Soft Comput (2015) 19:2423–2434
DOI 10.1007/s00500-014-1496-1

FOCUS

Border-sensitive learning in generalized learning vector quantization: an alternative to support vector machines

Marika Kaden · Martin Riedel · Wieland Hermann · Thomas Villmann

Published online: 20 November 2014

Abstract Learning vector quantization (LVQ) algorithms as powerful classifier models for class discrimination of vectorial data belong to the family of prototype-based classifiers with a learning scheme based on Hebbian learning as a widely accepted neuronal learning paradigm. Those classifier approaches estimate the class distribution and generate from this a class decision for vectors to be classified. The estimation can be done by the determination of class-typical sensitive prototypes inside the class distribution area like in LVQ or by detection of the class borders for class discrimination as preferred by support vector machines (SVMs). Both strategies provide advantages and disadvantages depending on the given classification task. Whereas LVQs are very intuitive and usually process the data during learning in the data space, frequently equipped with variants of the Euclidean metric, SVMs implicitly map the data into a high-dimensional kernel-induced feature space for better separation. In this Hilbert space, the inner product is compliant to the kernel. However, this implicit mapping makes a vivid interpretation more difficult. As an alternative, we propose in this paper two modifications of LVQ to make it comparable to SVM: first border-sensitive learning is introduced to achieve border-responsible prototypes comparable with support vectors in SVM. Second, kernel distances for differentiable kernels are considered, such that prototype learning takes place in a metric space isomorphic to the feature mapping space of SVM. Combination of both features gives a powerful prototype-based classifier while keeping the easy interpretation and the intuitive Hebbian learning scheme of LVQ.

Keywords Learning vector quantization · Border sensitive classification · Kernel distance · Support vector machines

Communicated by I. R. Ruiz.

M. Kaden · M. Riedel · T. Villmann (✉)
Computational Intelligence Group, University of Applied Sciences Mittweida, Mittweida, Germany
e-mail: thomas.villmann@hs-mittweida.de

W. Hermann
Department of Neurology, Paracelsus Hospital Zwickau, Zwickau, Germany

1 Introduction

Classification of vectorial data is still a challenging topic. For example, classification of hyper-spectral vectors in remote sensing image analysis requires precise learning of classifier models for frequently overlapping or non-linear class distributions (Villmann et al. 2003). Discrimination of patient records in medicine may demand subtle differentiation of features for correct disease diagnosis (Schleif et al. 2009; Villmann 2002; Wutzler et al. 2009). Adaptive models from machine learning such as learning vector quantizers (LVQ, Kohonen 1997), support vector machines (SVMs, Schölkopf and Smola 2002) or multilayer perceptrons (MLP, Haykin 1994) promise alternatives to traditional multivariate data analysis approaches like linear or quadratic discriminant analysis (LDA/QDA) (Duda and Hart 1973; Sachs 1992), when these classical statistical methods do not deliver results with sufficient precision.

LVQs as well as SVMs belong to prototype-based classifiers. LVQ algorithms generate under certain conditions class typical prototypes whereas for SVMs the resulting prototypes determine the class borders and are here called support vectors. These support vectors are always data points. They are identified by convex optimization providing a unique solution. Yet, LVQs as introduced by Kohonen (1986, 1990) realize an intuitive learning based on the Hebbian princi-

ple. A cost function-based variant was proposed by Sato and Yamada and denoted as generalized LVQ (GLVQ, Sato and Yamada 1996). Here, optimization is realized as stochastic gradient descent learning such that an optimum is achieved only with high probability instead of a unique solution obtained from SVM optimization. Further, LVQs handle the prototypes in the data space, usually equipped with the Euclidean metric, such that they are easy to interpret. In contrast, SVMs implicitly map the data into the feature mapping space (FMS) associated with the used kernel. This FMS is high dimensional, maybe infinite, and the mapping is generally non-linear. These SVM properties frequently lead to a superior performance compared to other classification models (Schölkopf and Smola 2002). Yet, there are other powerful classifiers in play including decision trees, random forest and deep architectures (Bengio 2009). A good comparison can be found in Caruana et al. (2006).

However, we restrict ourself to the aspect of precise learning of the class borders as one important aspect for good classification performance. In this sense, SVM still plays an important role. Yet, the number of support vectors in SVM, which can be taken as a measure for model complexity, may become large and cannot be explicitly controlled or determined in advance. The model complexity is only implicitly controlled by regularization and additional slack variables (Schölkopf and Smola 2002).

For LVQ approaches, in general the number of prototypes has to be fixed before model adaptation, i.e. before training. This might be an advantage, if restrictions to the complexity of the model appear as it might be the case in onboard technical systems with restricted memory like in robotics (Klingner et al. 2014). Yet, the basic LVQ scheme with fixed prototype number can be easily combined with growing networks like growing neural gas (GNG, Fritzke 1995) allowing a controlled increase of model complexity (Hammer et al. 2005a).

As mentioned above, support vectors detect the class borders such that SVMs maximize the class separation margin (Hastie et al. 2001) whereas GLVQ optimizes the hypothesis margin (Crammer et al. 2003). The border-sensitive behavior as well as the kernel feature mapping of SVMs contributes to their superior performance for many applications. Several attempts were made to integrate the kernel idea into GLVQ using approximation techniques in the related FMS (Qin and Suganthan 2004; Schleif et al. 2011).

In this paper, we deal with the other aspect—the border-sensitive learning for LVQ models. In particular, we propose two different methods to establish class border sensitivity in GLVQ. Thereby, the aim of the investigation is not to show better performance for the new LVQ variants compared to SVM. Rather than this goal we would like to show these variants as a matter of principle to border sensitive SVM, if explicit control of model complexity is demanded. This focus is triggered by the assumption that frequently class separation is favored versus class description. In general, both aspects are difficult to combine (Hammer et al. 2014).

The first one of those border-sensitive LVQ variants uses an additional penalty term for the cost function of GLVQ forcing explicitly the prototypes to move closer to the class borders such that a better sensibility is achieved. The second approach achieves the border sensitivity implicitly by a parameter control for the classifier function already implemented inside the standard GLVQ model. This latter strategy leads to an adaptation of prototypes only for those data points, which are close to the class borders and can be related to active learning schemes (Hasenjäger and Ritter 1998; Schleif et al. 2007). Hence, the prototypes learn only those data near the class borders and, therefore, are implicitly sensitized for the class decision boundaries. Yet, as pointed out in Hammer et al. (2014), border sensitivity does not automatically implies class-typical prototypes.

Both approaches are demonstrated for artificial, illustrating data sets as well as real world data.

2 Generalized learning vector quantization (GLVQ)

In this section, we briefly give the basic variants of LVQ according to Kohonen and Sato and Yamada to justify notations and descriptions. In particular, we assume a given training data set of vectors $\mathbf{v} \in V \subseteq \mathbb{R}^n$. The cardinality of V is denoted as $\#V$. The prototypes $\mathbf{w}_k \in \mathbb{R}^n$ of the LVQ model for data representation are collected in the set $W = \{\mathbf{w}_k \in \mathbb{R}^n,\ k = 1, \ldots, M\}$. Each training vector $\mathbf{v}$ belongs to a predefined class $x(\mathbf{v}) \in \mathcal{C} = \{1, \ldots, C\}$ out of C classes. The prototypes are labeled by $y(\mathbf{w}_k) \in \mathcal{C}$ such that at least one prototype is responsible for each class. Further, we suppose a dissimilarity measure $d(\mathbf{v}, \mathbf{w}_k)$ in the data space, frequently but not necessarily chosen as the squared Euclidean distance.

2.1 Kohonen's LVQ

LVQs as introduced by Kohonen (1992) are prototype-based classifiers with a predefined set of prototypes, which is optimized during learning and then serving as reference set for classification. The optimization takes place by *attraction and repulsion* of the prototypes for presented training samples in compliance with a nearest prototype principle (NPP). According to this principle, let $\mathbf{w}^+$ denote the nearest prototype for a given data sample (vector) $\mathbf{v}$ with respect to the dissimilarity measure d and, additionally, $y\left(\mathbf{w}^+\right) = x(\mathbf{v})$ is valid. Thus, $\mathbf{w}^+$ is the best matching prototype with correct class label also shortly denoted as best matching correct prototype. We define $d^+(\mathbf{v}) = d\left(\mathbf{v}, \mathbf{w}^+\right)$ as the respective dissimilarity degree. Analogously, $\mathbf{w}^-$ is the best matching prototype with a class label $y\left(\mathbf{w}^-\right)$ different from $x(\mathbf{v})$, i.e.

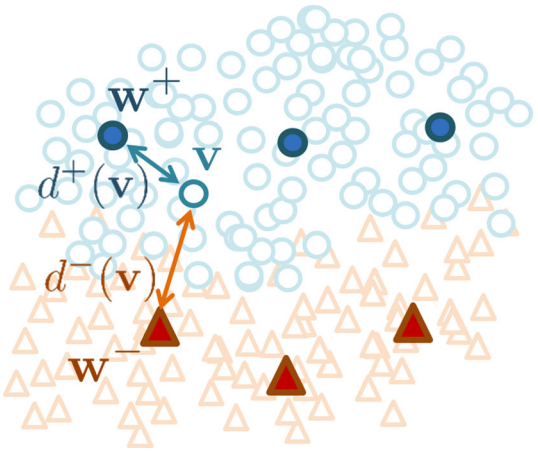

Fig. 1 Illustration of the winner determination of $\mathbf{w}^+$, the best matching correct prototype and the best matching incorrect prototype $\mathbf{w}^-$ together with their distances $d^+(\mathbf{v})$ and $d^-(\mathbf{v})$, respectively. The overall best matching prototype here is $\mathbf{w}^* = \mathbf{w}^+$

best matching incorrect prototype, and $d^-(\mathbf{v}) = d(\mathbf{v}, \mathbf{w}^-)$ is again the assigned dissimilarity degree, see Fig. 1.

Further, let

$$\mathbf{w}^* = \text{argmin}_{\mathbf{w}_k \in W}\left(d(\mathbf{v}, \mathbf{w}_k)\right) \tag{1}$$

be the overall best matching prototype (winner) without any label restriction and $d^*(\mathbf{v}) = d(\mathbf{v}, \mathbf{w}^*)$ the respective dissimilarity degree and $y^* = y(\mathbf{w}^*)$ indicates the respective class label of the winner. Hence, $\mathbf{w}^* \in W^* = \{\mathbf{w}^+, \mathbf{w}^-\}$ and W^* is denoted as the *winner subset* of the prototype set W. With these notations, the basic learning scheme in LVQ can be formulated as:

$$\triangle \mathbf{w}^* = \varepsilon \cdot \Psi\left(x(\mathbf{v}), y^*\right) \cdot \left(\mathbf{v} - \mathbf{w}^*\right) \tag{2}$$

with $0 < \varepsilon \ll 1$ being the learning rate (Biehl et al. 2014). The label evaluation function

$$\Psi\left(x(\mathbf{v}), y^*\right) = \delta_{x(\mathbf{v}),y^*} - \left(1 - \delta_{x(\mathbf{v}),y^*}\right) \tag{3}$$

determines the direction of the vector shift $\mathbf{v} - \mathbf{w}^*$ where $\delta_{x(\mathbf{v}),y^*}$ is the Kronecker symbol, such that $\delta_{x(\mathbf{v}),y^*} = 1$ holds for $x(\mathbf{v}) = y^*$ and zero elsewhere. This heuristic adaptation scheme leads to an approximation of a Bayes decision scheme (Kohonen 1997).

An improved convergence behavior is obtained for slight modifications of this basic scheme regarding, for example, an adaptive learning rate or an update also for the second winning prototype

$$\mathbf{w}^*_{2\text{nd}} = \text{argmin}_{\mathbf{w}_k \in W \setminus \{\mathbf{w}^*\}}\left(d(\mathbf{v}, \mathbf{w}_k)\right)$$

additionally to the overall winner $\mathbf{w}^*$. If $\mathbf{w}^*$ and $\mathbf{w}^*_{2\text{nd}}$ constitute the winner subset W^* Kohonen suggested a window rule

$$\min\left(\frac{d^+(\mathbf{v})}{d^-(\mathbf{v})}, \frac{d^-(\mathbf{v})}{d^+(\mathbf{v})}\right) \geq \frac{1-\omega}{1+\omega} \tag{4}$$

in the variant LVQ2.1. Prototype adaptation only takes place if this relation is fulfilled for a predefined value $0 < \omega < 1$ (Kohonen 1997), i.e. if the data sample $\mathbf{v}$ falls into a window around the decision border. Yet, this rule does not work for very high-dimensional data as explained in Witoelar et al. (2010).

After training, the response y^* of the LVQ yields the predicted classification of a data sample. According to the winner determination (1) for each data sample, the prototype set W determines a partition of the data space into the so-called receptive fields defined as:

$$R(\mathbf{w}_k) = \left\{\mathbf{v} \in V | \mathbf{w}_k = \mathbf{w}^*\right\} \tag{5}$$

also known as Voronoi tesselation. The dual graph $\mathcal{G}$, also denoted as Delaunay- or neighborhood graph, with prototype indices taken as the graph vertices, determines the class distributions via the class labels $y(\mathbf{w}_k)$ and the adjacency $\mathbf{G}$ matrix of $\mathcal{G}$ with elements $g_{ij} = 1$ iff $R(\mathbf{w}_i) \cap R\left(\mathbf{w}_j\right) \neq \emptyset$ and zero elsewhere. For given prototypes and data sample, the graph can be estimated using $\mathbf{w}^*$ and $\mathbf{w}^*_{2\text{nd}}$ (Martinetz and Schulten 1994).

It turns out that the window rule (4) may destabilize the learning process and, therefore, it was suggested to apply this rule only for a few learning steps after usual LVQ1 training to improve the performance (Kohonen 1997). This unstable behavior can be prevented or at least reduced, if the window rule is only applied if the receptive fields $R(\mathbf{w}^+)$ and $R(\mathbf{w}^-)$ are neighbored, i.e. $R(\mathbf{w}^+) \cap R(\mathbf{w}^-) \neq \emptyset$ (Kaden et al. 2014).

2.2 The basic GLVQ model

The aim of the *Generalized* LVQ introduced by Sato and Yamada Sato and Yamada (1996) was to keep the basic principle of attraction and repulsion in prototype-based classification learning but vanquishing the problem of the adaptation heuristic. In particular, stochastic gradient descent learning related to a well-defined cost function was identified as a powerful alternative. For this purpose, a classifier function

$$\mu(\mathbf{v}) = \frac{d^+(\mathbf{v}) - d^-(\mathbf{v})}{d^+(\mathbf{v}) + d^-(\mathbf{v})} \tag{6}$$

is considered, where $\mu(\mathbf{v}) \in [-1, 1]$ is valid and correct classification of a training sample $\mathbf{v}$ corresponds to $\mu(\mathbf{v}) \leq 0$. Then, a cost function

$$E_{\mathrm{GLVQ}}(W) = \frac{1}{\#V} \sum_{\mathbf{v} \in V} f(\mu(\mathbf{v})) \tag{7}$$

is defined with a monotonically increasing *transfer* or *squashing function* f. The squashing function is commonly chosen as a sigmoid function like

$$f(x) = \frac{1}{1 + \exp(-x)} \tag{8}$$

or as the identity function $f(x) = id(x) = x$.

Learning in GLVQ takes place as stochastic gradient descent on $E_{\mathrm{GLVQ}}(W)$. In particular, we have

$$\triangle \mathbf{w}^+ \sim \xi^+(\mathbf{v}) \cdot \frac{\partial d^+(\mathbf{v})}{\partial \mathbf{w}^+} \quad \text{and} \quad \triangle \mathbf{w}^- \sim \xi^-(\mathbf{v}) \cdot \frac{\partial d^-(\mathbf{v})}{\partial \mathbf{w}^-} \tag{9}$$

with the scaling factors

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

and

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

For the squared Euclidean metric, we obtain a vector shift according to

$$\frac{\partial d^{\pm}(\mathbf{v})}{\partial \mathbf{w}^{\pm}} = -2\left(\mathbf{v} - \mathbf{w}^{\pm}\right) \tag{10}$$

for the prototypes.

As shown in Crammer et al. (2003), GLVQ maximizes the hypothesis margin

$$M(\mathbf{v}, x(\mathbf{v})) = d^+(\mathbf{v}) - d^-(\mathbf{v}) \tag{11}$$

which refers to the distance of the closest prototype labeled with a different class from $\mathbf{v}$. Thus, it describes a 'security' of classification, i.e. it is related to the distance that the prototypes can be altered without changing the classification decision (Nova and Estévez 2013; Kaden et al. 2014). The hypothesis margin is associated with the generalization error bound independent of the data dimension but depending on the number of prototypes (Hammer et al. 2005b). Further, we remark that minimizing the cost function $E_{\mathrm{GLVQ}}(W)$ from (7) approximates the minimization of the misclassification rate (Kaden et al. 2014).

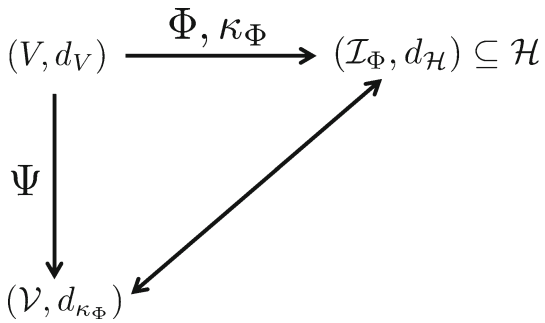

Fig. 2 Visualization of the relationship between the original data space V, the kernel mapping Φ and the kernel data space $\mathcal{V}$. The metric space $\mathcal{V}$ is isometric to the image $\mathcal{I}_{\kappa_\Phi}$ of V under the kernel map Φ under certain conditions, which itself uniquely corresponds to the kernel κ_Φ in a canonical manner

2.3 GLVQ and non-Euclidean distances

Depending on the classification task, other (differentiable) dissimilarity measures than the Euclidean may be more appropriate (Hammer and Villmann 2002; Villmann and Haase 2011). Quadratic forms $d_\Lambda(\mathbf{v}, \mathbf{w}) = (\mathbf{v}, \mathbf{w})^\top \Lambda (\mathbf{v}, \mathbf{w})$ are discussed in Bunte et al. (2012), and Schneider et al. (2009a, b, 2010). Here, the positive semi-definite matrix Λ is decomposed into $\Lambda = \Omega^\top \Omega$ with arbitrary matrices $\Omega \in \mathbb{R}^{m \times D}$ which can be adapted during the training. For classification visualization, the parameter m has to be two or three, the full problem is obtained for $m = D$. If data are matrices, distances based on Schatten norms or general matrix norms come into play (Horn and Johnson 2013; Schatten 1950), which show very good discriminative behavior (Gu et al. 2012). Alternatively, SVMs implicitly map the data and prototypes into a high-, maybe infinite-,dimensional function Hilbert space $\mathcal{H}$ by a generally non-linear mapping $\Phi : V \to \mathcal{I}_{\kappa_\Phi} \subseteq \mathcal{H}$ to achieve better class separability (Cristianini and Shawe-Taylor 2000; Shawe-Taylor and Cristianini 2004). This mapping corresponds uniquely in canonical manner to positive-definite universal kernels[1] $\kappa_\Phi(\mathbf{v}, \mathbf{w})$ (PU-kernels), such that $\langle \Phi(\mathbf{v}), \Phi(\mathbf{w}) \rangle_\mathcal{H} = \kappa_\Phi(\mathbf{v}, \mathbf{w})$ is valid for the inner product $\langle \bullet, \bullet \rangle_\mathcal{H}$ (Aronszajn 1950; Mercer 1909; Steinwart 2001) (Fig. 2). By means of this inner product, the metric $d_\mathcal{H}$ is determined and for the image $\mathcal{I}_{\kappa_\Phi}$ of the mapping Φ the equality $d_\mathcal{H}(\Phi(\mathbf{v}), \Phi(\mathbf{w})) = d_{\kappa_\Phi}(\mathbf{v}, \mathbf{w})$ holds with

$$d_{\kappa_\Phi}(\mathbf{v}, \mathbf{w}) = \sqrt{\kappa_\Phi(\mathbf{v}, \mathbf{v}) - 2\kappa_\Phi(\mathbf{v}, \mathbf{w}) + \kappa_\Phi(\mathbf{w}, \mathbf{w})} \tag{12}$$

is the so-called kernel distance (Villmann et al. 2013). In context of GLVQ, it is interesting to consider *differentiable*

[1] The theory of universal kernels is out of the focus of this paper, we explicitly refer to Steinwart (2001) and Micchelli et al. (2006) for precise definition and consideration of there properties. Here, we only remark that exponential kernels belong to the set of universal kernels.

PU-kernels (DPU-kernels), for which the derivative $\frac{\partial \kappa_\Phi(\mathbf{v},\mathbf{w})}{\partial \mathbf{w}}$ exists. We can define an accompanying formal data transformation $\Psi : V \longrightarrow \mathcal{V}$, where $\mathcal{V}$ is the data space equipped with the kernel metric d_{κ_Φ}. The formal map Ψ is bijective, continuous and non-linear iff Φ does (Steinwart 2001). Further, $\mathcal{V}$ is isometric and isomorphic to $\mathcal{I}_{\kappa_\Phi}$ and, hence, offers the same topological structure and richness as the image $\mathcal{I}_{\kappa_\Phi}$ as known from SVMs.

The differentiability of the kernel ensures the applicability of the stochastic gradient learning of GLVQ in $\mathcal{V}$, replacing the distance $d(\mathbf{v}, \mathbf{w})$ from the data space V contained in the calculation of the classifier function $\mu(\mathbf{v})$ from (6) by the kernel distance $d_{\kappa_\Phi}(\mathbf{v}, \mathbf{w})$ (Villmann et al. 2014). We denote this new data space $\mathcal{V}$ as *kernelized data space* and the respective GLVQ in $\mathcal{V}$ as kernel-GLVQ (KGLVQ). We remark that this approach does not require any approximation techniques as suggested in earlier kernelized GLVQ variants proposed in Qin and Suganthan (2004) and Schleif et al. (2011).

3 Class border sensitive learning in GLVQ

As we have seen in the previous section, KGLVQ is an extension of usual GLVQ to kernel data spaces $\mathcal{V}$. However, in general, the prototypes of KGLVQ as well as for GLVQ are not particularly sensitized to detect the class borders. This might be a disadvantage for KGLVQ compared to SVMs, if precise classification decisions are favored. In this section, we provide two possibilities to integrate class border sensitivity in GLVQ and, hence, also for KGLVQ. The first choice applies an additive attraction force for prototypes with different class responsibilities, such that the prototypes move closer to each other, which *implicitly* leads to an improved class border sensitivity. The second approach supposes a parametrized sigmoid transfer functions $f_\theta(\mu)$ in (7), where the θ parameter controls the class border sensitivity via the so-called active sets. These active sets appear as subsets of the whole training data set containing only those data samples close to the class borders. It turns out that only the data contained in the active set contribute to the prototype learning and, hence, the prototypes become particularly sensitive to these data subsets.

3.1 Border sensitive learning in GLVQ by an additive penalty function

Penalizing an undesirable behavior of a learning system is a common strategy in machine learning. In this context, class border sensitivity learning by an additive penalty term for *unsupervised* fuzzy-c-means models was proposed in Villmann et al. (2012) and Yin et al. (2012). Here, we adopt these ideas for class border sensitive learning in GLVQ, i.e. we also consider a penalty strategy for classification learning (P-GLVQ). For this purpose, we extend the cost function of GLVQ (7) by an additive penalty term $F_{\text{neigh}}(W, V)$ such that

$$E_{\text{P-GLVQ}}(W, \gamma) = (1-\gamma)\cdot E_{\text{GLVQ}}(W) + \gamma \cdot F_{\text{neigh}}(W, V) \tag{13}$$

is a convex sum. The parameter $\gamma \in [0, 1)$ is the sensitivity control parameter. The penalty term $F_{\text{neigh}}(W, V)$ is a *neighborhood-attentive attraction force* (NAAF)

$$F_{\text{neigh}}(W, V) = \sum_{\mathbf{v}\in V} \sum_{k:\mathbf{w}_k \in W^-(\mathbf{v})} h_\sigma^{NG}\left(k, \mathbf{w}^+, W^-(\mathbf{v})\right) d\left(\mathbf{w}^+, \mathbf{w}_k\right) \tag{14}$$

depending on the subset $W^-(\mathbf{v}) \subset W$ of all prototypes with incorrect class labels for a given data vector $\mathbf{v}$. The term $d\left(\mathbf{w}^+, \mathbf{w}_k\right)$ explicitly penalizes large distances between the best matching prototypes of the correct and incorrect class, i.e. large distances of these prototypes to the class borders.

This distance is weighted with the neighborhood function

$$h_{\sigma_-}^{NG}\left(k, \mathbf{w}^+, W^-(\mathbf{v})\right) = c_{\sigma_-}^{NG} \cdot \exp\left(-\frac{\left(rk_k\left(\mathbf{w}^+, W^-(\mathbf{v})\right) - 1\right)^2}{2\sigma_-^2}\right) \tag{15}$$

determining a rank-neighborhood between the prototypes in $W^-(\mathbf{v})$ via the dissimilarity rank function

$$rk_k\left(\mathbf{w}^+, W^-(\mathbf{v})\right) = \sum_{\mathbf{w}_l \in W^-(\mathbf{v})} H\left(d\left(\mathbf{w}^+, \mathbf{w}_k\right) - d\left(\mathbf{w}^+, \mathbf{w}_l\right)\right) \tag{16}$$

known from *Neural Gas* (NG, Martinetz et al. 1993). Here, H is the Heaviside function

$$H(x) = \begin{cases} 0 & \text{if } x \le 0 \\ 1 & \text{else.} \end{cases} \tag{17}$$

The NAAF causes an additional gradient term

$$\frac{\partial F_{\text{neigh}}(W, V)}{\partial \mathbf{w}_j} = h_{\sigma_-}^{NG}\left(j, \mathbf{w}^+, W^-(\mathbf{v})\right) \cdot \frac{\partial d\left(\mathbf{w}^+, \mathbf{w}_j\right)}{\partial \mathbf{w}_j} \tag{18}$$

for a given input vector $\mathbf{v}$ and $\mathbf{w}_j \in W^-(\mathbf{v})$, i.e. all incorrect prototypes are gradually moved towards the correct best matching prototype $\mathbf{w}^+$ according to their dissimilarity rank with respect to $\mathbf{w}^+$ but into the direction defined by the derivative $\frac{\partial d(\mathbf{w}^+, \mathbf{w}_j)}{\partial \mathbf{w}_j}$. For the squared Euclidean distance, this gives a gradual usual vector shift. In consequence, the closer the neighborhood is to the correct winning prototype

$\mathbf{w}^+$, the stronger is the resulting neighborhood attraction but controlled by the neighborhood range $\sigma_- > 0$. Further, the weighting coefficient γ regulates the overall influence of border sensitive learning in this model. Because, all prototypes in $W^-(\mathbf{v})$ belong to another class than $\mathbf{w}^+$, the introduced attraction force enhances prototypes positions close to the decision borders between classes. Hence, an implicit better class border sensitivity is achieved.

Otherwise, two new parameters have to be involved for the realization of such a strategy, which requires additional control in applications. Moreover, the margin optimization strategy is lost due to the penalty term $F_{\text{neigh}}(W, V)$ from (14) contributing to the over all costs (13). From this point of view, a more GLVQ-inherent approach is desired.

3.2 Class border sensitive learning by parametrized transfer functions f_θ in GLVQ

In this section, we provide an alternative approach for class border sensitive learning in GLVQ, which is more consistent to standard GLVQ than the previously penalty strategy. It uses a more inherent modification of standard GLVQ than P-GLVQ and, therefore, does not need an additional external force. For this purpose, we seize the thought presented in Strickert (2011) and Witoelar et al. (2010) that the shape of the transfer function f in (7) sensitively influences the decision certainty at the class borders. Therefore, we suppose f to be of the sigmoid type (8) as the usual opposite to the identity already suggested in the very first presentation of the GLVQ approach (Sato and Tsukumo 1994; Sato and Yamada 1995). Particularly, we can pay attention to the sensitivity behavior introducing the respective parametrized variant

$$f_\theta(x) = \frac{1}{1 + \exp\left(-\frac{x}{2\theta^2}\right)} \tag{19}$$

with the parameter θ determining the slope of f_θ, see Fig. 3.

The derivative $f'_\theta(\mu(\mathbf{v}))$ of the logistic function can be expressed in terms of the sigmoid function itself and reads as:

$$f'_\theta(\mu(\mathbf{v})) = \frac{f_\theta(\mu(\mathbf{v}))}{2\theta^2} \cdot (1 - f_\theta(\mu(\mathbf{v}))), \tag{20}$$

which appears in the scaling factors $\xi^\pm$ occuring in the updates (9) for the winning prototypes $\mathbf{w}^\pm$. Considering this derivative (see Fig. 4), we observe that $|\xi^\pm| \gg 0$ holds only for $|\mu(\mathbf{v})| \ll 1$, i.e. a significant prototype update only takes place for a small range of the classifier values μ in (6). This range also depends on the slope parameter θ. Therefore, we introduce the *active set* of the data contributing significantly to a prototype update during learning to be the set

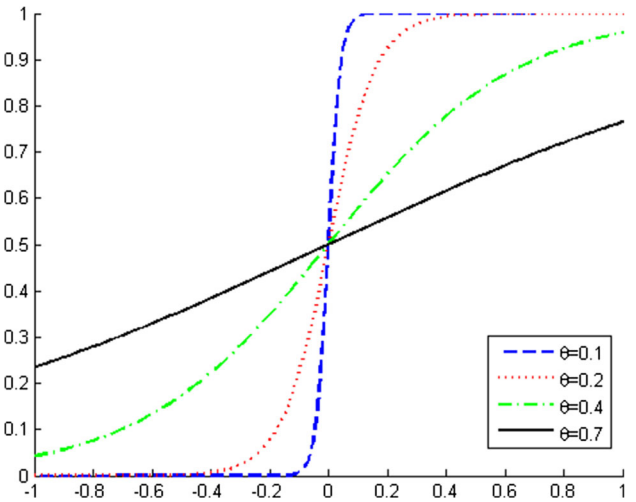

Fig. 3 Visualization of the parametrized sigmoid function $f_\theta(x)$ depending on the slope parameter θ

$$\hat{\Xi} = \left\{ \mathbf{v} \in V \,|\, \mu(\mathbf{v}) \in \left[-\frac{1-\mu_\theta}{1+\mu_\theta}, \frac{1-\mu_\theta}{1+\mu_\theta} \right] \right\} \tag{21}$$

with μ_θ chosen such that $f'_\theta(\mu) \approx 0$ is valid for $\mu \in \Xi = V \setminus \hat{\Xi}$. Thus, $|\xi^\pm| \approx 0$ is valid for all training sample not belonging to the active set. Hence, these data do not contribute to prototype learning; see Fig. 4.

Otherwise, data samples contained in the active set yield moderate prototype updates such that the prototypes become sensitized for them. Obviously, the active set $\hat{\Xi}$ is distributed along the class decision boundaries, because $f'_\theta(\mu) \gg 0$ is valid only here. Therefore, the active set $\hat{\Xi}$ is characterized by values $\mu(\mathbf{v}) \approx 0$ for $\mathbf{v} \in \hat{\Xi}$. Hence, this active set $\hat{\Xi}$ can be understood as another formulation of the window rule for LVQ2.1 given in (4) and taking there $\omega = \mu_\theta$ (Kohonen 1997). The learning of the parameter θ in GLVQ was explicitly addressed in Witoelar et al. (2010). Optimization for accuracy improvement was discussed in Strickert (2011).

These observations lead to the idea to control the border sensitivity of the GLVQ algorithm by the parameter θ, which obviously determines the width of the active set surrounding the class borders. Large values correspond to an overall learning whereas small θ values define small stripes as active sets. In consequence, only these data contribute to the prototype updates. In other words, according to (21), the active set is crisp but the possibilities for control are smooth such that we could speak about *thresholded active sets* $\hat{\Xi}_\theta$. Therefore, border sensitivity leads to prototypes sensitized to those data points close to the class borders depending on the control parameter θ. In this sense, the active set learning can be seen as a kind of *attention based or learning* (Hermann et al. 1994) or *active learning* (Hasenjäger and Ritter 1998; Hasenjäger et al. 1999; Schleif et al. 2007). We refer to this border sensitive approach as BS-GLVQ.

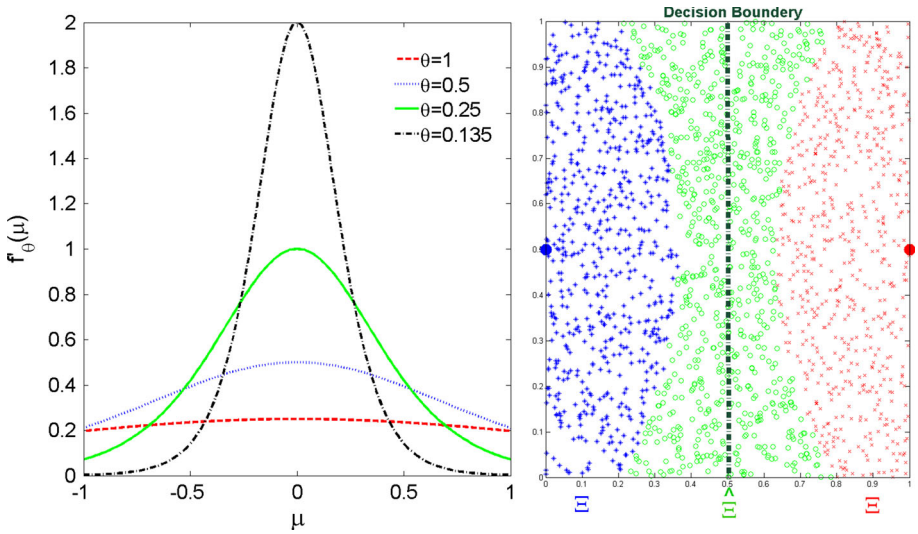

Fig. 4 *Left* derivatives $f'_\theta(\mu)$ for several θ values. *Right* visualization of the active set $\hat{\Xi}$ (*green circles*) for a illustrative two-class example with one prototype per class. The prototypes are the *big dots*. Only data points belonging to the active set contribute significantly to prototype learning such that they become sensitive for them (color figure online)

It should be explicitly mentioned at this point that BS-GLVQ is still optimizing the hypothesis margin $M(\mathbf{v}, x(\mathbf{v}))$ from (11) because the transfer function is still monotonically increasing.

Last but not least, we emphasize another advantage of this border sensitive approach: For this methodology, we can state that in the limit $\theta \searrow 0$ the respective cost function

$$E_{\mathrm{BS-GLVQ}}(W, \theta) = \frac{1}{\#V} \sum_{\mathbf{v} \in V} f_\theta(\mu(\mathbf{v})) \tag{22}$$

reflects the classification error (Kaden et al. 2014). This fact is based on the observation that in the limit $\theta \to 0$ the sigmoid f_θ from (19) becomes the Heaviside function $H(x)$ (17) (Kaden et al. 2014).

4 Illustrative example and application

In this section, we demonstrate the desired properties of the border sensitive variants of GLVQ. For this purpose, we start with two-dimensional artificial data sets such that the results can be easily visualized. Thereafter, we present results from a medical application. After this, we move to more sophisticated real-world examples and applications, one from image segmentation, the other one being a medical application in neurology. For all experiments, using a GLVQ-variant, the prototypes were initialized randomly as data points of the respective classes.

4.1 Illustrative toy examples

The first two-dimensional artificial data set is a three-class problem. The data classes are uniformly distributed as in the Czech-flag, see Fig. 5.

For each class, we generated 1,000 data points for training and 1,000 for testing. All GLVQ variants were trained in 2,000 epochs with constant learning rate $\varepsilon = 0.01$. The results are reported for the test data.

We compare both border sensitive GLVQ approaches P-GLVQ and BS-GLVQ with a standard GLVQ network. We refer to standard GLVQ for the variant with the identity transfer function $f(\mu) = \mu$. In case of the parametrized transfer function f_θ, we used the initial parameter $\theta_{\mathrm{init}} = 1.0$ decreased to $\theta_{\mathrm{fin}} = 0.1$ during learning (BS-GLVQ). The balancing parameter γ for P-GLVQ was set permanently to $\gamma = 0.5$.

We observe that both border sensitive models place the prototypes closer to the class borders than standard GLVQ, see Fig. 5. Moreover, the classification accuracy is improved: For the BS-GLVQ, we achieved 91.1 % and the sigmoid variant results 97.2 % whereas standard GLVQ gets only 89.7 %. Thus, class border sensitive models detect the noisy class borders more accurately.

The second artificial two-dimensional data set, depicted in Fig. 6, is a non-linear two-class problem denoted as Palau flag.

Again we generated for each class 1,000 data for training and 1,000 for testing. As before, the learning rate was fixed to be $\varepsilon = 0.01$ during the learning within 2,000 epochs and the results are reported for the test data. If standard GLVQ is applied with two prototypes, i.e. one prototype for each class, and the Euclidean distance used as dissimilarity measure, a test accuracy of only 67.1 % is obtained. If we switch in this standard GLVQ to a kernel distance $d_{\kappa_\Phi}(\mathbf{v}, \mathbf{w})$ according to (12) with an exponential kernel

$$\kappa_\Phi(\mathbf{v}, \mathbf{w}) = \exp\left(-\frac{(\mathbf{v} - \mathbf{w})}{2\sigma^2}\right) \tag{23}$$

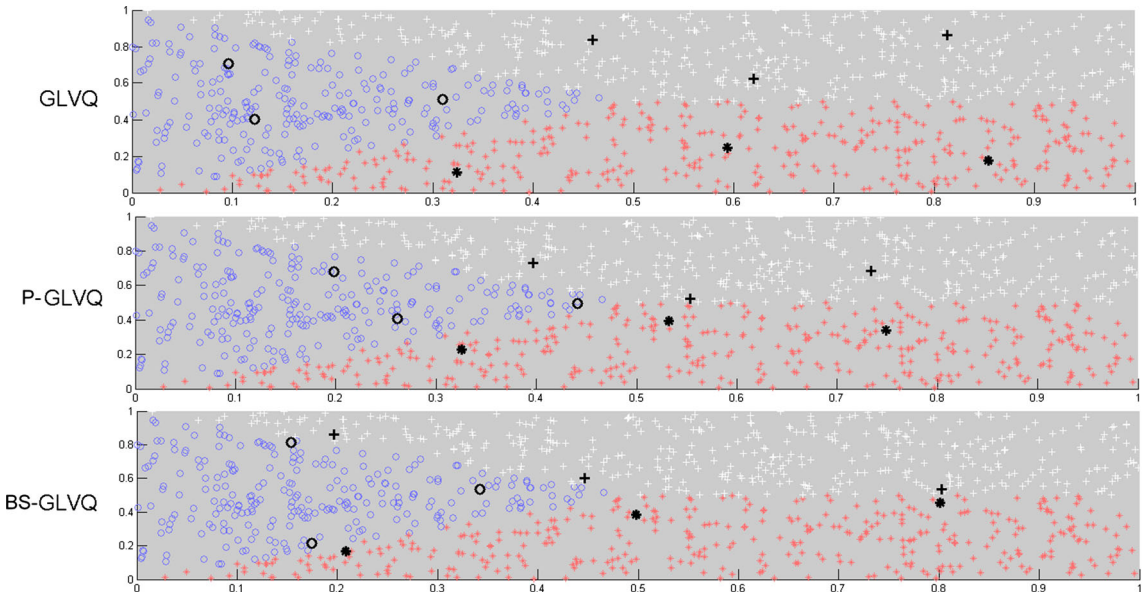

Fig. 5 Border sensitive learning for the class distribution example 'Czech-flag': Obtained prototype positions for standard GLVQ (*top*), P-GLVQ (*middle*) and BS-GLVQ (*bottom*). The Euclidean distance was used

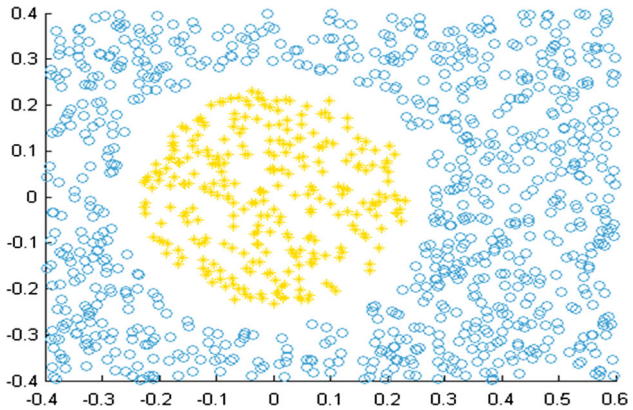

Fig. 6 Palau flag data set

the performance is increased to 95.5 %. The kernel width was automatically adapted according to the gradient learning $\frac{\partial E_{\text{BS-GLVQ}}(W,\theta)}{\partial \sigma}$ as described in Villmann et al. (2014). We refer to this kernelized GLVQ as KGLVQ.

In the next step, we used 2 prototypes for the inner class and 4 for the surrounding. Again we started with the Euclidean variant. Standard GLVQ achieved 97.1 % accuracy. Applying BS-GLVQ with slowly linearly decreasing θ parameter down to $\theta_{\text{fin}} = 0.16$, the performance is further increased to 99.1 %. Subsequently, we applied again the KGLVQ as before. For standard KGLVQ, a slightly improved accuracy of 97.4 % compared to standard GLVQ is obtained. The resulting prototype and incorrectly classified data points are visualized in Fig. 7.

However, the BS-GLVQ performance is not achieved. Incorporating now the border sensitive learning, i.e BS-KGLVQ is applied, BS-GLVQ is outperformed by a further improved accuracy of 99.5 % for a final θ parameter $\theta_{\text{fin}} = 0.11$. Yet, the prototypes move closer to the class borders to realize better sensitivity. In comparison to this BS-KGLVQ with the predefined number of only 6 prototypes, SVM approach (LIB-SVM, vers. 3.17, Chang and Lin 2011) yields an accuracy of 99.9 % but requires 16 support vectors serving as prototypes, see Fig. 8. Here, the kernel with was determined manually to be $\sigma = 0.7$ with regularizing parameter $C = 500$ for best performance.

4.2 Real-world datasets

4.2.1 Image segmentation data set

The image segmentation data set (ISDS) is from the UCI Repository (Blake and Merz 1998). It consist of 2, 310 data vectors of dimension $n = 16$ feature values of instances drawn from a database of 7 outdoor images. Each instance is a 3×3 region. The features describe structural and statistical properties of the instances like saturation, intensity, RGB values, hedge values and others. Each data vector is assigned to the original outdoor image as the respective class label. Thus, classification of these vector is a 7-class problem with 300 samples for each class.

The data were preprocessed applying a z-score transformation, i.e. normalized according to the variance. The mean is set to the zero vector, accordingly. The data set was processed by tenfold cross-validation. In each fold, the lower amount of 10 % of the data was taken as training samples, the

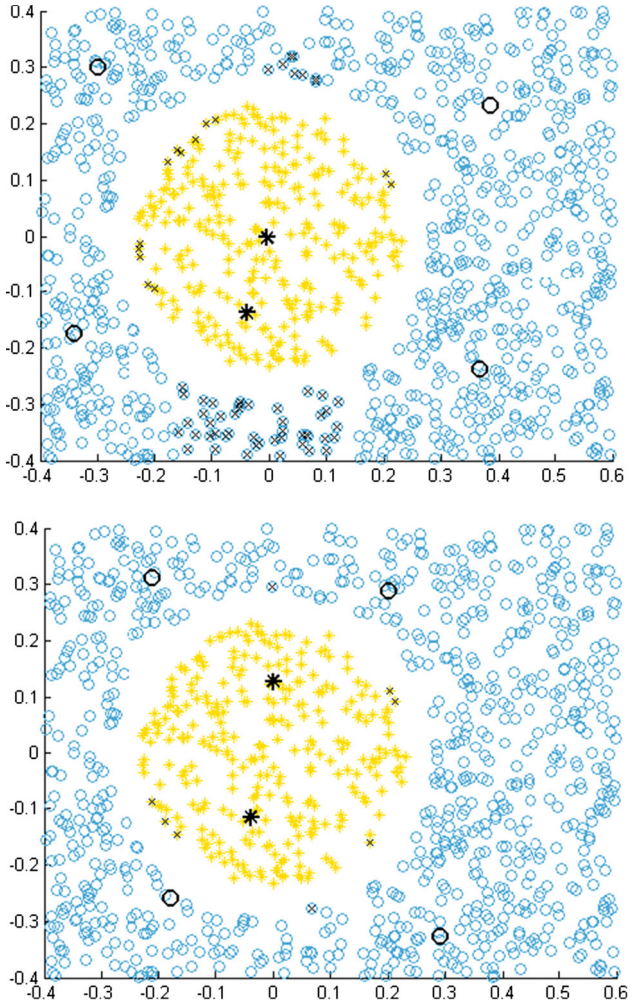

Fig. 7 Application of KGLVQ with $2+4$ prototypes to the Palau flag data set (*top*). Wrongly classified data points are marked by *crosses*. BS-KGLVQ achieves an improvement (*bottom*). Prototypes move closer to the class borders

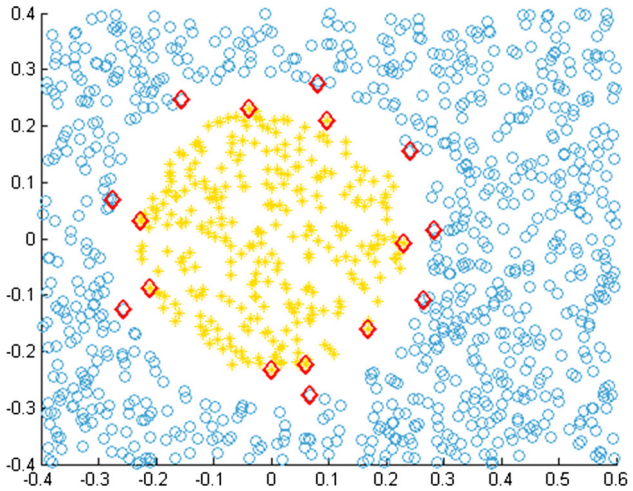

Fig. 8 Application of SVM to the Palau flag data set. Support vectors are marked by *diamonds*. BS-KGLVQ achieves an improvement (*bottom*). Prototypes move closer to the class borders compared to KGLVQ

remaining data were applied for testing. All reported results refer to the averaged test outcomes.

The main concern of these simulations is to verify experimentally the convergence of the cost function value to the classification error in case of decrease in the border-sensitivity control parameter θ.

We applied BS-GLVQ, with two prototypes per class. During the learning, the control parameter θ was fixed to $\theta_{\text{ini}} = 0.7$ during the first training phase. For this value, $f_\theta(\mu) \approx id(\mu) = \mu$ holds, i.e. BS-GLVQ behaves like standard GLVQ (quasi-linear activation). After this initial phase, the sensitivity parameter θ is slowly decreased to zero (Fig. 9).

We observe that the classification error rate, i.e. the ratio of the number of misclassified test data and the overall number of test data, decreases from 0.15 without border sensitive learning to 0.04 if the border sensitivity is considered. Further, for $\theta \searrow 0$ the cost function value $E_{\text{BS-KGLVQ}}(W, \theta)$ approaches the classification error as predicted by the theory.

Replacing the Euclidean distance by the kernel distance with self-adapting kernel width and keeping the remaining parameter settings unchanged lead to a slightly deteriorated classification error rate of 0.16 and 0.05, respectively.

4.2.2 *Classification of Wilson's disease based on electrophysiological data*

The last data set is regarded to detection of neurological manifestation of Wilson's disease. Wilson's disease is an autosomal-recessive disorder of copper metabolism, which shows disturbances in the liver function. This liver impairment leads to an accumulation of copper in the brain. In consequence, a reduced glucose metabolism is observed in several parts of the brain, the basal ganglia show hepatic and extrapyramidal motor symptoms and movement disorders are apparently for patients suffering from Wilson's disease (Barthel et al. 2001; Hermann et al. 2003, 2005). According to a clinical scheme suggested by Konovalov, patients can be divided into two main groups: patients with neurological and without neurological manifestations denoted as the neurological and the non-neurological group, respectively (Hermann et al. 2002). In addition to hepatolenticular degeneration in Wilson's disease, sensory and extrapyramidal motoric systems are also disturbed. The impairments of these nervous pathways can be detected by investigation of the latencies for evoked potentials in different brain regions (Günther et al. 2011). Collecting several evoked potentials according to a pre-defined medical examination yields a so-called electrophysiological impairment profile (EIP) for the patient. Yet, it is not clear so far, whether a precise classification of the EIPs according to their underlying neurological type is possible (Hermann et al. 2005).

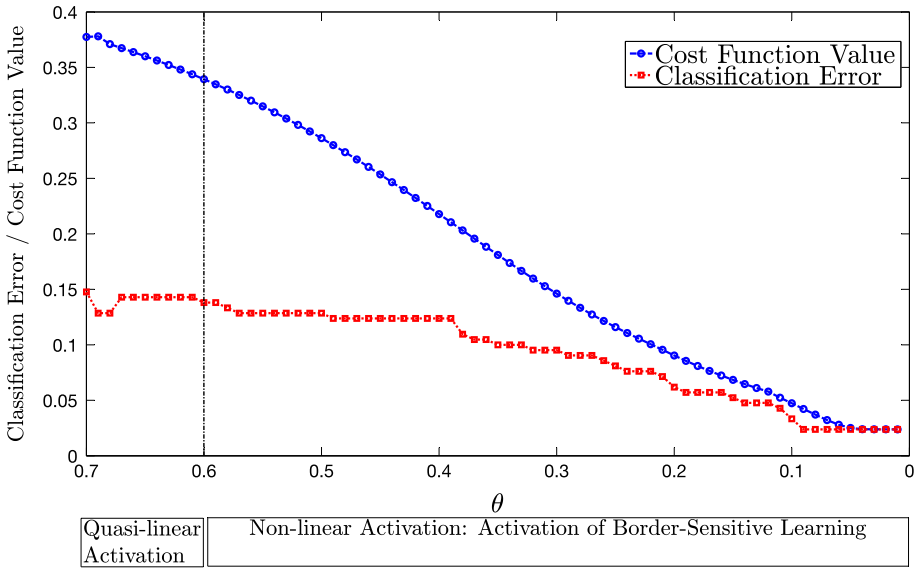

Fig. 9 BS-GLVQ with Euclidean distance for the ISDS. Starting with a quasi-linear activation corresponding to $\theta \approx 0.7$ (standard GLVQ), the sensitivity control parameter θ is slowly decreased in steps of 0.01. Each level is trained during 1,000 epochs with constant learning rate $\varepsilon = 0.01$. The process is accompanied by a decreasing classification error (*red square*). For $\theta \searrow 0$, the cost function value $E_{\text{BS-GLVQ}}(W, \theta)$ approaches the classification error. Curves were obtained from tenfold cross-validation

The aim of the study was to find a hint, whether such a distinction can be made. For this purpose, a database containing $M = 122$ five-dimensional EIPs was provided, generated according to the conditions defined in Hermann et al. (2003). Classic discriminant analysis did not reveal any significant distinction (Hermann et al. 2005).

We applied both border sensitive GLVQ algorithms. P-GLVQ was trained with balancing parameter $\gamma = 0.5$. BS-KGLVQ was applied with several θ parameter. For all models, we used 3,000 training epochs with constant learning rate $\varepsilon = 0.01$ and 6 prototypes per class. For comparison, an SVM with radial basis function kernel (rbf) was trained. The kernel width σ_{rbf} and the regularizing parameter C were manually tuned for best performance ($\sigma = 0.015, C = 200$). The data were preprocessed by a z-score transformation and classification results are obtained as tenfold cross validation. The respective results are depicted in Table 1.

BS-KGLVQ achieves drastically improved accuracies compared to standard KGLVQ, which accords approximately the BS-KGLVQ for $\theta = \frac{1}{\sqrt{2}}$, see Table 1. With increasing sensitivity (decreasing θ parameter), better accuracies are obtained. However, if the θ value drops down, the best performance of 89.2 % test accuracy is lost. This may be dedicated to the fact that the active set $\hat{\Xi}$ becomes too small for precise learning in those cases. For comparison, we also applied SVM to achieve a test accuracy of 87.4 %. Hence, without the border sensitivity feature, SVMs would be superior. Yet, incorporating border sensitive learning into GLVQ, KGLVQ variants outperform SVMs in this application. Further, we remark at this point that model complexity of the SVMs is at least three times larger (in average 45.5 support vectors for SVM) in comparison to the 12 prototypes used for the KGLVQ models.

Table 1 Accuracies and respective standard deviations for the Wilson's disease classification for the applied classifier models obtained by tenfold cross-validation

Dataset	BS-KGLVQ					SVM
	$\theta = \frac{1}{\sqrt{2}}$	$\theta = \frac{1}{\sqrt{5}}$	$\theta = \frac{1}{\sqrt{7}}$	$\theta = \frac{1}{\sqrt{10}}$	P-GLVQ	σ_{rbf}
Training	87.8 %	91.9 %	90.0 %	90.4 %	90.1 %	87.5 %
	(±0.013)	(±0.015)	(±0.015)	(±0.014)	(±0.011)	(±0.015)
Test	81.9 %	82.6 %	89.2 %	87.4 %	91.0 %	87.4 %
	(±0.086)	(±0.086)	(±0.083)	(±0.090)	(±0.090)	(±0.137)

Thereby, 'training' refers to the averaged *training* accuracy whereas 'test' is dedicated to the averaged test performance

Regarding the medical question, we have to state that although we achieved a quite high performance using BS-GLVQ, the obtained classification accuracies are not sufficiently high for a secure clinical discrimination. For this purpose, further investigations including an improved database and/or other dissimilarity measures are mandatory.

5 Conclusion and outlook

In this paper, we introduced two strategies for class border sensitive learning in GLVQ. The first one adds a penalty term to the cost function to force class border sensitivity of the prototypes, the second uses a parameter control of the sigmoid transfer function defining implicitly the so-called active sets as subsets of the whole training data, which only contribute significantly to prototype adaptation. The latter approach adopts an observation about certainty of class decision boundaries already investigated earlier but not utilized

for improved learning so far. This methodology realizes some kind of attention based or active learning as earlier proposed. The proposed strategies for border sensitive learning together with a kernelized variant of GLVQ offer a powerful alternative to SVMs. An advantage of the introduced approaches compared to SVM is the explicit control of the model complexity in GLVQ/KGLVQ, because the number of prototypes has to be chosen in advance for these models whereas in SVMs the number of support vector may become quite large in case of difficult classification tasks and cannot be explicitly controlled.

We applied and compared the new border sensitive GLVQ approaches for several data set: Artificial data sets were considered for illustration whereas real-world data set offers more challenging difficulties to be surmounted. In particular, a medical data set of neurophysiological data in case of Wilson's disease patients was investigated. Border sensitive KGLVQ variants achieve better results than SVMs with significant lower model complexity. Further, the classification results indicate that a discrimination between neurological and non-neurological type of Wilson's disease can be performed on the basis of electrophysiological impairment profiles. However, this hypothesis needs further (medical) investigations.

Acknowledgments M. Kaden and M. Riedel acknowledge funding by the European Social Fonds (ESF), Saxony, Germany.

References

Aronszajn N (1950) Theory of reproducing kernels. Trans Am Math Soc 68:337–404

Barthel H, Villmann T, Hermann W, Hesse S, Kühn HJ, Wagner A, Kluge R (2001) Different patterns of brain glucose consumption in Wilsons disease. Zeitschrift für Gastroenterologie 39:241

Bengio Y (2009) Learning deep architectures for AI. Found Trends Mach Learn 2(1):1–127

Biehl M, Hammer B, Villmann T (2014) Distance measures for prototype based classification. In: Petkov N (ed) Proceedings of the international workshop on brain-inspired computing 2013 (Cetraro/Italy). Springer, Berlin

Blake C, Merz C (1998) UCI repository of machine learning databases. University of California, Irvine, CA, Department of Information and Computer Science. http://www.ics.edu/mlearn/MLRepository.html

Bunte K, Schneider P, Hammer B, Schleif FM, Villmann T, Biehl M (2012) Limited rank matrix learning, discriminative dimension reduction and visualization. Neural Netw 26(1):159–173

Caruana R, Niculescu-Mizil A (2006) An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd international conference on machine learning. ACM, New York, pp 161–168

Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2(3:27):1–27

Crammer K, Gilad-Bachrach R, Navot A, Tishby A (2003) Margin analysis of the LVQ algorithm. In: Becker S, Thrun K, Obermayer K (eds.) Advances in neural information processing (Proc. NIPS 2002), vol 15. MIT Press, Cambridge, pp 462–469

Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge

Duda R, Hart P (1973) Pattern classification and scene analysis. Wiley, New York

Fritzke B (1995) A growing neural gas network learns topologies. In: Tesauro G, Touretzky DS, Leen TK (eds) Advances in neural information processing systems, vol 7. MIT Press, Cambridge, pp 625–632

Günther P, Villmann T, Hermann W (2011) Event related potentials and cognitive evaluation in Wilson's disease with and without neurological manifestation. J Neurol Sci [Turkish] 28(1):79–85

Gu Z, Shao M, Li L, Fu Y (2012) Discriminative metric: Schatten norms vs. vector norm. In: Proceedings of the 21st international conference on pattern recognition (ICPR 2012), pp 1213–1216

Hammer B, Nebel D, Riedel M, Villmann T (2014) Generative versus discriminative prototype based classification. In: Villmann T, Schleif FM, Kaden M, Lange M (eds) Advances in self-organizing maps and learning vector quantization: proceedings of 10th international workshop WSOM 2014, Mittweida. Advances in intelligent systems and computing, vol 295. Springer, Berlin, pp 123–132

Hammer B, Strickert M, Villmann T (2005) On the generalization ability of GRLVQ networks. Neural Process Lett 21(2):109–120

Hammer B, Strickert M, Villmann T (2005) Supervised neural gas with general similarity measure. Neural Process Lett 21(1):21–44

Hammer B, Villmann T (2002) Generalized relevance learning vector quantization. Neural Netw 15(8–9):1059–1068

Hasenjäger M, Ritter H (1998) Active learning with local models. Neural Process Lett 7:107–117

Hasenjäger M, Ritter H, Obermayer K (1999) Active learning in self-organizing maps. In: Oja E, Kaski S (eds) Kohonen maps. Elsevier, Amsterdam, pp 57–70

Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Springer, Heidelberg

Haykin S (1994) Neural networks—a comprehensive foundation. IEEE Press, New York

Hermann W, Barthel H, Hesse S, Grahmann F, Kühn HJ, Wagner A, Villmann T (2002) Comparison of clinical types of Wilson's disease and glucose metabolism in extrapyramidal motor brain regions. J Neurol 249(7):896–901

Hermann W, Günther P, Wagner A, Villmann T (2005) Klassifikation des Morbus Wilson auf der Basis neurophysiologischer Parameter. Der Nervenarzt 76:733–739

Hermann W, Villmann T, Grahmann F, Kühn H, Wagner A (2003) Investigation of fine motoric disturbances in Wilson's disease. Neurol Sci 23(6):279–285

Hermann W, Villmann T, Wagner A (2003) Elektrophysiologisches Schädigungsprofil von Patienten mit einem Morbus Wilson'. Der Nervenarzt 74(10):881–887

Hermann W, Wagner A, Kühn HJ, Grahmann F, Villmann T (2005) Classification of fine-motoric disturbances in Wilson's disease using artificial neural networks. Acta Neurologica Scandinavia 111(6):400–406

Herrmann M, Bauer HU, Der R (1994) The 'perceptual magnet' effect: a model based on self-organizing feature maps. In: Smith LS, Hancock PJB (eds) Neural computation and psychology. Springer, Stirling, pp 107–116

Horn R, Johnson C (2013) Matrix analysis, 2nd edn. Cambridge University Press, Cambridge

Kaden M, Hermann W, Villmann T (2014) Optimization of general statistical accuracy measures for classification based on learning vector quantization. In: Verleysen M (ed) Proceedings of European symposium on artificial neural networks, computational intelligence and machine learning (ESANN'2014). i6doc.com, Louvain-La-Neuve, Belgium, pp 47–52

Kaden M, Lange M, Nebel D, Riedel M, Geweniger T, Villmann T (2014) Aspects in classification learning—review of recent developments in learning vector quantization. Found Comput Decis Sci 39(2):79–105

Klingner M, Hellbach S, Riedel M, Kaden M, Villmann T, Böhme HJ (2014) RFSOM—extending self-organizing feature maps with adaptive metrics to combine spatial and textural features for body pose estimation. In: Villmann T, Schleif FM, Kaden M, Lange M (eds) Advances in self-organizing maps and learning vector quantization: proceedings of 10th international workshop WSOM 2014, Mittweida. Advances in intelligent systems and computing, vol 295. Springer, Berlin, pp 157–166

Kohonen T (1986) Learning vector quantization for pattern recognition. Report TKK-F-A601, Helsinki University of Technology, Espoo, Finland

Kohonen T (1990) Improved versions of learning vector quantization. In: Proceedings of IJCNN-90, international joint conference on neural networks, San Diego, vol I. IEEE Service Center, Piscataway, pp 545–550

Kohonen T (1995) Self-organizing maps. Springer Series in Information Sciences, vol 30. Springer, Berlin. (Second Extended Edition 1997)

Kohonen T, Kangas J, Laaksonen J, Torkkola K (1992) LVQ_PAK: a program package for the correct application of Learning Vector Quantization algorithms. In: Proceedings of IJCNN'92, international joint conference on neural networks, vol I. IEEE Service Center, Piscataway, pp 725–730

Martinetz T, Schulten K (1994) Topology representing networks. Neural Netw 7(2)

Martinetz TM, Berkovich SG, Schulten KJ (1993) 'Neural-gas' network for vector quantization and its application to time-series prediction. IEEE Trans Neural Netw 4(4):558–569

Mercer J (1909) Functions of positive and negative type and their connection with the theory of integral equations. Philos Trans R Soc Lond A 209:415–446

Micchelli C, Xu Y, Zhang H (2006) Universal kernels. J Mach Learn Res 7(26):051–2667

Nova D, Estévez P (2013) A review of learning vector quantization classifiers. Neural Comput Appl. doi:10.1007/s00521-013-1535-3

Qin A, Suganthan P (2004) A novel kernel prototype-based learning algorithm. In: Proceedings of the 17th international conference on pattern recognition (ICPR'04), vol 4, pp 621–624

Sachs L (1992) Angewandte statistik, 7th edn. Springer, Berlin

Sato A, Tsukumo J (1994) A criterion for training reference vectors and improved vector quantization. In: Proceedings of ICNN'94, international conference on neural networks. IEEE Service Center, Piscataway, pp 161–166

Sato A, Yamada K (1996) Generalized learning vector quantization. In: Touretzky DS, Mozer MC, Hasselmo ME (eds) Advances in neural information processing systems, vol 8. Proceedings of the 1995 conference. MIT Press, Cambridge, pp 423–429

Sato A, Yamada K (1995) A proposal of generalized learning vector quantization. Tech Rep IEICE 95(346):161–166

Schatten R (1950) A theory of cross-spaces. Annals of Mathematics Studies, vol 26. Princeton University Press, Princeton

Schleif FM, Hammer B, Villmann T (2007) Margin-based active learning for LVQ networks. Neurocomputing 70(7–9):1215–1224

Schleif FM, Villmann T, Hammer B, Schneider P (2011) Efficient kernelized prototype based classification. Int J Neural Syst 21(6):443–457

Schleif FM, Villmann T, Kostrzewa M, Hammer B, Gammerman A (2009) Cancer informatics by prototype networks in mass spectrometry. Artif Intell Med 45(2–3):215–228

Schölkopf B, Smola A (2002) Learning with kernels. MIT Press, Cambridge

Schneider P, Bunte K, Stiekema H, Hammer B, Villmann T, Biehl M (2010) Regularization in matrix relevance learning. IEEE Trans Neural Netw 21(5):831–840

Schneider P, Hammer B, Biehl M (2009a) Adaptive relevance matrices in learning vector quantization. Neural Comput 21:3532–3561

Schneider P, Hammer B, Biehl M (2009b) Distance learning in discriminative vector quantization. Neural Comput 21:2942–2969

Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis and discovery. Cambridge University Press, Cambridge

Steinwart I (2001) On the influence of the kernel on the consistency of support vector machines. J Mach Learn Res 2:67–93

Strickert M (2011) Enhancing M|G|RLVQ by quasi step discriminatory functions using 2nd order training. Machine Learning Reports 5 (MLR-06-2011), pp 5–15. ISSN: 1865–3960. http://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_06_2011.pdf

Villmann T (2002) Neural maps for faithful data modelling in medicine—state of the art and exemplary applications. Neurocomput 48(1–4):229–250

Villmann T, Geweniger T, Kästner M (2012) Border sensitive fuzzy classification learning in fuzzy vector quantization. Mach Learn Rep 6(MLR-06-2012):23–39. http://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_06_2012.pdf. ISSN:1865–3960

Villmann T, Haase S (2011) Divergence based vector quantization. Neural Computat 23(5):1343–1392

Villmann T, Haase S, Kaden M (2014) Kernelized vector quantization in gradient-descent learning. Neurocomputing (in press)

Villmann T, Haase S, Kästner M (2013) Gradient based learning in vector quantization using differentiable kernels. In: Estevez P, Principe J, Zegers P (eds) Advances in self-organizing maps: 9th international workshop WSOM 2012 Santiago de Chile. Advances in intelligent systems and computing, vol 198. Springer, Berlin, pp 193–204

Villmann T, Merényi E, Hammer B (2003) Neural maps in remote sensing image analysis. Neural Netw 16(3–4):389–403

Witoelar A, Gosh A, de Vries J, Hammer B, Biehl M (2010) Window-based example selection in learning vector quantization. Neural Comput 22(11):2924–2961

Wutzler U, Venner, Villmann T, Decker O, Ott U, Steiner T, Gumz A (2009) Recording of dissimulation and denial in the context of the psychosomatic evaluation at living kidney transplantation using the Minnesota Multiphasic Personality Inventory (MMPI). GMS Psycho Soc Med 6:1–11

Yin C, Mu S, Tian S (2012) Using cooperative clustering to solve multiclass problems. In: Wang Y, Li T (eds) Foundation of intelligent systems—proceedings of the sixth international conference on intelligent systems and knowledge engineering (ISKE 2011), Shanghai, China. Advances in intelligent and soft computing, vol. 122. Springer, Berlin, pp 327–334