learning Vector Quantization Using Differentiable Kernel

The idea of kernel mapping has a long tradition. Beginning with the theoretic work of Aronzaijn and Mercer about positive kernels and dedicated reproducing kernel Hilbert spaces (RKHS) a broad mathematical framework was established, which can be used to generate powerful machine learning algorithms [1], [18]. One of the most popular schemes in the context of classification tasks are *support vector machines* (SVM, [27]).

The basic idea in this model is the data are implicitly mapped into a high-dimensional (maybe infinite) feature space by a so-called kernel mapping, to become easily separable there. To be precisely, the data space V is assumed to be compact equipped with metric $d_V : V \times V \to \mathbb{R}^+$. We denote this metric space by (V, d_V). A function κ on V is a positive kernel

$$\kappa_\Phi : V \times V \to \mathbb{C} \tag{11}$$

if there exists a Hilbert space $\mathcal{H}$ and a feature map

$$\Phi : (V, d_V) \longrightarrow \mathcal{I}_{\kappa_\Phi} \subseteq \mathcal{H} \tag{12}$$

where $\mathcal{I}_{\kappa_\Phi}$ is the image of V under the mapping Φ. In $\mathcal{I}_{\kappa_\Phi}$ an inner product is determined by the kernel

$$\kappa_\Phi(\mathbf{v},\mathbf{w}) = \langle \Phi(\mathbf{v}), \Phi(\mathbf{w}) \rangle_{\mathcal{H}} \tag{13}$$

for all $\mathbf{v},\mathbf{w} \in V$ and $\langle \cdot,\cdot \rangle_{\mathcal{H}}$ is the inner product of the Hilbert space defining the metric

$$d_{\mathcal{H}}\left(\Phi(\mathbf{v}), \Phi(\mathbf{w})\right) = \sqrt{\kappa_\Phi(\mathbf{v},\mathbf{v}) - 2\kappa_\Phi(\mathbf{v},\mathbf{w}) + \kappa_\Phi(\mathbf{w},\mathbf{w})} \tag{14}$$

in $\mathcal{I}_{\kappa_\Phi}$. The positive definiteness of the kernel ensures the metric properties, whereas for general kernels (14) is only a semi-metric [1]. The map Φ is injective iff the kernel is universal [31].

So far these kernels can be used in GLVQ only using approximation techniques like the Nystrøm-approximation to deal with the problem of the infinite-dimensional representation of prototypes in $\mathcal{I}_{\kappa_\Phi}$ [26], [24], [23], which obviously leads to an information loss in general. To overcome this unsatisfying situation we consider another bijective map

$$\Psi : (V, d_V) \longrightarrow (V, d_{\kappa_\Phi}) \tag{15}$$

for universal kernels, such that the vector space objects are preserved where the dissimilarity measure d_{κ_Φ} is determined by $d_{\kappa_\Phi}(\mathbf{v},\mathbf{w}) = d_{\mathcal{H}}\left(\Phi(\mathbf{v}), \Phi(\mathbf{w})\right)$. It turns out that the map Ψ is continuous iff Φ does [31], and, therefore, it is injective because of the bijectivity. It can be easily shown that the metric space $\mathcal{V}_{\kappa_\Phi} = (V, d_{\kappa_\Phi})$ is isometric-isomorph to $\mathcal{I}_{\kappa_\Phi}$ [33]. Now, we further assume that the kernel $\kappa_\Phi(\mathbf{v},\mathbf{w})$ is differentiable at least with respect to the second argument $\mathbf{w}$. In this case we obtain

$$\frac{\partial d^2_{\kappa_\Phi}(\mathbf{v},\mathbf{w})}{\partial \mathbf{w}} = \frac{\partial \kappa_\Phi(\mathbf{w},\mathbf{w})}{\partial \mathbf{w}} - 2\frac{\partial \kappa_\Phi(\mathbf{v},\mathbf{w})}{\partial \mathbf{w}}, \tag{16}$$

which can immediately plugged into gradient based prototype adaptation (4) of GLVQ replacing any other metric. Hence, the prototypes belong itself to $\mathcal{V}_{\kappa_\Phi}$ and remain finite-dimensional as the original data vectors without any loss of information and structure compared to $\mathcal{I}_{\kappa_\Phi}$. We refer to this variant as *differentiable kernel* GLVQ (DK-GLVQ).

IV. Differentiable Kernel and Matrix Learning in GMLVQ

One of the most famous and well-known examples of universal positive kernels is the Gaussian kernel

$$\Gamma_\Phi\left(\mathbf{v},\mathbf{w}\right) = \exp\left(-\left(\frac{\mathbf{v}-\mathbf{w}}{\sqrt{2}\sigma}\right)^2\right) \tag{17}$$

with the width $\sigma > 0$. Obviously, it is differentiable, and, therefore could be used in the space $\mathcal{V}_{\kappa_\Phi} = (V, d_{\kappa_\Phi})$ for gradient based vector quantization like DK-GLVQ. Thereby, the optimization of the kernel width σ can be subject of adaptation by respective gradient learning in DK-GLVQ. However, determination of an optimum kernel width is frequently sensitive and unstable as it is known from radial basis function learning [14].

We remark that the term $\vartheta = \left(\frac{\mathbf{v}-\mathbf{w}}{\sqrt{2}\sigma}\right)^2$ in $\Gamma_\Phi\left(\mathbf{v},\mathbf{w}\right)$ is a scaled quadratic Euclidean distance. It turns out that $\Gamma_\Phi\left(\mathbf{v},\mathbf{w}\right)$ remains universal if ϑ is replaced by another (quadratic) metric [19]. Hence, we can combine the idea of matrix learning with DK-GLVQ. In particular, we consider the kernel

$$\Gamma_\Phi\left(\mathbf{v},\mathbf{w},\Omega\right) = \exp\left(-d_\Omega\left(\mathbf{v},\mathbf{w}\right)\right) \tag{18}$$

using the quadratic form (8) with derivatives

$$\frac{\partial \Gamma_\Phi\left(\mathbf{v},\mathbf{w},\Omega\right)}{\partial \mathbf{w}} = \Gamma_\Phi\left(\mathbf{v},\mathbf{w},\Omega\right) \cdot 2\Omega\left(\mathbf{v}-\mathbf{w}\right) \tag{19}$$

and

$$\frac{\partial \Gamma_\Phi\left(\mathbf{v},\mathbf{w},\Omega\right)}{\partial \mathbf{\Omega}_{\mathbf{r_1},\mathbf{r_2}}} = -2\Gamma_\Phi\left(\mathbf{v},\mathbf{w},\Omega\right) \cdot \left[\mathbf{\Omega}\left(\mathbf{v}-\mathbf{w}\right)\right]_{r_1}\left[\mathbf{v}-\mathbf{w}\right]_{r_2} \tag{20}$$

to be needed for prototype and matrix updates via (16) and (4) as well as (9). The optimization of the Ω-matrix again is self-regularizing as it is known from GMLVQ [6], [7] and therefore more stable than simple kernel-width learning.

We refer to this model as *differentiable kernel* GMLVQ (DK-GMLVQ).

Data sets	GLVQ	DK-GLVQ	GMLVQ		DK-GMLVQ	
			$m = 2$	$m = n$	m=2	m=n
PIMA	75.1 ± 0.028	76.2 ± 0.031	77.87 ± 0.016	77.74 ± 0.031	77.21 ± 0.008	78.26 ± 0.025
WDBC	93.49 ± 0.016	94.2 ± 0.010	94.48 ± 0.016	94.73 ± 0.016	95.60 ± 0.019	95.43 ± 0.025

Table I
TEST ACCURACIES [%] WITH STANDARD DEVIATION FOR THE TWO BENCHMARK UCI-DATA SETS OF THE DIFFERENT METHODS, EACH TRAINED WITH ONE PROTOTYPE PER CLASS.

V. Application of the DK-GMLVQ

We tested the DK-GMLVQ algorithm for several real world data sets. The first two data sets are from the well-known UCI database to be comparable with other investigations and approaches. The more challenging data set consists of hyperspectral vectors of different coffee types, referred as coffee data set. For this data set we specifically investigate the properties of the DK-GMLVQ compared to its non-kernelized counterpart GMLVQ.

A. Results for UCI data

In this first consideration we compare the DK-GMLVQ with other classification algorithms for two benchmark UCI-data sets: the Wisconsin-Breast-Cancer-data (WDBC) and the Indian diabetes data set (PIMA). The data sets contain 562 and 768 data vectors with 32 and 8 data dimensions, respectively, and each divided into two classes (healthy/ill). The presented results are obtained from a three-fold cross validation. For each simulation we used only one prototype per class. The results are depicted in Tab. V-A.

It turns out that the DK-GLVQ and the DK-GMLVQ outperform their counterparts GLVQ and GMLVQ, respectively. Hence, the non-linear mapping Ψ (15) provides the better class separation possibility as known from the underlying feature map Φ (12) of SVMs. Therefore, a SVM

was additionally trained for comparison for both problems: We applied the recently proposed Extreme Learning Kernel (ELM, [11]). The ELM kernel is actually a d-facto parameter free kernel with the same classification performance as the RBF-kernel with optimal Gaussian width. SVM models are obtained by use of a Sequential Minimization Optimization (SMO) optimizer as proposed and the ELM kernel [22]. The achieved SVM accuracies are $97.7\% \pm 1.45$ and $76.4\% \pm 4.2$ for WDBC and PIMA, respectively, which is comparable to DK-GLVQ and DK-GMLVQ. As it was already explained in [26], the numbers of support vectors is very high for both problems, i.e. 512 and 691.

For GMLVQ as well as DK-GMLVQ we also investigated the case that the matrix Ω in the distance measure (8) is of limited size $2 \times n$, which would be necessary for visualization. We observe that the accuracy loss is not dramatic for this scenario, i.e. a linear transformation of the data into the space $\mathbb{R}^2$ is possible while keeping the classification performance.

B. Results for Coffee Data

In this application we classified hyperspectral short-wave infrared range (SWIR) spectra of different coffee types (Bonga Forest - black, Ethiopia Sidamo Grande - green, Espresso Columbia - blue, Australia Skybury- magenta, Ganos Espresso Cuba - red). Hyperspectral processing along with an appropriate analysis of the acquired high-dimensional spectra has proven to be a suitable and very powerful method to quantitatively assess the biochemical composition of a wide range of biological samples [12], [30], [3]. By utilizing a hyperspectral camera (HySpex SWIR-320m-e, Norsk Elektro Optikk A/S) we obtained a rather extensive data base of spectra of five different coffee types (5000 spectra for each class). The acquired spectra are in the SWIR between 970 nm and $2,500$ nm at 6 nm resolution yielding 256 bands per spectrum. Proper image calibration was done by using a standard reflection pad (polytetrafluoroethylene, PTFE)[2]. After appropriate image segmentation the obtained spectra were normalized according to the l_2-norm. The mean spectra of the five types are visualized in Fig.1.

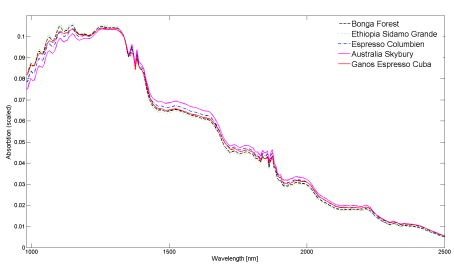

Figure 1. Mean spectra of the five investigated coffee types.

Again, we have taken only one prototype per class. For each class, 1000 spectra were randomly selected for training. The remaining spectra were applied for testing. The achieved test accuracies are displayed in Tab.V-B.

Datasets	GLVQ	DK-GLVQ	GMLVQ		DK-GMLVQ	
			$m = 2$	$m = n$	m=2	m=n
Coffee	83.29	80.0	88.51	88.97	90.84	91.38

Table II
TEST ACCURACIES [%] FOR THE COFFEE DATA SET OF THE DIFFERENT METHODS, EACH TRAINED WITH ONE PROTOTYPE PER CLASS.

Additionally, we investigated the obtained correlation matrices $\Lambda = \Omega^\top \Omega$ for the GMLVQ and the DK-GMLVQ, which are visualized in Fig.'s 2 and 3.

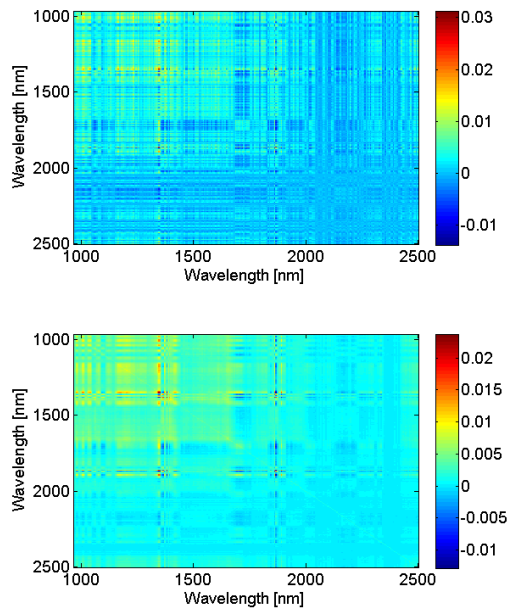

Figure 2. Correlation matrices $\Lambda = \Omega^\top \Omega$, $\Omega \in \mathbb{R}^{m \times n}$ for the coffee data set for GMLVQ with $m = 2$ (limited rank, top) and $m = n = 256$ (full-rank, bottom).

The learned matrices provide information about the correlations between the spectral bands useful for classification. We observe that the full-ranked GMLVQ leads to a smother correlation matrix. Yet, the essential structure information is preserved approximately in case of the limited rank matrix Ω. Further, the matrices can be used for classification dependent data visualization. In Fig. 4 the data are plotted by $\Omega \mathbf{v}$ according to (8) for the limited rank $\Omega \in \mathbb{R}^{2 \times n}$ and according to a principle component projection using the eigenvectors of Ω in case of the full rank $\Omega \in \mathbb{R}^{n \times n}$.

We compare these visualizations with a (unsupervised) principle component projection displayed in Fig. 5. The difference is obvious: In the unsupervised PCA-projection the classes are heavily overlapping whereas after classification optimized projection the classes are better separated in the visualization space. However, the classes are not completely separated. The black class (Ganos Espresso Cuba) in Fig. 4 is strongly overlapping with the green class (Bonga Forest).

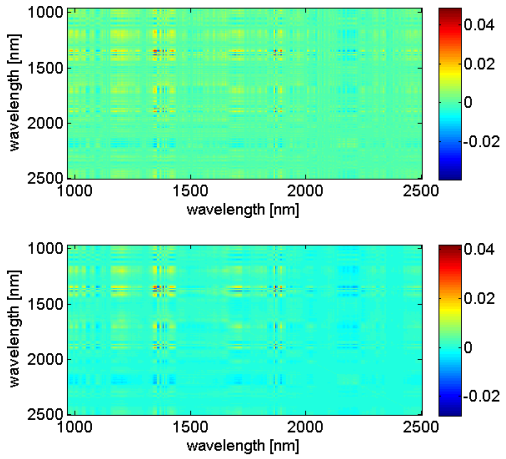

Figure 3. Correlation matrices $\Lambda = \Omega^\top \Omega$, $\Omega \in \mathbb{R}^{m \times n}$ for the coffee data set for DK-GMLVQ with $m = 2$ (limited rank, top) and $m = n = 256$ (full-rank, bottom).

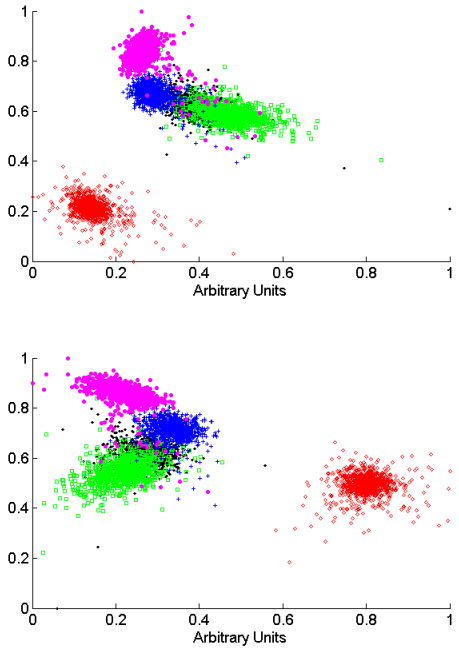

Figure 4. Data projection of coffee data by means of GMLVQ: top - projection of the data by $\Omega \mathbf{v}$ using the learned limited rank matrix $\Omega \in \mathbb{R}^{2 \times n}$; bottom - projection of the data according to the eigenanalysis of $\Lambda = \Omega^\top \Omega$ with full rank matrix $\Omega \in \mathbb{R}^{n \times n}$ and $n = 256$.

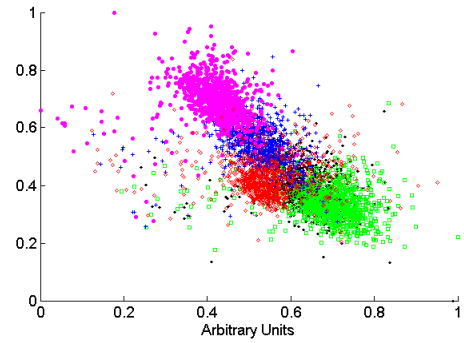

Figure 5. Data projection of coffee data according to an unsupervised principle component analysis.

A slight improvement is obtained if the DK-GMLVQ is used and the above visualization techniques are applied there. The respective data visualizations are depicted in Fig.6.

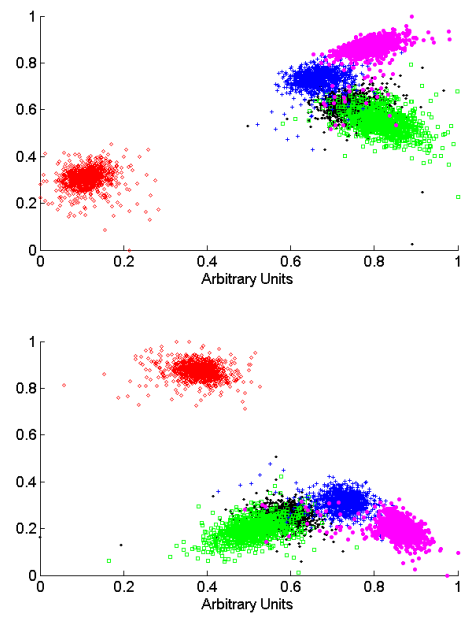

Figure 6. Data projection of coffee data based on DK-GMLVQ: top - projection of the data by $\Omega \mathbf{v}$ using the learned limited rank matrix $\Omega \in \mathbb{R}^{2 \times n}$; bottom - projection of the data according to the eigenanalysis of $\Lambda = \Omega^\top \Omega$ with full rank matrix $\Omega \in \mathbb{R}^{n \times n}$ and $n = 256$.

We observe a slightly improved separation between the black and the green class, however, far away from an optimal separation. However, this is in agreement with a consideration of the respective confusion matrices for both classifiers, see Tab. V-B. The classifiers are not able to separate the Bonga Forest (black) type from Ethiopia Sidamo Grande (green) with sufficient precision. Otherwise, we can conclude from the visualizations that the other coffee types can be clearly separated using the spectral information.

	coffee types	black	green	blue	magenta	red
GMLVQ	black	81.5	14.1	4.4	0.1	0
	green	26.0	73.0	1.0	0	0
	blue	5.7	0.2	92.9	1.2	0
	magenta	0.8	0.1	1.6	97.5	0
	red	0	0	0	0	100
DK-GMLVQ	black	82.8	15.5	1.5	0.1	0
	green	17.5	82.2	0.3	0	0
	blue	4.7	0.1	94.1	1.1	0
	magenta	0.7	0.1	1.5	97.7	0
	red	0	0	0	0	100

Table III
RELATIVE CONFUSION MATRIX [%] OF GMLVQ AND DK-GMLVQ

The application of differentiable kernel leads to a small improvement in confusion matrix as well as in the respective class visualizations. In particular, the green coffee type is significantly less frequent misclassified as a Bonga Forest type (black).

VI. CONCLUSION

In this contribution we consider the application of differentiable kernels in GMLVQ for classification and class visualization. Using this method we are able to obtain classification performances comparable to other widely applied classifiers including SVM. Moreover, using the visualization properties of the underlying GMLVQ we are able to optimize the class visualization also for this kind of kernel-based classification.

ACKNOWLEGMENT

Thanks for allocation of the coffe data set to Prof. Udo Seiffert (Fraunhofer IFF Magdeburg/Germany) and Ganos Kaffee-Rösterei Leipzig providing the coffee.

REFERENCES

[1] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[2] A. Backhaus, F. Bollenbeck, and U. Seiffert. High-throughput quality control of coffee varieties and blends by artificial neural networks and hyperspectral imaging. In *Proceedings of the 1st International Congress on Cocoa, Coffee and Tea, CoCoTea 2011*, page accepted for publication, 2011.

[3] A. Backhaus, F. Bollenbeck, and U. Seiffert. Robust classification of the nutrition state in crop plants by hyperspectral imaging and artificial neural networks. In *Proceedings of the 3rd IEEE Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing WHISPERS 2011*, page 9. IEEE Press, 2011.

[4] J. Bezdek. A convergence theorem for the fuzyy ISODATA clustering algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(1):1–8, 1980.

[5] J. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum, New York, 1981.

[6] M. Biehl, K. Bunte, F.-M. Schleif, P. Schneider, and T. Villmann. Large margin discriminative visualization by matrix relevance learning. In H. Abbass, D. Essam, and R. Sarker, editors, *Proc. of the International Joint Conference on Neural Networks (IJCNN), Brisbane*, pages 1873–1880, Los Alamitos, 2012. IEEE Computer Society Press.

[7] M. Biehl, B. Hammer, F.-M. Schleif, P. Schneider, and T. . Villmann. Stationarity of matrix relevance learning vector quantization. *Machine Learning Reports*, 3(MLR-01-2009):1–17, 2009. ISSN:1865-3960, http://www.uni-leipzig.de/compint/mlr/mlr$_01_2$009.*pdf*.

[8] K. Bunte, B. Hammer, A. Wismüller, and M. Biehl. Adaptive local dissimilarity measures for discriminative dimension reduction of labeled data. *Neurocomputing*, 73:1074–1092, 2010.

[9] K. Bunte, P. Schneider, B. Hammer, F.-M. S. T. Villmann, and M. Biehl. Limited rank matrix learning, discriminative dimension reduction and visualization. *Neural Networks*, 26(1):159–173, 2012.

[10] K. Crammer, R. Gilad-Bachrach, A.Navot, and A.Tishby. Margin analysis of the LVQ algorithm. In S. Becker, S. Thrun, and K. Obermayer, eds., *Advances in Neural Information Processing (Proc. NIPS 2002)*, volume 15, pages 462–469, Cambridge, MA, 2003. MIT Press.

[11] B. Frénay and M. Verleysen. Parameter-free kernel in extreme learning for non-linear support vector regression. *Neurocomputing*, 74(16):2526–2531, 2011.

[12] H. Grahn and P. Geladi, editors. *Techniques and Applications of Hyperspectral Image Analysis*. Wiley, 2007.

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

[14] S. Haykin. *Neural Networks. A Comprehensive Foundation*. Macmillan, New York, 1994.

[15] M. Kästner, B. Hammer, M. Biehl, and T. Villmann. Functional relevance learning in generalized learning vector quantization. *Neurocomputing*, 90(9):85–95, 2012.

[16] T. Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995. (Second Extended Edition 1997).

[17] T. M. Martinetz, S. G. Berkovich, and K. J. Schulten. 'Neural-gas' network for vector quantization and its application to time-series prediction. *IEEE Trans. on Neural Networks*, 4(4):558–569, 1993.

[18] J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society, London, A*, 209:415–446, 1909.

[19] C. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *Journal of Machine Learning Research*, 7:26051–2667, 2006.

[20] E. Mwebaze, P. Schneider, F.-M. Schleif, J. Aduwo, J. Quinn, S. Haase, T. Villmann, and M. Biehl. Divergence based classification in learning vector quantization. *Neurocomputing*, 74(9):1429–1435, 2011.

[21] E. Pekalska and R. Duin. *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications*. World Scientific, 2006.

[22] J. Platt. *Fast Training of Support Vector Machines Using Sequential Minimal Optimization*. MIT Press, Cambridge, MA, USA, 1999.

[23] A. Qin and P. Suganthan. A novel kernel prototype-based learning algorithm. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)*, volume 4, pages 621–624, 2004.

[24] A. K. Qin and P. N. Suganthan. Kernel neural gas algorithms with application to cluster analysis. In *ICPR (4)*, pages 617–620, 2004.

[25] 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.

[26] F.-M. Schleif, T. Villmann, B. Hammer, and P. Schneider. Efficient kernelized prototype based classification. *International Journal of Neural Systems*, 21(6):443–457, 2011.

[27] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.

[28] P. Schneider, K. Bunte, H. Stiekema, B. Hammer, T. Villmann, and M. Biehl. Regularization in matrix relevance learning. *IEEE Transactions on Neural Networks*, 21(5):831–840, 2010.

[29] P. Schneider, B. Hammer, and M. Biehl. Adaptive relevance matrices in learning vector quantization. *Neural Computation*, 21:3532–3561, 2009.

[30] U. Seiffert, F. Bollenbeck, H.-P. Mock, and A. Matros. Clustering of crop phenotypes by means of hyperspectral signatures using artificial neural networks. In *Proceedings of the 2nd IEEE Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing WHISPERS 2010*, pages 31–34. IEEE Press, 2010.

[31] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

[32] T. Villmann and S. Haase. Divergence based vector quantization. *Neural Computation*, 23(5):1343–1392, 2011.

[33] T. Villmann and S. Haase. A note on gradient based learning in vector quantization using differentiable kernels for Hilbert and Banach spaces. *Machine Learning Reports*, 6(MLR-02-2012):1–29, 2012. ISSN:1865-3960, http://www.techfak.uni-bielefeld.de/ ~ fschleif/mlr/mlr_02_2012.pdf.

[34] T. Villmann, S. Haase, and M. Kästner. Gradient based learning in vector quantization using differentiable kernels. In P. Estevecz, editor, *Proc. of the 9th Workshop on Self-Organizing Maps, Santiago de Chile*, LNCS, page submitted, Berlin, 2012. Springer.