FOUNDATIONS OF COMPUTING AND DECISION SCIENCES
Vol. 39 (2014) No. 2

DOI: 10.2478/fcds-2014-0006 ISSN 0867-6356
e-ISSN 2300-3405

Aspects in Classification Learning - Review of Recent Developments in Learning Vector Quantization

M. Kaden, M. Lange, D. Nebel, M. Riedel, T. Geweniger, and T. Villmann *

Abstract. Classification is one of the most frequent tasks in machine learning. However, the variety of classification tasks as well as classifier methods is huge. Thus the question is coming up: which classifier is suitable for a given problem or how can we utilize a certain classifier model for different tasks in classification learning. This paper focuses on learning vector quantization classifiers as one of the most intuitive prototype based classification models. Recent extensions and modifications of the basic learning vector quantization algorithm, which are proposed in the last years, are highlighted and also discussed in relation to particular classification task scenarios like imbalanced and/or incomplete data, prior data knowledge, classification guarantees or adaptive data metrics for optimal classification.

Keywords: learning vector quantization, non-standard metrics, classification, classification certainty, statistics

1 Introduction

Machine learning of complex classification tasks is still an challenging problem. The data sets may originate from different scientific fields like biology, medicine, finance and other. They can vary in several aspects like complexity/dimensionality, data structure/type, precision, class imbalances, prior knowledge to name just a few. Thus, the requirements for successful classifier models are multiple. They should be precise and stable in learning behavior as well as easy to understand and interpret. Additional features are desirable. To those eligible properties belong aspects of classification visualization, classification reasoning, classification significance and classification

*Computational Intelligence Group at the University of Applied Sciences Mittweida, Dept. of Mathematics, Technikumplatz 17, 09648 Mittweida, Saxonia - Germany,
corresponding author TV - email: thomas.villmann@hs-mittweida.de, www: https://www.mni.hs-mittweida.de/webs/villmann.html

certainties. Further, the classifier result should be independent on the certain realization of the data distribution but rather robust against noisy data and inaccurate learning samples. These properties are subsumed in the generalization ability of the model. Other model features of interest are the training complexity, the possibility of re-learning if new training data become available and a fast decision process for unknown data to be classified in the working phase.

Although, the task of classification learning seems to be simple and clearly defined as the minimization of the classification error or, equivalently, the maximization of the accuracy. This might be not the complete truth. In case of imbalanced contradicting training data of two classes, a good strategy to maximize the accuracy is to ignore the minor class and concentrate learning only to the major class. Those problems frequently occur in medicine and health sciences, where only a few samples are available for sick patients compared to the number of healthy persons. Another problem is that misclassifications for several classes may cause different costs. For example, patients suffering from a non-detected illness cause high therapy cost later whereas healthy persons misclassified as infected would require additional but cheaper medical tests. For those cases classification also has to deal with minimization of the respective costs in these scenarios. Thus, classifier models have to be designed to handle different classification criteria. Besides these objectives also other criteria might be of interest like classifier model complexity, the interpretability of the results or the suitability for real time applications [3].

According to these features, there exists a broad variety of classifiers ranging from statistical models like Linear and Quadratic Discriminant Analysis (LDA/QDA, [29, 76]) to adaptive algorithms like the Multilayer Perceptron (MLP, [75]), the k-Nearest Neighbor (kNN, [22]), Support Vector Machines (SVMs, [83]), or the Learning Vector Quantization (LVQ, [52]). SVMs were originally deigned only for two-class problems. For multi-class problems greedy strategy like cascades of one-versus-all approaches exist [41]. LDA and QDA are inappropriate for many non-linear classification tasks. MLPs converge slowly in learning in general and suffer from difficult model design (number of units in each layer, optimal number of hidden layers) [12]. Here deep architecture may offer an alternative [4]. Yet, the interpretation of the classification decision process in MLPs is difficult to explain based on the mathematical rule behind - they work more or less as black-box tools [41]. As an alternative, SVMs frequently achieve superior results and allow easy interpretation. SVMs belong to prototype-based models. They translate the classification task into a convex optimization problem based on the kernel trick, which consists in an implicit mapping of the data into a maybe infinite-dimensional kernel-mapping space [24, 93]. Non-linear problems can be resolved using non-linear kernels [83]. Classification guarantees are given in terms of margin analysis [100, 101], i.e. SVMs maximize the separation margin [40]. The decision process is based on the prototypes, determined during the learning phase. These prototypes are called support vectors and are data points defining the class borders in the mapping space and, hence, are not class-typical. The disadvantage of SVM models is their model complexity, which might be large for complicate classification tasks compared to the number of training samples. Further, a control of the complexity by relaxing strategies is difficult [50].

A classical and one of the most popular classification methods is the k-Nearest-Neighbor (kNN) approach [22, 26], which can achieve close to Bayes optimal classification if k is selected appropriately [40]. Drawbacks of this approach are the sensitivity with respect to outliers and the resulting risk of overfitting and the computational effort in the working phase. There exist several approaches to reduce these problems using condensed training sets and improved selection strategies [18, 39, 110] as pointed out in [9]. Nevertheless, kNN frequently serves as a baseline.

LVQs as introduced by T. KOHONEN can be seen as nearest neighbor classifiers based on a predefined set of prototypes optimized during learning and serving as reference set [53]. More precisely, the nearest neighbor paradigm becomes a *nearest prototype principle* (NPP). Although, the basic LVQ schemes are heuristically motivated approximating a Bayes decision, LVQs are one of the most successful classifiers [52]. A variant of this scheme is the Generalized LVQ (GLVQ,[77]), which keeps the basic ideas of the intuitive LVQ but introduces a cost function approximating the overall classification, which is optimized by gradient descent learning. LVQs are easy to interpret and the prototypes serve as class-typical representatives of their classes under certain conditions. GLVQ also belong to margin optimizer based on the hypothesis margin [23]. The hypothesis margin is related to the distance that the prototypes can be altered without changing the classification decision [68]. Therefore, GLVQ can be seen as an alternative to SVMs [34, 35].

In the following we will review the developments of LVQ-variants for classification task proposed during the last years in relation to several aspects of classification learning. Naturally, this collection of aspects cannot be complete. But at least, it highlights some of the most relevant aspects. Just before, we give a short explanation of the basic LVQ variants and GLVQ.

2 Basic LVQ variants

In this section we briefly give the basic variants of LVQ to justify notations and descriptions. We suppose to have a training data set of vectors $\mathbf{v} \in V \subseteq \mathbb{R}^n$ and let N_V denote the cardinality of 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 $x(\mathbf{v}) \in \mathcal{C} = \{1, \ldots, C\}$. The prototypes are labeled by $y(\mathbf{w}_k) \in \mathcal{C}$ such that each class is represented by at least one prototype.

One can distinguish at least two main branches of LVQ the margin optimizer and the probabilistic variants [68]. The basic schemes for both variants are explained in the following.

2.1 LVQ as Margin Optimizer

Now we assume a dissimilarity measure $d(\mathbf{v}, \mathbf{w}_k)$ in the data space, frequently but not necessarily chosen as the squared Euclidean distance

$$d_E(\mathbf{v}, \mathbf{w}_k) = (\mathbf{v} - \mathbf{w}_k)^2 = \sum_{j=1}^{n} (v_j - w_j)^2 . \tag{1}$$

According to the nearest prototype principle (NPP), let $\mathbf{w}^+$ denote the nearest prototype for a given data sample (vector) $\mathbf{v}$ according to the dissimilarity measure d with $y(\mathbf{w}^+) = x(\mathbf{v})$, i.e. the best matching prototype with correct class label also shortly denoted as best matching correct prototype. We define $d^+(\mathbf{v}) = d(\mathbf{v}, \mathbf{w}^+)$ as the respective dissimilarity degree. Analogously, $\mathbf{w}^-$ is the best matching prototype with a class label $y(\mathbf{w}^-)$ different from $x(\mathbf{v})$, i.e. best matching incorrect prototype, and $d^-(\mathbf{v}) = d(\mathbf{v}, \mathbf{w}^-)$ is again the assigned dissimilarity degree, see Fig.1.

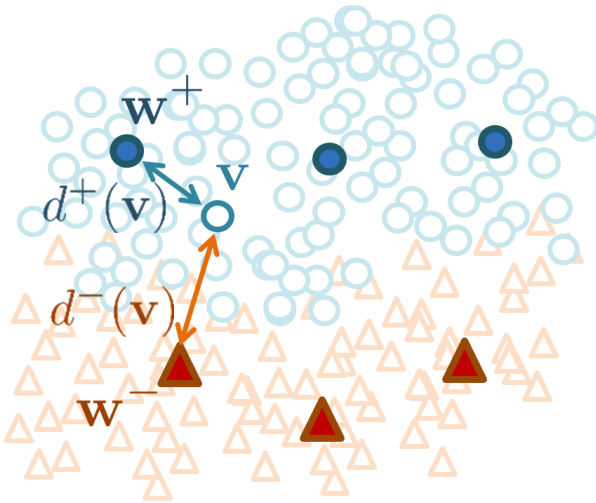

Figure 1: Illustration of the winner determination of $\mathbf{w}^+$, the best matching correct prototype and the best matching incorrect prototype $\mathbf{w}^-$together with their distances $d^+(\mathbf{v})$ and $d^-(\mathbf{v})$, respectively. The overall best matching prototype here is $\mathbf{w}^*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

Further, let

$$\mathbf{w}^* = \mathrm{argmin}_{\mathbf{w_k} \in \mathrm{W}} \left(\mathrm{d}(\mathbf{v}, \mathbf{w_k}) \right) \tag{2}$$

indicate the overall best matching prototype (BMP) without any label restriction accompanied by the dissimilarity degree $d^*(\mathbf{v}) = d(\mathbf{v}, \mathbf{w}^*)$. Hence, $\mathbf{w}^* \in \{\mathbf{w}^+, \mathbf{w}^-\}$. [1] Further, let be $y^* = y(\mathbf{w}^*)$. Thus the response of the classifier during the working

[1] Formally, w^* depends on $\mathbf{v}$, i.e $w^* = w^*(\mathbf{v})$. We omit this dependency in the notation but keep it always in mind.

phase is y^* obtained by the competition (2). According to the BMP for each data sample, we obtain a partition of the data space into receptive fields defined as

$$R(\mathbf{w}_k) = \{\mathbf{v} \in V | \mathbf{w}_k = \mathbf{w}^*\} \tag{3}$$

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 $R(\mathbf{w}_i) \cap R(\mathbf{w}_j) \neq \emptyset$ and zero elsewhere. For given prototypes and data sample the graph can be estimated using $\mathbf{w}^*$ and

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

as the second best matching prototype [59].

LVQ algorithms constitute a competitive learning according to the NPP over the randomized order of the available training data samples based on the basic intuitive principle *attraction and repulsion* of prototypes depending on their class agreement for a given training sample.

LVQ1 as the most simple LVQ only updates the BMP depending on the class label evaluation

$$\triangle \mathbf{w}^* = -\varepsilon \cdot \Psi(x(\mathbf{v}), y^*) \cdot (\mathbf{v} - \mathbf{w}^*) \tag{4}$$

with $0 < \varepsilon \ll 1$ being the learning rate. The adaptation

$$\mathbf{w}^* := \mathbf{w}^* - \triangle \mathbf{w}^* \tag{5}$$

realizes the Hebbian learning as a vector shift. The value

$$\Psi(x(\mathbf{v}), y^*) = \delta_{x(\mathbf{v}),y^*} - \left(1 - \delta_{x(\mathbf{v}),y^*}\right) \tag{6}$$

determines the direction of the vector shift $\mathbf{v} - \mathbf{w}^*$ where $\delta_{x(\mathbf{v}),y^*}$ is the Kronecker symbol such that $\delta_{x(\mathbf{v}),y^*} = 1$ for $x(\mathbf{v}) = y^*$ and zero elsewhere. The update (4) describes a *Winner Takes All* (WTA) rule moving the BMP closer to or away from the data vector if their class labels agree or disagree, respectively. Formally it can be written as

$$\triangle \mathbf{w}^* = \varepsilon \cdot \Psi(x(\mathbf{v}), y^*) \cdot \frac{1}{2} \cdot \frac{\partial d_E(\mathbf{v}, \mathbf{w}^*)}{\partial \mathbf{w}^*} \tag{7}$$

relating them to the derivative of $d_E(\mathbf{v}, \mathbf{w}^*)$. LVQ2.1 and LVQ3 differ from LVQ1 in this way that also the second best matching prototype is considered or adaptive learning rates come into play, for a detailed description we refer to [52].

As previously mentioned, the basic LVQ-models introduced by Kohonen are only heuristically motivated to approximate a Bayes classification scheme in an intuitive manner. Therefore, Sato&Yamada proposed a variant denoted as *Generalized* LVQ (GLVQ,[77]), such that stochastic gradient descent learning becomes available. For this purpose a classifier function

$$\mu(\mathbf{v}) = \frac{d^+(\mathbf{v}) - d^-(\mathbf{v})}{d^+(\mathbf{v}) + d^-(\mathbf{v})} \tag{8}$$

is introduced, where $\mu(\mathbf{v}) \in [-1, 1]$ is valid and correct classification corresponds to $\mu(\mathbf{v}) < 0$. The resulting cost function to be minimized is

$$E_{GLVQ}(W, V) = \frac{1}{2 \cdot N_V} \sum_{\mathbf{v} \in V} f(\mu(\mathbf{v})) \tag{9}$$

where f is a monotonically increasing transfer or squashing function frequently chosen as the identity function $f(x) = id(x) = x$ or a sigmoid function like

$$f_\Theta(x) = \frac{1}{1 + \exp\left(-\frac{x}{2\Theta^2}\right)} \tag{10}$$

with the parameter Θ determining the slope [109], see Fig.(2).

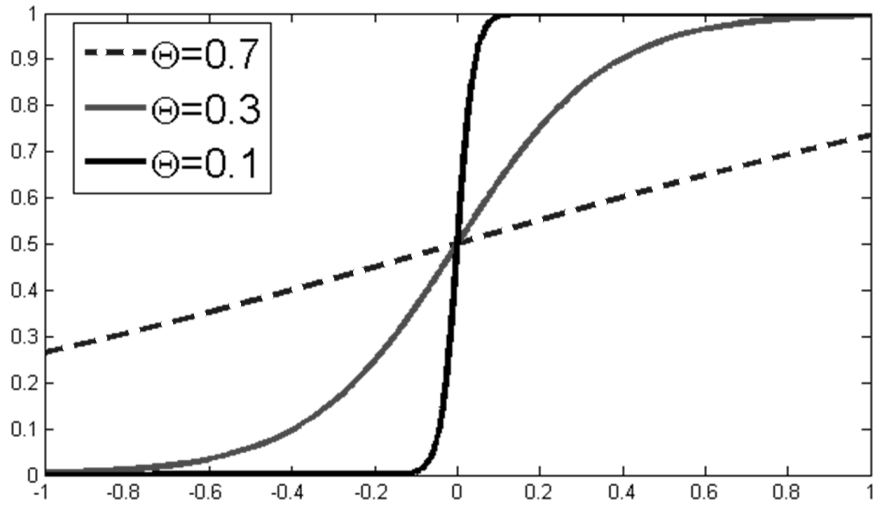

Figure 2: Shape of the sigmoid function $f_\Theta(x)$ from (10) depending on the slope parameter Θ

As before, N_V denotes the cardinality of the data set V. The prototype update, realized as a stochastic gradient descent step, writes as

$$\triangle \mathbf{w}^\pm \propto \varepsilon \cdot \xi^\pm \cdot \frac{\partial d_E^\pm(\mathbf{v})}{\partial \mathbf{w}^\pm} \tag{11}$$

with

$$\xi^\pm = \frac{\partial f}{\partial \mu} \cdot \frac{\partial \mu}{\partial d_E^\pm} = \mp 2 \cdot \frac{\partial f}{\partial \mu} \cdot \frac{d^\mp(\mathbf{v})}{\left(d^+(\mathbf{v}) + d^-(\mathbf{v})\right)^2} \tag{12}$$

for both $\mathbf{w}^+$ and $\mathbf{w}^-$. Again we observe that the update of the prototypes follows the basic principle of LVQ-learning and also involves the derivative of the dissimilarity measure.

As shown in [23], GLVQ maximizes the hypothesis margin, which is associated with the generalization error bound independent from the data dimension but depending on the number of prototypes.

2.2 Probabilistic variants of LVQ

Two probabilistic variants of LVQ were proposed by SEO&OBERMAYER. Although independently introduced, they are closely related. The first one, Soft Nearest Prototype Classifier (SNPC, [89]) is also based on the NPP. We consider probabilistic assignments

$$u_\tau\left(k|\mathbf{v}\right) = \frac{\exp\left(-\frac{d_E(\mathbf{v},\mathbf{w}_k)}{2\tau^2}\right)}{\sum_{j=1}^{M}\exp\left(-\frac{d_E(\mathbf{v},\mathbf{w}_j)}{2\tau^2}\right)} \tag{13}$$

that a data vector $\mathbf{v} \in V$ is assigned to the prototype $\mathbf{w}_k \in W$. The parameter τ determines the width of the Gaussian and should be chosen in agreement with the variance of the data.

In medicine, medical doctors judge the proximity of patients to given standards and define *local costs*

$$lc\left(\mathbf{v}, W\right) = \sum_{k=1}^{M} u_\tau\left(k|\mathbf{v}\right) \cdot \left(1 - \delta_{x(\mathbf{v}),y(\mathbf{w}_k)}\right) \tag{14}$$

for classification of this training sample. The cost function of SNPC is

$$E_{SNPC}\left(W, V\right) = \sum_{\mathbf{v}\in V} lc\left(\mathbf{v}\right) \tag{15}$$

which can be optimized by stochastic gradient descent learning with respect to the prototypes.

A generative mixture model for LVQ with an explicit discriminative cost function has been proposed in [90] denoted as Robust Soft LVQ (RSLVQ). For this purpose, the probability that a data sample $\mathbf{v} \in V$ is generated by the prototype set W is introduced as

$$p(\mathbf{v}|W) = \sum_{j=1}^{M} p(\mathbf{w}_j) \cdot p(\mathbf{v}|\mathbf{w}_j) \tag{16}$$

with prior probabilities $p(\mathbf{w}_j)$ typically chosen as constant and the conditional probabilities $p(\mathbf{v}|\mathbf{w}_j)$ determined as $p(\mathbf{v}|\mathbf{w}_j) = u_\tau\left(j|\mathbf{v}\right)$ for Euclidean data and depending on the Gaussian width τ. Taking the labels into account we have

$$p(\mathbf{v}, x\left(\mathbf{v}\right)|W) = \sum_{j=1}^{M} \delta_{x(\mathbf{v}),y(\mathbf{w}_j)} \cdot p(\mathbf{w}_j) \cdot p(\mathbf{v}|\mathbf{w}_j) \tag{17}$$

such that marginalization gives $p(x\left(\mathbf{v}\right)|W) = \sum_{j=1}^{M} \delta_{x(\mathbf{v}),y(\mathbf{w}_j)} \cdot p(\mathbf{w}_j)$. This yields

$$p(x\left(\mathbf{v}\right)|\mathbf{v}, W) = \frac{p(\mathbf{v}, x\left(\mathbf{v}\right)|W)}{p(\mathbf{v}|W)} \tag{18}$$

as class probability. For i.i.d. data the cost function to be minimized in RSLVQ is the sum of the log-likelihood ratios

$$E_{RSLVQ}(W,V) = \sum_{\mathbf{v}\in V} \ln p(x\,(\mathbf{v})\,|\mathbf{v},W) \tag{19}$$

which can be optimized again by stochastic gradient descent learning for Euclidean data.

Both probabilistic approaches keep the basic LVQ-learning principle of attraction and repulsion, we refer to [90, 89].

3 Characterization of Classification Tasks and their Relation to LVQ-variants

In this section we will collect and characterize problems and tasks related to classification learning and provide respective LVQ variants. Further, we consider aspects of appropriate dissimilarities and respective LVQ-variants, if structural knowledge about the data is available or if restrictions apply. Yet, this collection is neither assumed to be complete nor comprehensive. The aim is just to show that these issues can be treated by variants of the basic LVQ schemes.

3.1 Structural Aspects for Data Sets and Appropriate Dissimilarities

3.1.1 Restricted Data - Dissimilarity Data

For most of the LVQ-schemes, vector data are supposed. Yet, non-vectorized occur in many applications, e.g. text classification, categorical data, or gene sequences. Those data can be handled by embedding techniques applied in LVQ or by median variants, if the pairwise dissimilarities collected in the dissimilarity matrix $\mathbf{D} \in \mathbb{R}^{N\times N}$ are provided. For example, one popular method to generate such dissimilarities for text data (or gene sequences) is the normalized compression distance [21]. The eigenvalues of $\mathbf{D}$ determine, whether an embedding is possible: Let be n_+, n_-, n° be the number of positive, negative and zero eigenvalues of (symmetric) $\mathbf{D}$ collected in the signature vector $\Sigma = (n_+, n_-, n^\circ)$ and $D_{ii} = 0$. If $n_- = n^\circ = 0$ an Euclidean embedding is always possible and prototypes are the convex linear combination $\mathbf{w}_k = \sum_{j=1}^{N} \alpha_{kj}\mathbf{v}_j$ with $\alpha_{kj} \geq 0$ and $\sum_{j=1}^{N} \alpha_{kj} = 1$ [5]. The squared Euclidean distances between data samples and prototypes can be calculated as

$$d_{\mathbf{D}}\,(\mathbf{v}_j, \mathbf{w}_k) = [\mathbf{D}\alpha_k]_j - \frac{1}{2}\alpha_k^\top \mathbf{D}\alpha_k$$

and replace the d_E in the above cost function for GLVQ. Gradient descent learning can then be carried out as gradient learning for the coefficient vectors α_k using the

derivatives $\frac{\partial d_{\mathbf{D}}(\mathbf{v}_j,\mathbf{w}_k)}{\partial \alpha_k}$ [112]. This methodology is also referred as *relational learning* paradigm. If such an embedding is not possible or does not show a reasonable meaning, median variants have to be applied, i.e. the prototypes have to be restricted to be data samples. Respective variants for RSLVQ and GLVQ based on a generalized Expectation-Maximization (EM) scheme are proposed in [64, 66]. The respective median approach for SNPC is considered in [65].

Examples for those dissimilarities or metrics, which are not differentiable, are the edit distance or compression distance based on the Kolmogorov-complexity for text comparisons [21], or *locality improved kernels* (LIK-kernels) used in gene analysis [36].

3.1.2 Structurally Motivated Dissimilarities

If additional knowledge about data is available it might be advantageously to make use of this information. For vectorial data $\mathbf{v} \in \mathbb{R}^n$ representing discretized probability density functions $v(t) \geq 0$ with $v_j = v(t_k)$ and $\sum_{j=1}^{n} v_j = c = 1$, divergences $D(\mathbf{v}||\mathbf{w})$ may be a more appropriate dissimilarity measure than the Euclidean distance. For example, grayscale histograms of grayscale images can be seen as such discrete densities. More general, if we assume $c \geq 1$, the data vectors constitute discrete representations of positive measures and generalized divergences come into play, e.g. the generalized *Kullback-Leibler-divergence* (gKLD) is given by

$$D_{gKLD}(\mathbf{v}||\mathbf{w}) = \sum_{j=1}^{n} \left[v_j \cdot \log\left(\frac{v_j}{w_j}\right) - (v_j - w_j) \right] \tag{20}$$

as explained in [20]. For differentiable divergences $D(\mathbf{v}||\mathbf{w}_k)$ with respect to the prototype vector $\mathbf{w}_k$, it can be easily plugged into the above cost functions of the several LVQ-variants from Sec. 2 for stochastic gradient descent learning. The derivative for the generalized Kullback-Leibler-divergence is

$$\frac{\partial D_{gKLD}(\mathbf{v}||\mathbf{w})}{\partial \mathbf{w}} = -\frac{\mathbf{v}}{\mathbf{w}} + 1\,.$$

Other popular divergences are the *Rényi-divergence*

$$D_{\alpha}(\mathbf{v}||\mathbf{w}) = \frac{1}{\alpha - 1} \log\left(\sum_{j=1}^{n} v_j^{\alpha} w_j^{1-\alpha} \right) \tag{21}$$

applied in information theoretic learning (ITL, [70, 69]) with $\alpha > 0$ with the derivative

$$\frac{\partial D_{\alpha}(\mathbf{v}||\mathbf{w})}{\partial \mathbf{w}} = -\frac{\mathbf{v}^{\alpha} \circ \mathbf{w}^{-\alpha}}{\sum_{j=1}^{n} v_j^{\alpha} w_j^{1-\alpha}}$$

using the Hadamard product $\mathbf{v} \circ \mathbf{w}$, and the *Cauchy-Schwarz-divergence*

$$D_{CS}(\mathbf{v}||\mathbf{w}) = \frac{1}{2} \log\left[\left(\sum_{j=1}^{n} v_j^2 \right) \cdot \left(\sum_{j=1}^{n} w_j^2 \right) \right] - \log\left(\sum_{j=1}^{n} v_j w_j \right) \tag{22}$$

also proposed in ITL with the derivative

$$\frac{\partial D_{cs}\left(\mathbf{v}||\mathbf{w}\right)}{\partial \mathbf{w}} = \frac{\mathbf{w}}{\left(\sum_{j=1}^{n} w_j^2\right)} - \frac{\mathbf{v}}{\left(\sum_{j=1}^{n} v_j w_j\right)} \,.$$

An ITL-LVQ-classifier similar to SNPC based on the Rényi-divergence with $\alpha = 2$ as the most convenient case was presented in [98], whereas the Cauchy-Schwarz-divergence was used in a fuzzy variant of ITL-LVQ-classifiers in [106]. A comprehensive overview of differentiable divergences together with derivatives for prototype learning can be found in [102] and an explicit application for GLVQ was presented in [63].

In biology and medicine, frequently data vectors are compared in terms of a correlation measure $\varrho\left(\mathbf{v}, \mathbf{w}\right)$ [76, 97]. Most prominent correlation values are the *Spearman-rank-correlation* and the *Pearson-correlation*. The latter one is defined as

$$\varrho_P\left(\mathbf{v}, \mathbf{w}\right) = \frac{\sum_{k=1}^{n}\left(v_k - \mu_{\mathbf{v}}\right) \cdot \left(w_k - \mu_{\mathbf{w}}\right)}{\sqrt{\sum_{k=1}^{n}\left(v_k - \mu_{\mathbf{v}}\right)^2 \cdot \sum_{k=1}^{n}\left(w_k - \mu_{\mathbf{w}}\right)^2}} \tag{23}$$

with $\mu_{\mathbf{v}} = \frac{1}{n}\sum_{j=1}^{n} v_i$ and $\mathbf{w}$ defined analogously. The Pearson-correlations is differentiable according to

$$\frac{\partial \varrho_P\left(\mathbf{v}, \mathbf{w}\right)}{\partial \mathbf{w}} = \varrho_P\left(\mathbf{v}, \mathbf{w}\right) \cdot \left(\frac{1}{\mathcal{B}}\mathbf{v} - \frac{1}{\mathcal{D}}\mathbf{w}\right) \tag{24}$$

with the abbreviations $\mathcal{B} = \sum_{k=1}^{n}\left(v_k - \mu_{\mathbf{v}}\right) \cdot \left(w_k - \mu_{\mathbf{w}}\right)$ and $\mathcal{D} = \sum_{k=1}^{n}\left(w_k - \mu_{\mathbf{w}}\right)^2$ [97]. Therefore, the Pearson-correlations can immediately be applied in gradient based learning for the LVQ-classifiers [96] whereas Spearman-correlation needs an approximation technique, because an explicit derivation with respect to $\mathbf{w}$ does not exit due to the inherent rank function [95, 49]. Related to these approaches are covariances for the dissimilarity judgment, which were considered in the context of vector quantization learning in [62, 54].

3.2 Fuzzy Data and Fuzzy Classification Approaches related to LVQ

The processing of data with uncertain class knowledge for training samples and probabilistic classification of unknown data in the working phase of a classifier belong to the challenging tasks in machine learning and vector quantization. Standard LVQ and GLVQ are restricted to deal with exact class decisions for training data and return crisp decisions. Unfortunately, these requirements for training data are not always fulfilled due to uncertainties for those data. Yet, SNPC and RSLVQ allow processing of fuzzy data. For example, the local costs (14) in SNPC can be fuzzyfied replacing the the crisp decision realized according to the Kronecker-value $\delta_{x(\mathbf{v}),y(\mathbf{w}_k)}$ by fuzzy assignments $\alpha_{x(\mathbf{v}),y(\mathbf{w}_k)} \in [0, 1]$ [79, 107]. Information theoretic learning vector quantizers for fuzzy classification were considered in [106] and a respective RSLVQ investigation was proposed in [30, 85].

Otherwise, if the class label of the training data are fuzzy, further modification of LVQ approaches are required relaxing the strict assignments of prototypes to certain classes. This attempt was done in [31] for FSNPC. An alternative for those problems might be a combination of unsupervised vector quantization together with an supervised fuzzy classifier extension based on self-organizing maps (SOM, [51, 52]) and neural gas (NG, [60, 59]) as proposed in [105].

Comparison of fuzzy classification results is mandatory as for crisp classification. Therefore, reliable and compatible evaluation measures are necessary. Statistical measures like the κ- index or the κ-Fleiss-index for comparison of two and more classification solutions, respectively, are well-known and accepted in statistics for crisp classification [14, 17, 76]. Their extensions regarding fuzzy classifications are investigated in [32, 111].

4 Attempts to Improve the Classifier Performance

Several aspects can be identified to improve classifier performance. These issues are not only pure classification accuracy and false positive/negative rates but also comprise facets like interpretability and class representation, model size, classification guarantee and other [3, 45, 44]. In the following we will graze some of these aspects without any claim of completeness.

4.1 Robustness, Classification Certainty and Border Sensitivity

Several aspects can be identified when discussing robustness and assessment of classification certainty of a classifier model. For SVMs, most questions are answered by the underlying theory of convex optimization and structural risk minimization providing also generalization bounds [40, 83, 100]. For GLVQ generalization bounds were considered in [23, 35, 34]. However, LVQ-methods depend sensitively on the initialization of the prototypes optimized during the stochastic online learning process, which is in contrast to the well-determined convex optimization of applied in SVMs. Several attempts were proposed to make progress regarding this problem ranging from intelligent initialization to the *harmonic to minimum LVQ algorithm* (H2M-LVQ, [71]). This latter approach starts with a different cost function compared to GLVQ incorporating the harmonic average distance instead of d^+ and d^- in (9). According to this average, the whole distance information between the presented data sample and all prototypes is taken into account, which reduces the initialization sensitivity. During training progress, a smooth transition to the usual minimum distance GLVQ takes place to end up with standard GLVQ. A more intuitive approach for initialization insensitive GLVQ is to adopt the idea of neighborhood cooperativeness in neural maps also for the prototype in GLVQ. Thus, not only the best prototype $\mathbf{w}^+$ is updated but also all other prototypes of the correct class proportional to their dissimilarity degree, for example by a rank based scheme known from neural gas. The respective algorithm is denoted as supervised neural gas (SNG, [36]).

In mathematical statistics, classification and discrimination ability is also assessed in terms of significance level and confidence intervals. Beside the previously mentioned generalization bounds, the research for these aspects of LVQ schemes is underestimated so far [92, 1]. A related feature also to the confidence concept in statistics is conformal prediction which provides together with a classification decision of the classifier a value describing the certainty of the decision [91, 108]. A LVQ-realization was proposed in [82], a respective approach for SNG was presented in [81]. Another recently investigated approach is based on so-called *reject options* based on considerations for reject tradeoff for optimum recognition error [19]. Rejection measures return a value $r(\mathbf{v})$ indicating the certainty of the classification of a data point $\mathbf{v} \in V$. For example, in RSLVQ as a probabilistic classifier model we can take

$$r_{RSLVQ}(\mathbf{v}) = \mathrm{argmax}_k \left(p(y(\mathbf{w}_k) | \mathbf{v}, W) \right)$$

where lower values correspond to lower certainty [28]. For GLVQ one can choose

$$r_{GLVQ}(\mathbf{v}) = \frac{|d^+(\mathbf{v}) - d^-(\mathbf{v})|}{2 \|\mathbf{w}^+ - \mathbf{w}^-\|^2}$$

if the respective receptive fields from (3) are neighbored in Delaunay-graph $\mathcal{G}$, i.e.

$$R(\mathbf{w}^-) \cap R(\mathbf{w}^+) \neq \emptyset \tag{25}$$

is valid. Related to those rejection options is a class-dependent outlier detection with

$$r_{class-out}(\mathbf{v}) = -d^+(\mathbf{v})$$

as rejection measure.

For a certain classification it is important to detect precisely the class borders. In SVMs this concept is realized by the support vectors, which are extreme points of the class distributions. One aim of LVQ approaches is to represent the classes by class typical prototypes. For a more detail discussion see, Sec.4.2. Here we want to emphasize that class border sensitive LVQ-models can be demanded. The first attempt in this direction was the *window rule* for LVQ2.1. According to this ruel, learning takes only place if the training sample $\mathbf{v}$ falls into a window according to

$$\min \left(\frac{d^+(\mathbf{v})}{d^-(\mathbf{v})}, \frac{d^-(\mathbf{v})}{d^+(\mathbf{v})} \right) \geq \frac{1-\omega}{1+\omega} \tag{26}$$

in the variant LVQ2.1. Prototype adaptation only takes place if this relation is fulfilled for a predefined value $0 < \omega < 1$ [52], 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 [109]. Further, the window rule (26) 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 [52]. However, 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. $R(\mathbf{w}^+) \cap R(\mathbf{w}^-) \neq \emptyset$.

A more simple and intuitive border sensitive learning can be achieved in GLVQ. For this purpose, we consider the squashing function $f_\Theta(x)$ from (10) depending on the slope parameter Θ. The prototype update (11) is proportional to the derivative

$$f'_\Theta(\mu(\mathbf{v})) = \frac{f_\Theta(\mu(\mathbf{v}))}{2\Theta^2}(1 - f_\Theta(\mu(\mathbf{v})))$$

via the scaling factors $\xi^\pm$ from (12). For small slope parameter values $0 < \Theta \ll 1$ only those data points generate a non-vanishing update, for which the classifier function $\mu(\mathbf{v})$ from (8) is close enough to zero [8], i.e. the data sample is close to a class border, see Fig. (3).

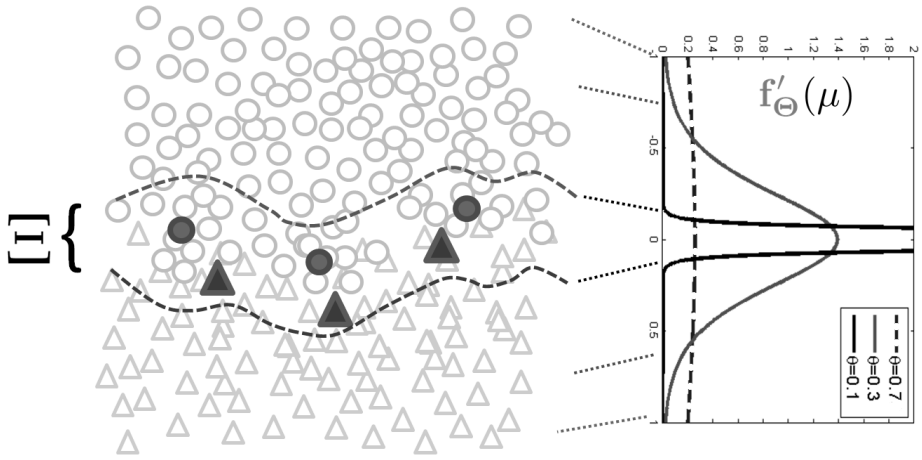

Figure 3: Illustration of the border-sensitive LVQ.

The respective data points are denoted as *active set* Ξ contributing to the prototype learning. Thus, the active set determines the border sensitivity of the GLVQ-model. In consequence, small Θ-values realize *border sensitive learning* for GLVQ and prototypes are certainly forced to move to the class borders [48].

4.2 Generative versus Discriminative models, Asymmetric error Assessment and Statistical Classification by LVQ-models

As pointed out in [73], there is a discrepancy between generative and discriminative features in prototype-based classification schemes, in particular for class overlapping data. The generative aspects reflect the class-wise representation of the data by the respective class prototypes emphasizing interpretable prototypes, whereas the discriminative part ensures best possible class separability. In LVQ-models, discriminative part is mainly realized by the repellent prototype update for the best matching incorrect prototype $\mathbf{w}^-$ as for example in LVQ2.1 or GLVQ, which can be seen as a kind learning from mistakes [87]. The generative aspect is due to the attraction of the best

matching prototype $\mathbf{w}^+$ with correct class label. A detailed consideration of balancing both aspects for GLVQ and RSLVQ can be found in [73]. There, the balancing is realized by a decomposition of the cost functions into a generative and a discriminative part. For example, the generative part in GLVQ for class representation takes into account the class-wise quantization error

$$E_{GLVQ}^{repr}(W,V) = \frac{1}{2}\sum_{\mathbf{v}\in V} d^+(\mathbf{v})$$

adopted from unsupervised vector quantization, whereas the original GLVQ cost function $E_{GLVQ}(W,V)$ from (9) plays the role of the discriminative part [73]. Combining both aspect yields a different weighting of $d^+(\mathbf{v})$ and $d^-(\mathbf{v})$.

Other weighting and scaling may emphasize other aspects. Class-dependent weighting and asymmetric error assessment of $f(\mu(\mathbf{v}))$ in GLVQ by a composed scaling factor

$$s(\mathbf{v}) = \beta(x(\mathbf{v})) \cdot \gamma(y(\mathbf{w}^-), x(\mathbf{v}))$$

was suggested in [46], where $\beta(x(\mathbf{v})) > 0$ are class-priors weighting the misclassifications of classes in the cost function 9. The $\gamma(y(\mathbf{w}^-), x(\mathbf{v}))$-factor allows to model class-dependent misclassification cost and thus enable to integrate asymmetric misclassification costs.

Related to these aspects is the *Receiver Operating Characteristic* (ROC, [15, 27]) for balancing the efficiency (true positive rate - TP-rate) and false positive rate (FP-rate), see Fig.(4).

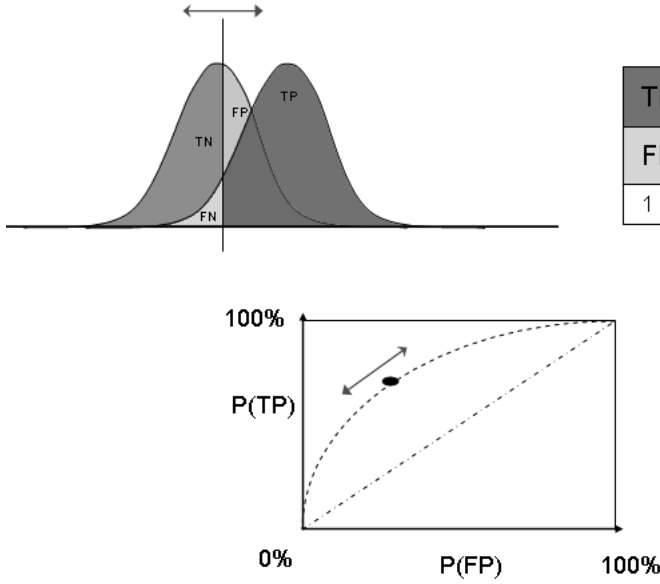

Figure 4: Illustration of the Receiver Operating Characteristic and the confusion matrix with *true positives* (TP), *false positives* (FP), f*alse negatives* (FN) and *true negatives* (TN) (from: http://de.wikipedia.org/wiki/Receiver_Operating_Characteristic, 15.01.2014).

ROC analysis plays an important role in *binary* classification assessment in particular in medicine and social sciences. Originally, ROC is an important tool for performance comparison of classifiers [27].

Recent successful LVQ/GLVQ approaches for medical applications also utilize this methodology for improved LVQ analysis and classifier comparison [6, 7, 11]. In particular, ROC-curves are considered to be an appropriate tool for classifier performance comparison [15], which are based on the evaluation of true and false positive rates. Frequently, the area under ROC-curve (AUC) is calculated as the respective statistical quantity for comparison [43, 67, 99].

Unfortunately, original GLVQ as proposed in [77] does not optimize the classification error rather than the cost function $E_{GLVQ}(W, V)$ from (9). Hence, the performance cannot judged consistently neither in terms of the statistical quantities provided by the confusion matrix nor by the ROC analysis.

However, if the parametrized sigmoid function $f_\Theta(x)$ is used for GLVQ, then the cost function becomes Θ-dependent $E_{GLVQ}(W, V, \Theta)$. It turns out that for $\Theta \searrow 0$ the sigmoid function $f_\Theta(x)$ converges to the Heaviside function $H(x)$ such that the cost function E_{GLVQ} approximates the *misclassification rate*. Using this observation

one can redefine the classifier function as

$$\mu_\Theta(\mathbf{v}) = f_\Theta(-\mu(\mathbf{v})) \tag{27}$$

with $\mu_\Theta(\mathbf{v}) \approx 1$ iff the data point $\mathbf{v}$ is correctly classified and $\mu_\Theta(\mathbf{v}) \approx 0$ otherwise, such that the new cost function $E_{GLVQ}(\mu_\Theta(\mathbf{v}))$ approximates the *classifications accuracy*

$$AC = \frac{TP + TN}{N_V} \tag{28}$$

with TP and TN are the number of true positives and true negatives, respectively, as considered in Fig.4. Again, N_V is the cardinality of the full data set V. In a similar way all quantities of a confusion matrix (see Fig. (4)) and combinations thereof can be obtained as a cost function for a GLVQ-like classifier keeping the idea of prototype learning [48]. In particular, many statistical quantities used in medicine, bioinformatics and social sciences for classification assessment like *precision* π and *recall* ρ defined by

$$\pi = \frac{TP}{TP + FP} \text{ and } \rho = \frac{TP}{TP + FN}$$

can be explicitly optimized by a GLVQ-like classifier. Also the well-known F_β-measure

$$F_\beta = \frac{\left(1 + \beta^2\right) \cdot \pi \cdot \rho}{\beta^2 \cdot \pi + \rho} \tag{29}$$

developed by C.J. van Rijsbergen [74] and frequently applied in engineering can serve as a cost function in this scheme [44]. For the common choice $\beta = 1$, F_β is the fraction of the harmonic and the arithmetic mean of precision and recall, i.e. $\beta > 0$ controls the balance of both values.

Further, we can draw the conclusion that with this statistical GLVQ-interpretation, a classifier evaluation in terms of statistical quality measures based on the confusion matrix as well as ROC-analysis becomes a consistent framework. As mentioned above, ROC-curve comparison is usually done investigating the respective AUC-differences. Other investigation focus on precision-recall-curves [25].

Recently, a GLVQ-approach for direct optimization of the AUC was proposed in [10]. This approach directly optimizes AUC using the probability interpretation of AUC as emphasized in [27, 38].

4.3 Appropriate Metrics and Metric Adaptation for Vector Data

Beside the data structure dependent dissimilarities and metric already discussed in Sec. (3.1.2), we now briefly consider non-standard (non-Euclidean) metrics for vector and matrix data, which can be used in LVQ-classifiers for appropriate separation. Thereby, one fascinating behavior of parametrized metrics is the possibility of a task dependent adaptation to achieve a better classification performance. For LVQ-classifiers, this topic was initialized by the pioneering works [13] and [37] about *relevance learning* in GLVQ denoted as *Generalized Relevance LVQ* (GRLVQ). In

this work the usually applied squared Euclidean metric in GLVQ is replaced by the weighted variant

$$d_\lambda(\mathbf{v},\mathbf{w}) = \sum_{j=1}^{n} \lambda_j^2 (v_j - w_j)^2 \tag{30}$$

with the normalization $\sum_{j=1}^{n} \lambda_j^2 = 1$. Together with the prototype adaptation for a presented training sample $\mathbf{v}$ with label $x(\mathbf{v})$, also the relevance weights λ_j are optimized according to

$$\triangle\lambda_j \propto \lambda_j \left[\xi^+ \left(v_j - w_j^+\right)^2 - \xi^- \left(v_j - w_j^-\right)^2\right] \tag{31}$$

to improve the classification performance. Here we applied the stochastic gradient $\frac{\partial E_{GRLVQ}}{\partial \lambda_j}$ and the derivative

$$\frac{\partial d_\lambda(\mathbf{v},\mathbf{w})}{\partial \lambda_j} = 2\lambda_j (v_j - w_j)^2 \ . \tag{32}$$

The generalization of this relevance learning is the matrix variant (abbreviated as GMLVQ) using the metric

$$d^\Omega(\mathbf{v},\mathbf{w}) = \sum_{i=1}^{m} \left([\Omega(\mathbf{v}-\mathbf{w})]_i\right)^2 \tag{33}$$

with $\Omega \in \mathbb{R}^{m\times n}$ as a linear data mapping [86, 87, 16]. The derivative reads as

$$\frac{\partial d^\Omega(\mathbf{v},\mathbf{w})}{\partial \Omega_{kl}} = 2 \cdot [\Omega(\mathbf{v}-\mathbf{w})]_k [\mathbf{v}-\mathbf{w}]_l$$

where $[\mathbf{v}-\mathbf{w}]_j$ denotes the jth component of the vector $\mathbf{v}-\mathbf{w}$. For quadratic $\Omega \in \mathbb{R}^{n\times n}$ the regularization condition $\det(\Lambda) = \det\left(\Omega^\top\Omega\right)$ has to be enforced [84].

Many interesting variants have been proposed including prototype- or class-specific matrices. Recently, the extension to vector Minkowski-p-metrics

$$d_p^\Omega(\mathbf{v},\mathbf{w}) = \sqrt[p]{\sum_{i=1}^{m} (|z_i|)^p} \tag{34}$$

were considered in [57, 58] with the linear mapping

$$\mathbf{z} = \Omega(\mathbf{v}-\mathbf{w}) \ . \tag{35}$$

Minkowski-p-norms allow further flexibility according to their underlying unit balls Fig.5.

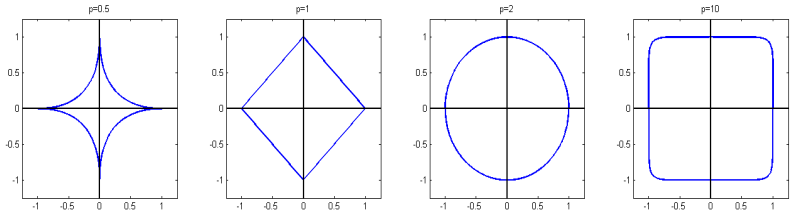

Figure 5: Unit balls for several Minkowski-p-norms $\|\mathbf{x}\|_p$ (34): from left to right $p = 0.5, \ldots = 1, \ldots = 2, \ldots = 10$.

In particular, all values $0 < p \leq \infty$ are allowed [55]. For example, values $p < 1$ emphasize small deviations. Thereby, for $p \neq 2$ the respective spaces are only Banach spaces, which are equiped with a semi-inner product instead of the usual Euclidean inner product.

Kernel distances became aware also for LVQ approaches due to the great success of SVMs. Positive definite kernel function $\kappa_\Phi(\mathbf{v}, \mathbf{w})$ correspond to kernel feature maps $\Phi : V \to \mathcal{I}_\Phi \subseteq \mathcal{H}$ in a canonical manner [2, 83]. The data are mapped into an associated Hilbert space $\mathcal{H}$ such that for the respective inner product $\langle \bullet, \bullet \rangle_{\mathcal{H}}$ in $\mathcal{H}$ the relation

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

is valid. Therefore, a kernel distance is defined by the inner product, which can be calculated as

$$d_{\kappa_\Phi}(\Phi(\mathbf{v}), \Phi(\mathbf{w})) = \sqrt{\kappa(\mathbf{v}, \mathbf{v})^2 - 2\kappa(\mathbf{v}, \mathbf{w}) + \kappa(\mathbf{w}, \mathbf{w})^2} \tag{36}$$

for images $\Phi(\mathbf{v})$ and $\Phi(\mathbf{w})$. First integration attempts of kernel distances into GLVQ were suggested in [72] and [80] using various approximation techniques to determine the gradient learning in the kernel associated Hilbert space $\mathcal{H}$. An elementary alternative is the utilization of *differentiable* universal kernels [103] based on the theory of universal kernels [61, 88, 94]. This approach allows the adaptation of the prototypes in the original data space but equipped with the kernel distance generated by the differentiable kernel, i.e. the metric space (V, d_{κ_Φ}) [104, 103]. Hence, such a distance is also differentiable according to (36), see Fig.6.

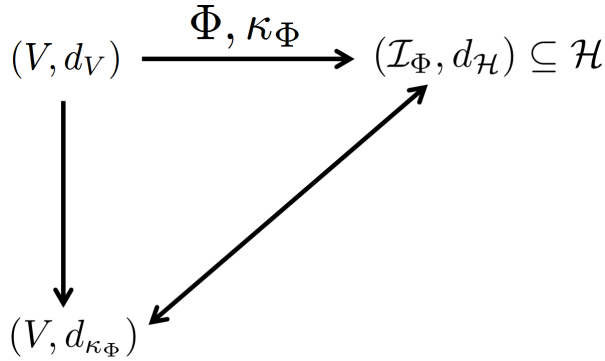

Figure 6: Utilization of differentiable kernels κ_Φ and respective kernel distances d_{κ_Φ} in vector quantization instead of the usual data metric d_V. SVMs operate in $\mathcal{I}_\Phi$ based on the inner product, whereas differentiable kernels may be applied directly in gradient descent learning for GLVQ living the metric space (V, d_{κ_Φ}).

For example, exponential kernels are universal, which can be used together with the above mentioned Minkowski-p-norms and the linear data mapping (35), revealing

$$\kappa_p^\Omega(\mathbf{v}, \mathbf{w}) = \exp\left(-\left(d_p^\Omega(\mathbf{v}, \mathbf{w})\right)^p\right)$$

as an adaptive kernel with kernel parameters Ω [47].

The natural extension of vector quantization is matrix quantization. For example, grayscale images of bacterial structures in biology have to be classified or hand written digit recognition. One possibility is to extract certain features related to the task. Another possibility would be to take the images as matrices and application of *matrix norms* for comparison. Matrix norms differ from usual norms by the additional property of sub-multiplicity $\|A \cdot B\| \leq \|A\| \cdot \|B\|$, such that the matrix norm becomes compliant with the matrix multiplication [42]. One of the most prominent class of matrix norms are *Schatten-p*-norms [78], which are closely related to Minkowski-p-norms. The Schatten-p-norm $s_p(A)$ of a matrix A is defined as

$$s_p(A) = \sqrt[p]{\sum_{k=1}^{n} (\sigma_k)^p} \tag{37}$$

where the $\sigma_k(A)$ are the singular values of A, i.e. the squared singular values $(\sigma_k(A))^2$ are the eigenvalues of $\Omega = A^* A \in \mathbb{R}^{n \times n}$ and where A^* denotes the conjugate complex of A [78]. With this matrix norm, the vector space of complex matrices $\mathbb{C}^{m \times n}$ becomes a Banach space $\mathfrak{B}_{m,n}$. As for vector norms, the value $p = 2$ is associated with a Hilbert space. Schatten-norms were considered for improved LVQ classification compared to vector norms in image data analysis in [33]. Further properties of the respective Banach spaces were studied in [56].

5 Summary

In this review paper we give a summary over interesting developments in learning vector classification systems. Of course, such a survey can neither be complete nor an in-depth analysis. This is more a starting point for further reading for interested researcher and operators in practice. It does not replace own experiences but it may help to find suggestions for specific tasks.

Acknowledgement

The authors are grateful to long year cooperations and friendships to many researcher, which provided substantial results mentioned in this review paper. We thank (in alphabetical order) Michael Biehl, Kerstin Bunte, Barbara Hammer, Sven Haase, Frank-Michael Schleif, Petra Schneider, Udo Seiffert and Marc Strickert for many inspiring, interesting and blitheful discussions while having coffee, wine and whiskey as well as delicious dinners together as legal and inspiring doping for ongoing exciting research.

References

[1] F. Aiolli and A. Sperduti. A re-weighting strategy for improving margins. *Artifiical Intelligence*, 137:197–216, 2002.

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

[3] A. Backhaus and U. Seiffert. Classification in high-dimensional spectral data: Accuracy vs. interpretability vs. model size. *Neurocomputing*, page in press, 2014.

[4] Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009.

[5] B.Hammer and A.Hasenfuss. Relational neural gas. *Künstliche Intelligenz*, pages 190–204, 2007.

[6] M. Biehl. Admire LVQ: Adaptive distance measures in relevance Learning Vector Quantization. *KI - Künstliche Intelligenz*, 26:391–395, 2012.

[7] M. Biehl, K. Bunte, and P. Schneider. Analysis of flow cytometry data by matrix relevance learning vector quantization. *PLoS ONE*, 8(3):e59401, 2013.

[8] M. Biehl, A. Ghosh, and B. Hammer. Dynamics and generalization ability of LVQ algorithms. *Journal of Machine Learning Research*, 8:323–360, 2007.

[9] M. Biehl, B. Hammer, and T. Villmann. Distance measures for prototype based classification. In N. Petkov, editor, *Proceedings of the International Workshop on Brain-Inspired Computing 2013 (Cetraro/Italy)*, page in press. Springer, 2014.

[10] M. Biehl, M. Kaden, and T. Villmann. Statistical quality measures and

ROC-optimization by learning vector quantization classifiers. In H. Kestler, M. Schmid, H. Binder, and B. Bischl, editors, *Proceedings of the 46th Workshop on Statistical Computing (Ulm/Reisensburg 2014)*, number 2014-xxx in Ulmer Informatik-Berichte, page accepted. University Ulm, Germany, 2014.

[11] M. Biehl, P. Schneider, D. Smith, H. Stiekema, A. Taylor, B. Hughes, C. Shackleton, P. Stewart, and W. Arlt. Matrix relevance LVQ in steroid metabolomics based classification of adrenal tumors. In M. Verleysen, editor, *Proc. of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2012)*, pages 423–428, Louvain-La-Neuve, Belgium, 2012. i6doc.com.

[12] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[13] T. Bojer, B. Hammer, D. Schunk, and T. von Toschanowitz K. Relevance determination in learning vector quantization. In *9th European Symposium on Artificial Neural Networks. ESANN'2001. Proceedings. D-Facto, Evere, Belgium*, pages 271–6, 2001.

[14] C. Bouveyron, S. Girard, and C. Schmid. High-dimensional data clustering. *Computational Statistics and Data Analysis*, 57(1):502–519, 2007.

[15] A. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1149–1155, 1997.

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

[17] T. Calinski and J. Harabacz. A dendrite method for cluster analysis. *Communications in Statistics*, 3:1–27, 1974.

[18] K. Chidanananda and G. Krishna. The condensed nearest neighbor rule using the concept of mutual nearest neighborhood. *IEEE Transactions on Information Theory*, 25:488–490, 1979.

[19] C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transaction on Information Theory*, 16(1):41–46, 1970.

[20] A. Cichocki, R. Zdunek, A. Phan, and S.-I. Amari. *Nonnegative Matrix and Tensor Factorizations*. Wiley, Chichester, 2009.

[21] R. Cilibrasi and P. Vitányi. Clustering by compression. *IEEE Transactions on Information Theory*, 51(4):1523–1545, 2005.

[22] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.

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

[24] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[25] J. Davis and M. Goadrich. The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 233–240, New York, NY, USA, 2006. ACM.

[26] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. Wiley, New

York, 1973.

[27] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861–874, 2006.

[28] L. Fischer, D. Nebel, T. Villmann, B. Hammer, and H. Wersing. Rejection strategies for learning vector quantization Ũ a comparison of probabilistic and deterministic approaches. In T. Villmann, F.-M. Schleif, M. Kaden, and M. Lange, editors, *Advances in Self-Organizing Maps: 10th International Workshop WSOM 2014 Mittweida*, Advances in Intelligent Systems and Computing, page accepted, Berlin, 2014. Springer.

[29] R. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[30] T. Geweniger, P. Schneider, F.-M. Schleif, M. Biehl, and T. Villmann. Extending RSLVQ to handle data points with uncertain class assignments. *Machine Learning Reports*, 3(MLR-02-2009):1–17, 2009. ISSN:1865-3960, http://www.uni-leipzig.de/˜compint/mlr/mlr_02_2009.pdf.

[31] T. Geweniger and T. Villmann. Extending FSNPC to handle data points with fuzzy class assignments. In M. Verleysen, editor, *Proc. of European Symposium on Artificial Neural Networks (ESANN'2010)*, pages 399–404, Evere, Belgium, 2010. d-side publications.

[32] T. Geweniger, D. Zühlke, B. Hammer, and T. Villmann. Median fuzzy c-means for clustering dissimilarity data. *Neurocomputing*, 73(7–9):1109–1116, 2010.

[33] Z. Gu, M. Shao, L. Li, and Y. Fu. Discriminative metric: Schatten norms vs. vector norm. In *Proc. of The 21st International Conference on Pattern Recognition (ICPR 2012)*, pages 1213–1216, 2012.

[34] B. Hammer, M. Strickert, and T. Villmann. Relevance LVQ versus SVM. In L. Rutkowski, J. Siekmann, R. Tadeusiewicz, and L. Zadeh, editors, *Artificial Intelligence and Soft Computing (ICAISC 2004)*, Lecture Notes in Artificial Intelligence 3070, pages 592–597. Springer Verlag, Berlin-Heidelberg, 2004.

[35] B. Hammer, M. Strickert, and T. Villmann. On the generalization ability of GRLVQ networks. *Neural Processing Letters*, 21(2):109–120, 2005.

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

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

[38] J. Hanley and B. McNeil. The meaning and use of the area under a receiver operating characteristic. *Radiology*, 143:29–36, 1982.

[39] P. Hart. The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, 14:515–516, 1968.

[40] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Verlag, Heidelberg-Berlin, 2001.

[41] S. Haykin. *Neural Networks - A Comprehensive Foundation*. IEEE Press, New York, 1994.

[42] R. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2013.

[43] J. Huang and C. X. Ling. Using AUC and accuracy in evaluating learning

algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3):299–310, 2005.

[44] M. Kaden, W. Hermann, and T. Villmann. Optimization of general statistical accuracy measures for classification based on learning vector quantization. In M. Verleysen, editor, *Proc. of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2014)*, pages 47–52, Louvain-La-Neuve, Belgium, 2014. i6doc.com.

[45] M. Kaden and T. Villmann. A framework for optimization of statistical classification measures based on generalized learning vector quantization. *Machine Learning Reports*, 7(MLR-02-2013):69–76, 2013. ISSN:1865-3960, http://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_02_2013.pdf.

[46] M. Kaden and T. Villmann. Attention based classification learning in GLVQ and asymmetric classification error assessment. In T. Villmann, F.-M. Schleif, M. Kaden, and M. Lange, editors, *Advances in Self-Organizing Maps: 10th International Workshop WSOM 2014 Mittweida*, Advances in Intelligent Systems and Computing, page accepted, Berlin, 2014. Springer.

[47] M. Kästner, D. Nebel, M. Riedel, M. Biehl, and T. Villmann. Differentiable kernels in generalized matrix learning vector quantization. In *Proc. of the Internacional Conference of Machine Learning Applications (ICMLA'12)*, pages 1–6. IEEE Computer Society Press, 2012.

[48] M. Kästner, M. Riedel, M. Strickert, W. Hermann, and T. Villmann. Border-sensitive learning in kernelized learning vector quantization. In I. Rojas, G. Joya, and J. Cabestany, editors, *Proc. of the 12th International Workshop on Artificial Neural Networks (IWANN)*, volume 7902 of *LNCS*, pages 357–366, Berlin, 2013. Springer.

[49] M. Kästner, M. Strickert, D. Labudde, M. Lange, S. Haase, and T. Villmann. Utilization of correlation measures in vector quantization for analysis of gene expression data - a review of recent developments. *Machine Learning Reports*, 6(MLR-04-2012):5–22, 2012. ISSN:1865-3960, http://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_04_2012.pdf.

[50] S. Keerthi, O. Chapelle, and D. DeCoste. Building support vector machines with reduced classifier complexity. *Journal of Machine Learning Research*, 7:1493–1515, 2006.

[51] T. Kohonen. Automatic formation of topological maps of patterns in a self-organizing system. In E. Oja and O. Simula, editors, *Proc. 2SCIA, Scand. Conf. on Image Analysis*, pages 214–220, Helsinki, Finland, 1981. Suomen Hahmontunnistustutkimuksen Seura r. y.

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

[53] T. Kohonen, J. Kangas, J. Laaksonen, and K. Torkkola. LVQ_PAK: A program package for the correct application of Learning Vector Quantization algorithms. In *Proc. IJCNN'92, International Joint Conference on Neural Networks*, volume I, pages 725–730, Piscataway, NJ, 1992. IEEE Service Center.

[54] M. Lange. Partielle Korrelationen und Partial Mutual Information zur Analyse von fMRT-Zeitreihen. Master's thesis, University of Applied Sciences Mittweida,

Mittweida, Saxony, Germany, 2012.

[55] M. Lange, M. Biehl, and T. Villmann. Non-Euclidean principal component analysis by Hebbian learning. *Neurocomputing*, page in press, 2014.

[56] M. Lange, D. Nebel, and T. Villmann. Non-Euclidean principal component analysis for matrices by Hebbian learning. In L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. Zadeh, and J. Zurada, editors, *Artificial Intelligence and Soft Computing - Proc. the International Conference ICAISC, Zakopane*, volume 1 of *LNAI 8467*, pages 77–88, Berlin Heidelberg, 2014. Springer.

[57] M. Lange and T. Villmann. Derivatives of l_p-norms and their approximations. *Machine Learning Reports*, 7(MLR-04-2013):43–59, 2013. ISSN:1865-3960, http://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_04_2013.pdf.

[58] M. Lange, D. Zühlke, O. Holz, and T. Villmann. Applications of l_p-norms and their smooth approximations for gradient based learning vector quantization. In M. Verleysen, editor, *Proc. of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2014)*, pages 271–276, Louvain-La-Neuve, Belgium, 2014. i6doc.com.

[59] T. Martinetz and K. Schulten. Topology representing networks. *Neural Networks*, 7(2), 1994.

[60] 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.

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

[62] M.Strickert, B. Labitzke, A. Kolb, and T. Villmann. Multispectral image characterization by partial generalized covariance. In M. Verleysen, editor, *Proc. of European Symposium on Artificial Neural Networks (ESANN'2011)*, pages 105–110, Louvain-La-Neuve, Belgium, 2011. i6doc.com.

[63] 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.

[64] D. Nebel, B. Hammer, and T. Villmann. Supervised generative models for learning dissimilarity data. In M. Verleysen, editor, *Proc. of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2014)*, pages 35–40, Louvain-La-Neuve, Belgium, 2014. i6doc.com.

[65] D. Nebel and T. Villmann. About the equivalence of robust soft learning vector quantization and soft nearest prototype classification. *Machine Learning Reports*, 7(MLR-02-2013):114–118, 2013. ISSN:1865-3960, http://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_02_2013.pdf.

[66] D. Nebel and T. Villmann. A median variant of generalized learning vector quantization. In M. Lee, A. Hirose, Z.-G. Hou, and R. Kil, editors, *Proceedings of International Conference on Neural Information Processing (ICONIP)*, volume II of *LNCS*, pages 19–26, Berlin, 2013. Springer-Verlag.

[67] N. Niang and G. Saporta. Supervised classification and AUC. In *Workshop Franco-Brésilien sur la fouille de données*, pages 32–33, 2009.

[68] D. Nova and P. Estévez. A review of learning vector quantization classifiers.

Neural Computation and Applications, 2013.
[69] J. Principe. *Information Theoretic Learning*. Springer, Heidelberg, 2010.
[70] J. C. Principe, J. F. III, and D. Xu. Information theoretic learning. In S. Haykin, editor, *Unsupervised Adaptive Filtering*. Wiley, New York, NY, 2000.
[71] A. Qin and P. Suganthan. Initialization insensitive LVQ algorithm based on cost-function adaptation. *Pattern Recognition*, 38:773 Ũ–776, 2004.
[72] 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.
[73] M. Riedel, D. Nebel, T. Villmann, and B. Hammer. Generative versus discriminative prototype based classification. In T. Villmann, F.-M. Schleif, M. Kaden, and M. Lange, editors, *Advances in Self-Organizing Maps: 10th International Workshop WSOM 2014 Mittweida*, Advances in Intelligent Systems and Computing, page accepted, Berlin, 2014. Springer.
[74] C. Rijsbergen. *Information Retrieval*. Butterworths, London, 2nd edition edition, 1979.
[75] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psych. Rev.*, 65:386–408, 1958.
[76] L. Sachs. *Angewandte Statistik*. Springer Verlag, 7-th edition, 1992.
[77] 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.
[78] R. Schatten. *A Theory of Cross-Spaces*, volume 26 of *Annals of Mathematics Studies*. Princeton University Press, 1950.
[79] F.-M. Schleif, T. Villmann, and B. Hammer. Prototype based fuzzy classification in clinical proteomics. *International Journal of Approximate Reasoning*, 47(1):4–16, 2008.
[80] 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.
[81] F.-M. Schleif, T. Villmann, M. Kostrzewa, B. Hammer, and A. Gammerman. Cancer informatics by prototype networks in mass spectrometry. *Artificial Intelligence in Medicine*, 45(2-3):215–228, 2009.
[82] F.-M. Schleif, X. Zhu, and B. Hammer. A conformal classifier for dissimilarity data. In L. I. I. Maglogiannis, H. Papadopoulos, K. Karatzas, and S. Siouta, editors, *Proceedings of AIAI 2012, Halkidiki, Greece*, volume 382 of *IFIP Advances in Information and Communication Technology*, pages 234–243, Berlin, 2012. Springer.
[83] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.
[84] 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.
[85] P. Schneider, T. Geweniger, F.-M. Schleif, M. Biehl, and T. Villmann. Multivariate class labeling in Robust Soft LVQ. In M. Verleysen, editor, *Proc.*

of European Symposium on Artificial Neural Networks (ESANN'2011), pages 17–22, Louvain-La-Neuve, Belgium, 2011. i6doc.com.

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

[87] P. Schneider, B. Hammer, and M. Biehl. Distance learning in discriminative vector quantization. *Neural Computation*, 21:2942–2969, 2009.

[88] C. Scovel, D. Hush, I. Steinwart, and J. Theiler. Radial kernels and their reproducing kernel Hilbert spaces. *Journal of Complexity*, 26:641–660, 2010.

[89] S. Seo, M. Bode, and K. Obermayer. Soft nearest prototype classification. *IEEE Transaction on Neural Networks*, 14:390–398, 2003.

[90] S. Seo and K. Obermayer. Soft learning vector quantization. *Neural Computation*, 15:1589–1604, 2003.

[91] G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

[92] R. Shaffer, S. Rose-Pehrsson, and R. A. McGill. Probabilistic neural networks for chemical sensor array pattern recognition: Comparison studies, improvements and automated outlier rejection. Technical Report NRL/FR/6110-98-9879, Naval Research Laboratory, Washington, DC, 1998.

[93] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis and Discovery*. Cambridge University Press, 2004.

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

[95] M. Strickert and K. Bunte. Soft rank neighbor embeddings. In M. Verleysen, editor, *Proc. of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2013)*, pages 77–82, Louvain-La-Neuve, Belgium, 2013. i6doc.com.

[96] M. Strickert, F.-M. Schleif, U. Seiffert, and T. Villmann. Derivatives of Pearson correlation for gradient-based analysis of biomedical data. *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial*, (37):37–44, 2008.

[97] M. Strickert, F.-M. Schleif, T. Villmann, and U. Seiffert. Unleashing pearson correlation for faithful analysis of biomedical data. In M. Biehl, B. Hammer, M. Verleysen, and T. Villmann, editors, *Similarity-based Clustering*, volume 5400 of *LNAI*, pages 70–91. Springer, Berlin, 2009.

[98] K. Torkkola. Feature extraction by non-parametric mutual information maximization. *Journal of Machine Learning Research*, 3:1415–1438, 2003.

[99] S. Vanderlooy and E. Hüllermeier. A critical analysis of variants of the AUC. *Machine Learning*, 72:247–262, 2008.

[100] V. Vapnik. *Statistical Learning Theory*. Wiley and Sons, New York, 1998.

[101] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

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

[103] T. Villmann, S. Haase, and M. Kaden. Kernelized vector quantization in gradient-descent learning. *Neurocomputing*, page in press, 2014.

[104] T. Villmann, S. Haase, and M. Kästner. Gradient based learning in vector quantization using differentiable kernels. In P. Estevez, J. Principe, and P. Zegers, editors, *Advances in Self-Organizing Maps: 9th International Workshop WSOM 2012 Santiage de Chile*, volume 198 of *Advances in Intelligent Systems and Computing*, pages 193–204, Berlin, 2013. Springer.

[105] T. Villmann, B. Hammer, F.-M. Schleif, T. Geweniger, and W. Herrmann. Fuzzy classification by fuzzy labeled neural gas. *Neural Networks*, 19:772–779, 2006.

[106] T. Villmann, B. Hammer, F.-M. Schleif, W. Hermann, and M. Cottrell. Fuzzy classification using information theoretic learning vector quantization. *Neurocomputing*, 71:3070–3076, 2008.

[107] T. Villmann, F.-M. Schleif, and B. Hammer. Prototype-based fuzzy classification with local relevance for proteomics. *Neurocomputing*, 69(16–18):2425–2428, October 2006.

[108] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer, Berlin, 2005.

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

[110] Y. Wu, K. Ianakiev, and V. Govindaraju. Improved k-nearest neighbor classification. *Pattern Recognition*, 35:2311 – 2318, 2002.

[111] D. Zühlke, T. Geweniger, U. Heimann, and T. Villmann. Fuzzy Fleiss-Kappa for comparison of fuzzy classifiers. In M. Verleysen, editor, *Proc. of the European Symposium on Artificial Neural Networks (ESANN'2009)*, pages 269–274, Evere, Belgium, 2009. d-side publications.

[112] X. Zhu, F.-M. Schleif, and B. Hammer. Semi-supervised vector quantization for proximity data. In M. Verleysen, editor, *Proc. of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2013)*, pages 89–94, Louvain-La-Neuve, Belgium, 2013. i6doc.com.

Received January, 2014