Neurocomputing 74 (2011) 1429–1435

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# Divergence-based classification in learning vector quantization

E. Mwebaze [a,b], P. Schneider [b,c], F.-M. Schleif [d], J.R. Aduwo [a], J.A. Quinn [a], S. Haase [e], T. Villmann [e], M. Biehl [b,*]

[a] Faculty of Computing & IT, Makerere University, P.O. Box 7062, Kampala, Uganda
[b] Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen P.O. Box 407, 9700AK Groningen, The Netherlands
[c] School of Clinical & Experimental Medicine, University of Birmingham, Birmingham B15 2TT, UK
[d] CITEC, University of Bielefeld, Universitätsstr. 21-23, 33615 Bielefeld, Germany
[e] Department of MPI, University of Applied Sciences, Technikumplatz 17, 09648 Mittweida, Germany

ARTICLE INFO

Available online 21 February 2011

*Keywords:*
Classification
Learning vector quantization
Prototype-based classifiers
Similarity measures
Distance measures

ABSTRACT

We discuss the use of divergences in dissimilarity-based classification. Divergences can be employed whenever vectorial data consists of non-negative, potentially normalized features. This is, for instance, the case in spectral data or histograms. In particular, we introduce and study divergence based learning vector quantization (DLVQ). We derive cost function based DLVQ schemes for the family of $\gamma$-divergences which includes the well-known Kullback–Leibler divergence and the so-called Cauchy–Schwarz divergence as special cases. The corresponding training schemes are applied to two different real world data sets. The first one, a benchmark data set (Wisconsin Breast Cancer) is available in the public domain. In the second problem, color histograms of leaf images are used to detect the presence of cassava mosaic disease in cassava plants. We compare the use of standard Euclidean distances with DLVQ for different parameter settings. We show that DLVQ can yield superior classification accuracies and Receiver Operating Characteristics.

© 2011 Elsevier B.V. All rights reserved.

## 1. Introduction

Distance-based classification schemes can be implemented efficiently in the framework of the popular learning vector quantization (LVQ). LVQ systems are flexible, easy to implement, and can be applied in multi-class problems in a straightforward fashion. Because LVQ prototypes are determined in the feature space of observed data, the resulting classifiers can be interpreted intuitively. Consequently, LVQ classifiers are widely used in a variety of areas including image processing tasks, medical applications, control of technical processes, or bioinformatics. An extensive bibliography including applications can be found in [1].

A key step in the design of any LVQ system is the choice of an appropriate distance measure. Most frequently, practical prescriptions make use of Euclidean metrics or more general Minkowski measures, without further justification. Generalized Euclidean measures and adaptive versions thereof have been introduced in the framework of relevance learning, see [2–7] for examples.

Here, we explore an alternative class of distance measures which relates to approaches based on statistics or information theory. So-called divergences, for instance the most popular Kullback–Leibler divergence, quantify the dissimilarity of probability distributions or positive measures. They can immediately be employed as distances in supervised or unsupervised vector quantization, provided the feature vectors and prototypes consist of non-negative, potentially normalized components. Note that throughout this article we use the terms *distance* and *distance measure* in their general meaning, which does not imply symmetry or other metric properties.

Information theoretic distance measures have been discussed in the context of various machine learning frameworks, previously. This includes prototype-based clustering, classification, or dimension reduction, see [8–12] for just a few recent examples. Frequently, divergences are employed to quantify the similarity of the prototype density with the observed distribution of data. Note that, here, we use divergences to quantify directly the distance between individual data points and prototype vectors. Moreover, we derive gradient-based update schemes which exploit the differentiability of the divergences.

After setting up the general framework, we present the family of so-called $\gamma$-divergences as a specific example. It is further specified by choice of a parameter $\gamma$ and includes the well-known Kullback–Leibler and the so-called Cauchy–Schwarz divergence as special cases.

We develop the corresponding divergence-based LVQ (DLVQ) schemes and apply them to two different classification problems.

* Corresponding author. Tel.: +31 50 363 3997.
*E-mail address:* m.biehl@rug.nl (M. Biehl).

0925-2312/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2010.10.016

First, the Wisconsin Breast Cancer data set from the UCI data repository [13] is revisited. The second data set relates to the identification of the cassava mosaic disease based on color histograms representing leaf images [14]. Performances are evaluated in terms of Receiver Operator Characteristics and compared with the standard LVQ scheme using Euclidean distance. The influence of the parameter $\gamma$ is studied and we show that data set dependent optimal values can be identified.

In the next section we outline how divergences can be incorporated into the general framework of LVQ and we derive the corresponding gradient-based training. In Section 3 we introduce the considered classification problems and data sets. Computer experiments are described and results are presented in Section 4, before we conclude with a summary and outlook.

## 2. Divergence-based learning vector quantization

For a particular classification task, we assume that a set of labeled example data is available:

$$\{\mathbf{x}^\mu, y^\mu\}_{\mu=1}^{P},$$

where the $\mathbf{x}^\mu \in \mathbb{R}^N$ are feature vectors and the labels $y^\mu \in \{1,2,\ldots,C\}$ specify their class membership.

In an LVQ system we denote by $W=\{(\mathbf{w}_j, c(\mathbf{w}_j)\}_{j=1}^{M}$ a set of $M$ prototype vectors $\mathbf{w}_j \in \mathbb{R}^N$ which carry labels $c(\mathbf{w}_j) \in \{1,2,\ldots,C\}$. Note that one or several prototypes can be assigned to each class. Prototype vectors are identified in feature space and serve as typical representatives of their classes.

Together with a given distance measure $d(\mathbf{x},\mathbf{w})$, they parameterize the classification scheme. Most frequently, a *Winner-Takes-All* scheme is applied: an arbitrary input $\mathbf{x}$ is assigned to the class $c(\mathbf{w}_L)$ of the closest prototype with $d(\mathbf{x},\mathbf{w}_L) \le d(\mathbf{x},\mathbf{w}_j)$ for all $j$.

The purpose of training is the computation of suitable prototype vectors based on the available example data. The ultimate goal, of course, is generalization: the successful application of the classifier to novel, unseen data. LVQ training can follow heuristic ideas as in Kohonen's original LVQ1 [15]. A variety of modifications has been suggested in the literature, aiming at better convergence or favorable generalization behavior. A prominent and appealing example is the cost function based generalized learning vector quantization (GLVQ) [16]. We will resort to the latter as an example framework in which to introduce and discuss divergence-based LVQ. We would like to point out, however, that differentiable divergences could be incorporated into a large variety of cost function based or heuristic training prescriptions.

GLVQ training is guided by the optimization of a cost function of the form

$$E(W)=\sum_{\mu=1}^{P}\Phi\left(\frac{d(\mathbf{x}^\mu,\mathbf{w}_J)-d(\mathbf{x}^\mu,\mathbf{w}_K)}{d(\mathbf{x}^\mu,\mathbf{w}_J)+d(\mathbf{x}^\mu,\mathbf{w}_K)}\right), \tag{1}$$

where $\mathbf{w}_J$ denotes the closest correct prototype with $c(\mathbf{w}_J)=y^\mu$ and $\mathbf{w}_K$ is the closest incorrect prototype ($c(\mathbf{w}_K)\neq y^\mu$). Note that the argument of $\Phi$ in Eq. (1) is restricted to the interval $[-1,+1]$. While $\Phi$ is in general a non-linear (e.g. sigmoidal) function [16], we consider here the simple case $\Phi(x)=x$.

In principle, a variety of numerical optimization procedures is available for the minimization of the cost function (1). On-line training using stochastic gradient descent is a particularly simple method which has proven useful in many practical applications. In stochastic gradient descent, a single, randomly selected example $\mathbf{x}$ is presented and the corresponding winners $\mathbf{w}_J,\mathbf{w}_K$ are updated incrementally by

$$\Delta\mathbf{w}_J=\frac{-\eta\, d_K(\mathbf{x})}{(d_J(\mathbf{x})+d_K(\mathbf{x}))^2}\frac{\partial}{\partial\mathbf{w}_J}\, d_J(\mathbf{x}),\quad \Delta\mathbf{w}_K=\frac{+\eta\, d_J(\mathbf{x})}{(d_J(\mathbf{x})+d_K(\mathbf{x}))^2}\frac{\partial}{\partial\mathbf{w}_K}\, d_K(\mathbf{x}) \tag{2}$$

where $d_L(\mathbf{x})=d(\mathbf{x},\mathbf{w}_L)$ and $\partial/\partial\mathbf{w}_L$ denotes the gradient with respect to $\mathbf{w}_L$. The learning rate $\eta$ controls the step size of the algorithm. Training is performed in so-called epochs, each of which presents all examples in the training data in a randomized order.

Practical prescriptions are obtained by inserting a specific distance $d(\mathbf{x},\mathbf{w})$ and its gradient. Meaningful dissimilarities should satisfy the conditions

$$d(\mathbf{x},\mathbf{w})\ge 0 \quad \text{for all } \mathbf{x},\ \mathbf{w} \text{ and } d(\mathbf{x},\mathbf{w})=0 \text{ for } \mathbf{w}=\mathbf{x}.$$

Note that the LVQ framework does not require that the distance measure satisfies metric properties such as triangular inequalities or symmetry. In both, training and working phase, only distances between data and prototype vectors have to be evaluated, the distances between two prototypes or two feature vectors are never used.

In the following we assume that the data consists of vectors of non-negative components $x_j\ge 0$ which are normalized to $\sum_{j=1}^{N} x_j=1$. Potential extensions to non-normalized positive data are discussed at the end of this section.

Normalized non-negative $x_j$ can be interpreted as probabilities. This interpretation may be just formal, as for instance in the first example data set (WBC) that we consider in the following section. Probabilistic interpretations appear natural in many cases, for instance, whenever the vectors $\mathbf{x}$ represent histograms or spectra. An important example for the former is the characterization of images by normalized gray value or color histograms, which we also consider in our second example data set, see Section 3. Frequently, spectral data is conveniently normalized to constant total intensity and is employed for classification in a large variety of fields including remote sensing or bioinformatics [1]. Assuming normalized non-negative data suggests, of course, the consideration of prototype vectors which satisfy the same constraints. Hence, we enforce

$$w_j\ge 0 \quad \text{and} \quad \sum_{j=1}^{N} w_j=1 \tag{3}$$

explicitly after each training step, Eq. (2).

Under the above assumptions, information theory provides a multitude of potentially useful dissimilarity measures. Different classes of divergences and their mathematical properties are detailed in [17,18], while [19] presents first examples of divergence-based LVQ.

We compare results of DLVQ with the standard choice, i.e. (squared) Euclidean distance:

$$D_{eu}(\mathbf{x},\mathbf{w})=\tfrac{1}{2}(\mathbf{x}-\mathbf{w})^2,\quad \frac{\partial D_{eu}(\mathbf{x},\mathbf{w})}{\partial w_k}=-(x_k-w_k). \tag{4}$$

Inserting the derivative with respect to $\mathbf{w}$ into the general framework leads to the familiar GLVQ algorithm which moves the updated prototypes either towards or away from the presented feature vector, depending on their class labels [16].

Our first results concern the so-called Cauchy–Schwarz (CS) divergence, as introduced in [20]:

$$D_{cs}(\mathbf{x},\mathbf{w})=\tfrac{1}{2}\log[\mathbf{x}^2\mathbf{w}^2]-\log\mathbf{x}^T\mathbf{w},\quad \frac{\partial D_{cs}(\mathbf{x},\mathbf{w})}{\partial w_k}=\frac{w_k}{\mathbf{w}^2}-\frac{x_k}{\mathbf{x}^T\mathbf{w}}. \tag{5}$$

It is particularly simple and, like the Euclidean measure, obeys the symmetry relation $d(\mathbf{x},\mathbf{w})=d(\mathbf{w},\mathbf{x})$.

It might be interesting to note that, apart from the logarithm, the CS-measure is formally identical to the Pearson Correlation [21] in the case of zero mean data. Pearson Correlation has been used in the context of LVQ, see [21]. However, it is important

to realize that the CS-divergence is applied only to non-negative and, consequently, non-zero mean data.

In the following, we extend our preliminary studies of the CS-divergence [19] and consider the more general family of $\gamma$-divergences, see [17,18] for the mathematical background:

$$D_\gamma(\mathbf{y},\mathbf{z}) = \frac{1}{\gamma+1}\log\left[\left(\sum_j y_j^{\gamma+1}\right)^{1/\gamma}\cdot\left(\sum_j z_j^{\gamma+1}\right)\right] - \log\left[\left(\sum_j y_j z_j^{\gamma}\right)^{1/\gamma}\right] \tag{6}$$

for $\mathbf{y},\mathbf{z}\in\mathbb{R}^N$. The precise form of this dissimilarity measure is controlled by the parameter $\gamma>0$. Note that for $\gamma=1$ one obtains the symmetric Cauchy–Schwarz divergence as a special case, while the limit $\gamma\to 0$ yields the popular Kullback–Leibler divergence.

In general, for $\gamma\neq 1$, $d_\gamma(\mathbf{x},\mathbf{y})\neq d_\gamma(\mathbf{y},\mathbf{x})$. The asymmetry is also reflected in the derivatives with respect to the first or second argument, respectively. Consequently, the use of $d(\mathbf{x},\mathbf{w})=D_\gamma(\mathbf{x},\mathbf{w})$ yields a DLVQ scheme different from the one derived for $d(\mathbf{x},\mathbf{w})=D_\gamma(\mathbf{w},\mathbf{x})$.

In the following we consider only training processes which make consistent use of either $D_\gamma(\mathbf{w},\mathbf{x})$ or $D_\gamma(\mathbf{x},\mathbf{w})$, respectively. One and the same measure is employed throughout the entire training process and when evaluating the classification performance. Accordingly, one of the following derivatives has to be inserted into Eq. (2) to yield the actual DLVQ prescription [17,18]:

$$\frac{\partial D_\gamma(\mathbf{w},\mathbf{x})}{\partial w_j} = \frac{1}{\gamma}\frac{w_j^\gamma}{\sum_k w_k^{\gamma+1}} - \frac{1}{\gamma}\frac{x_j^\gamma}{\sum_k w_k x_k^\gamma} \tag{7}$$

$$\frac{\partial D_\gamma(\mathbf{x},\mathbf{w})}{\partial w_j} = \frac{w_j^\gamma}{\sum_k w_k^{\gamma+1}} - \frac{x_j w_j^{\gamma-1}}{\sum_k x_k w_k^\gamma}. \tag{8}$$

Note that $\gamma$-divergences with $\gamma>0$ are invariant under rescaling of the arguments: $D_\gamma(\lambda\mathbf{y},\mu\mathbf{z})=D_\gamma(\mathbf{y},\mathbf{z})$ for $\lambda,\mu>0$ [17,18]. Hence, the normalization of the feature vectors, $\sum_j x_j=1$, is not required in the formalism and has no effect on the results presented here. This invariance does not hold for more general dissimilarities, as discussed in [17].

## 3. Data sets and classification problems

Two different real world data sets serve as a testbed for the suggested DLVQ algorithms.

### 3.1. Wisconsin Breast Cancer (WBC) data

We first apply DLVQ to a popular benchmark problem: The Wisconsin Breast Cancer (original) data set (WBC) from the UCI data repository [13]. Disregarding 16 vectors which contain missing values, the WBC set provides 683 examples in 9 dimensions. The data contains labels corresponding to *malignant* (239 examples) and *benign* samples (444 examples). Single features correspond to different score values between 1 and 10, see [13] for their definition. This does not imply a natural interpretation of feature vectors as histograms or probabilities. However, the application of the divergence-based formalism is possible and it is justified to the same degree as the more popular choice of Euclidean distances. For a more detailed description of this data set and further references we refer the reader to [13] and [22]. We apply a normalization such that $\sum_j x_j=1$ for the following analysis.

### 3.2. Cassava mosaic disease (CMD) data

The second data set corresponds to features extracted from leaf images of cassava plants as provided by the Namulonge National Crops Resources Research Institute, Uganda. Sample images represent 92 healthy plants and 101 plants infected with the cassava mosaic disease. For example images and further details of the image acquisition, see [14].

Standard processing techniques were employed to remove background and clutter and in order to obtain a set of characteristic features from the leaf images. When aiming at optimal classification performance, various sets of features may be taken into account [14]. Here we limit the analysis to the aspect of discolorization caused by the disease. For the application of DLVQ we consider normalized histograms with 50 bins representing the distribution of hue values in the corresponding image. Example hue histograms can also be found in [14]. The application of DLVQ appears natural in this problem, as the information is represented by normalized histograms reflecting statistical properties of the data.

## 4. Computer experiments and results

For the following evaluation and comparison of algorithms, we split the available data randomly into training (90% of the data) and test set (10%). If not stated otherwise, all results reported in the following were obtained as averages over 25 randomized splits. In both cases, we consider the simplest possible LVQ system with one prototype per class, only. Their initial positions are obtained as the mean of 50% randomly selected examples from each class in the respective training set.

The effect of the learning rate on performance depends on properties of the data set and on the distance measure in use. For simplicity, training is performed at constant learning rates in this first study. In order to facilitate a fair comparison, we determined a close to optimal learning rate from preliminary runs with respect to the achieved accuracies after a fixed number of training epochs. For each algorithm we selected the best learning rate from the set $\{10^{-3},10^{-4},\ldots,10^{-12}\}$. Results presented in the following are obtained after 200 training epochs for the WBC data and after 1500 epochs with CMD data. The learning rates employed for the WBC data set were $\eta=10^{-4}$ when using Euclidean distances and $\eta=10^{-6}$ for $\gamma$-divergences. In the CMD data set, learning rates of $\eta=10^{-5}$ (Euclidean measure) and $\eta=10^{-6}$ ($\gamma$-divergences) have been used.

After training we determine training and test set accuracies of the classifiers and we report the average values obtained over the validation runs. When comparing classifiers, it is important to take into account that greater overall test accuracies do not necessarily indicate *better* performance. The Winner-Takes-All LVQ classifier represents just one working point, i.e. one combination of class 1 error and class 2 error. In particular, for unbalanced data sets a more detailed evaluation of the classification performance is instrumental.

In order to obtain further insight, we introduce a bias $\theta$ to the classification rule after training: an input vector $\mathbf{x}$ is assigned to class 1 if

$$d(\mathbf{x},\mathbf{w}_1) < d(\mathbf{x},\mathbf{w}_2)+\theta, \tag{9}$$

where $\mathbf{w}_i$ is the closest prototype representing class $i$. The bias of the resulting classification towards one of the classes depends on the sign and magnitude of the threshold. By varying $\theta$, the full receiver operating characteristics (ROC) of the classifier can be obtained [23,24]. Results presented in Fig. 1 display a threshold-average over the validation runs [24]. In the ROC, *false positive rates* correspond to the fraction of truly benign cases (WBC data) or healthy plants (CMD) which are misclassified. Correspondingly, the *true positive rate* gives the rate at which truly malignant (WBC) or diseased plants (cassava) are correctly classified. As an

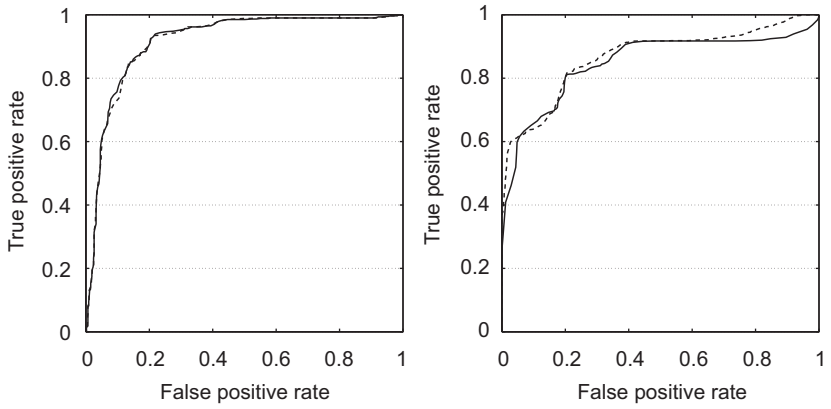



**Fig. 1.** ROC curves for the WBC data set (left panel) and CMD data set (right panel). For the WBC data, results are shown on average over 100 randomized training set compositions, whereas for the CMD data we have performed 200 randomized runs. In both cases, the ROC curves were threshold-averaged [24]. Results are displayed for the GLVQ variants based on Euclidean distances (solid lines) and Cauchy–Schwarz divergence (dashed lines).

**Table 1**
Numerical results for WBC and CMD data sets: mean accuracies in the unbiased LVQ classifier and AUC with respect to training and test sets, respectively. Numbers in parentheses give the standard deviation as observed over the validation runs.

| | Training acc. | Test acc. | AUC (training) | AUC (test) |
|---|---|---|---|---|
| **WBC** | | | | |
| $D_{eu}(\mathbf{x},\mathbf{w})$ | 0.850 (0.040) | 0.845 (0.041) | 0.924 (0.004) | 0.918 (0.004) |
| $D_{cs}(\mathbf{x},\mathbf{w})$ | 0.864 (0.003) | 0.853 (0.007) | 0.923 (0.005) | 0.916 (0.005) |
| **CMD** | | | | |
| $D_{eu}(\mathbf{x},\mathbf{w})$ | 0.790 (0.005) | 0.782 (0.007) | 0.856 (0.006) | 0.848 (0.007) |
| $D_{cs}(\mathbf{x},\mathbf{w})$ | 0.807 (0.0002) | 0.805 (0.0004) | 0.872 (0.003) | 0.867 (0.003) |

important and frequently employed measure of performance we also determine the corresponding area under curve (AUC) with respect to training set and test set performance.

### 4.1. Euclidean distance and Cauchy–Schwarz divergence

We first compare the two symmetric distance measures discussed here: standard Euclidean metrics and the Cauchy–Schwarz divergence. Fig. 1 displays the ROC with respect to test set performances in the WBC benchmark problem (left panel) and the CMD data set (right panel).

Table 1 summarize numerical findings in terms of the observed training and test accuracies for the unbiased LVQ classifier with $\theta = 0$ and the AUC of the receiver operator characteristics.

Note that for the WBC (original) data set, higher accuracies have been reported in the literature, see [13] for references. Here, we consider only the simplest LVQ setting in order to compare the use of Euclidean and divergence-based distance measures. The optimization of the DLVQ performance and the comparison with competing methods will be addressed in forthcoming studies.

In the WBC problem we do not observe drastic performance differences between the considered distance measures. The Cauchy–Schwarz-based DLVQ scheme does outperform standard Euclidean LVQ in the CMD data set as signaled by the greater AUC values. This reflects the fact that the CMD data set contains genuine histograms, for which the use of DLVQ seems natural.

### 4.2. The family of $\gamma$-divergences

The precise form of the $\gamma$-divergence is specified by the parameter $\gamma > 0$ in Eq. (6). In addition, we can select the measure $D_\gamma(\mathbf{x},\mathbf{w})$ or $D_\gamma(\mathbf{w},\mathbf{x})$, respectively. We display the mean test set accuracies as well as the AUC of the ROC as functions of $\gamma$ for both variants of the $\gamma$-divergence in Fig. 2 (WBC data) and in Fig. 3 (CMD).

In both data sets we do observe a dependence of the AUC performance on the value of $\gamma$ with a more or less pronounced optimum in a particular choice of the parameter. This is not necessarily paralleled by a maximum of the corresponding test set accuracy for unbiased LVQ classification as the latter represents only one particular working point of the ROC.

Note that in the range of values $\gamma$ displayed in Fig. 3 (right panel), corresponding to the use of $D_\gamma(\mathbf{w},\mathbf{x})$, the AUC appears to saturate for large values of the parameter. Additional experiments, however, show that performance decreases weakly when $\gamma$ is increased further.

For both data sets, the influence of $\gamma$ appears to be stronger and the best achievable AUC is slightly larger in the DLVQ variant using $D_\gamma(\mathbf{x},\mathbf{w})$. Table 2 summarizes numerical results in terms of the best observed test set AUC and the corresponding values of $\gamma$ as found for WBC and CMD data in both variants of the $\gamma$-divergence.

## 5. Summary and conclusion

We have presented DLVQ as a novel framework for distance-based classification. The use of divergences as distance measures is, in principle, possible for all data sets that contain non-negative feature values. It appears particularly suitable for the classification of histograms, spectra, or similar data structures for which divergences have been designed, originally.

As a specific example for this versatile framework we have considered the family of $\gamma$-divergences which contains the so-called Cauchy–Schwarz divergence as a special case and approaches the well-known Kullback–Leibler divergence in the limit $\gamma \to 0$. We would like to point out that a large variety of differentiable measures could be employed analogously, an overview of suitable divergences is given in [17,18].

The aim of this work is to demonstrate the potential usefulness of the approach. To this end, we considered two example data sets. The Wisconsin Breast Cancer (original) data is available from the UCI Machine Learning Repository [13] and serves as a popular benchmark problem for two-class classification. The second data set comprises histograms which represent leaf images for the purpose of the detection of the cassava mosaic disease [14].

In case of the WBC data we observe little differences in performance quality when comparing standard Euclidean metrics based LVQ with DLVQ employing the Cauchy–Schwarz

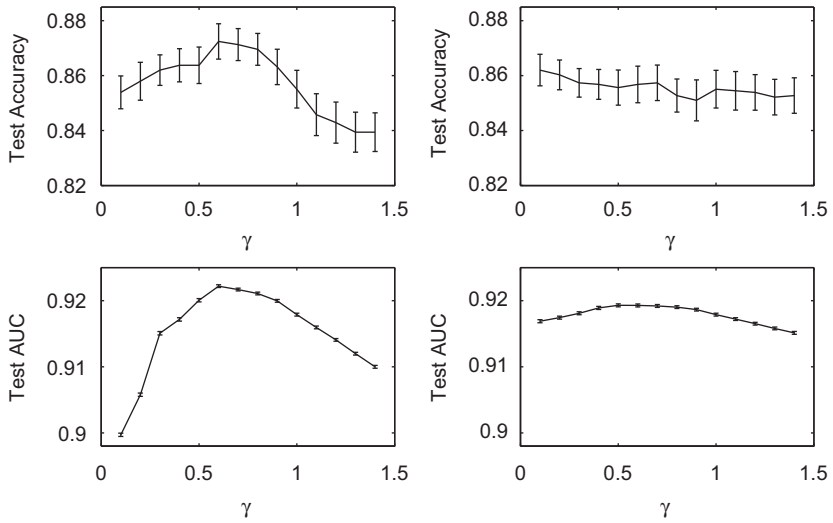



**Fig. 2.** Overall test set accuracies for the unbiased LVQ system with $\theta = 0$ (upper panels) and AUC of the ROC (lower panels) as a function of $\gamma$ for the WBC data set. The left panel displays results for the distance $D_\gamma(\mathbf{x},\mathbf{w})$, while the right panel corresponds to the use of $D_\gamma(\mathbf{w},\mathbf{x})$. Error bars mark the observed standard errors of mean.

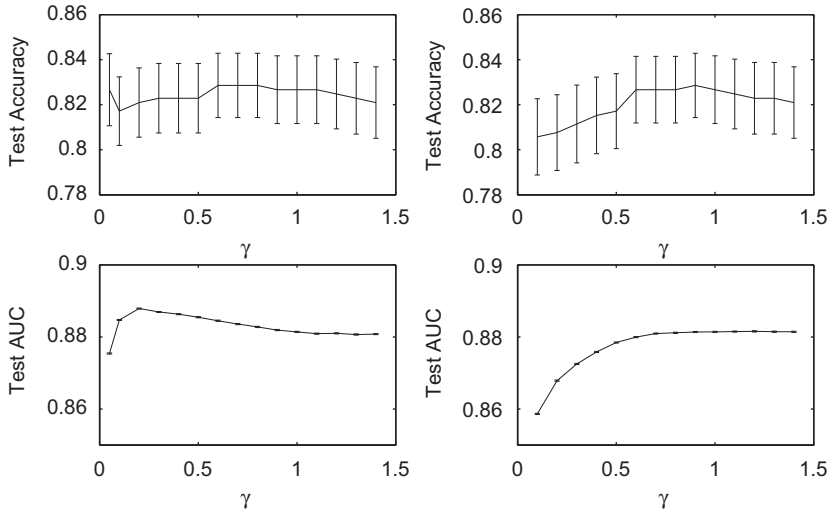



**Fig. 3.** Same as Fig. 2, but here for the CMD data set. The left panel corresponds to the use of $D_\gamma(\mathbf{w},\mathbf{x})$, while results displayed in the right panel were obtained for $D_\gamma(\mathbf{w},\mathbf{x})$.

**Table 2**
Best performance in terms of the mean test set AUC and corresponding value of $\gamma$ for the WBC and CMD data sets. Values in parenthesis correspond to the observed standard deviations.

| | $\gamma$ | AUC (test) |
|---|---|---|
| **WBC** | | |
| $D_\gamma(\mathbf{x},\mathbf{w})$ | 0.6 | 0.922 (0.004) |
| $D_\gamma(\mathbf{w},\mathbf{x})$ | 0.5 | 0.919 (0.005) |
| **CMD** | | |
| $D_\gamma(\mathbf{x},\mathbf{w})$ | 0.2 | 0.888 (0.003) |
| $D_\gamma(\mathbf{w},\mathbf{x})$ | 1.2 | 0.882 (0.004) |

divergence. When using the more general $\gamma$-divergences, a weak dependency on $\gamma$ is found which seems to allow for improving the performance slightly by choosing the parameter appropriately.

In contrast to the WBC, the CMD data set consists of genuine histogram data and the use of divergences appears more natural. In fact, we find improvement over the standard Euclidean measure already for the symmetric Cauchy–Schwarz divergence. Further improvement can be achieved by choosing an appropriate value of $\gamma$ in both variants of the non-symmetric distance measure. The dependence on $\gamma$ and its optimal choice is found to be data set specific.

The application of DLVQ appears most promising for problems that involve data with a natural interpretation as probabilities or positive measures. In forthcoming projects we will address more such data sets. Potential applications include image classification based on histograms, supervised learning tasks in a medical context, or the analysis of spectral data as in bioinformatics or remote sensing.

Besides the more extensive study of practical applications, future research will also address several theoretical and conceptual issues. The use of divergences is not restricted to the GLVQ formulation we have discussed here, it is possible to introduce DLVQ in a much broader context of heuristic or cost function based LVQ algorithms. Within several families of divergences it appears feasible to employ hyperparameter learning in order to determine, for instance, the optimal $\gamma$ directly in the training process, see [25] for a similar problem in the context of Robust Soft LVQ [26]. The use of asymmetric distance measures raises interesting questions concerning the interpretability of the LVQ prototypes, we will address this issue in forthcoming projects, as well.

Finally, the incorporation of relevance learning [2–6] into the DLVQ framework is possible for measures that are invariant under rescaling of the data, such as the $\gamma$-divergences investigated here. Relevance learning in DLVQ bears the promise to yield very powerful LVQ training schemes.

## Acknowledgments

Frank-Michael Schleif was supported by the Federal Ministry of Education and Research, Germany, FZ: 0313833, and the German Research Foundation (DFG) in the project *Relevance Learning for Temporal Neural Maps* under project code HA2719/4-1. Ernest Mwebaze was supported by The Netherlands Organization for International Cooperation in Higher Education (NUFFIC), Project NPT-UGA-238: *Strengthening ICT Training and Research Capacity in Uganda.*

## References

[1] Neural Networks Research Centre. Bibliography on the self-organizing map (SOM) and learning vector quantization (LVQ). Helsinki University of Technology, available at: 〈http://liinwww.ira.uka.de/bibliography/Neural/SOM.LVQ.html〉, 2002.
[2] T. Bojer, B. Hammer, D. Schunk, K. Tluk von Toschanowitz, Relevance determination in learning vector quantization, in: M. Verleysen (Ed.), Proceedings of the European Symposium on Art Neural Networks (ESANN), d-side, 2001, pp. 271–276.
[3] B. Hammer, T. Villmann, Generalized relevance learning vector quantization, Neural Networks 15 (8–9) (2002) 1059–1068.
[4] P. Schneider, M. Biehl, B. Hammer, Adaptive relevance matrices in learning vector quantization, Neural Computation 21 (2009) 3532–3561.
[5] P. Schneider, M. Biehl, B. Hammer, Distance learning in discriminative vector quantization, Neural Computation 21 (2009) 2942–2969.
[6] P. Schneider, K. Bunte, H. Stiekema, B. Hammer, T. Villmann, M. Biehl, Regularization in matrix relevance learning, IEEE Transactions on Neural Networks 21 (5) (2010) 831–840.
[7] K. Bunte, B. Hammer, A. Wismüller, M. Biehl, Adaptive local dissimilarity measures for discriminative dimension reduction of labeled data, Neurocomputing 73 (7–9) (2010) 1074–1092.
[8] E. Jang, C. Fyfe, H. Ko, Bregman divergences and the self organising map, in: C. Fyfe, D. Kim, S.-Y. Lee, H. Yin (Eds.), Intelligent Data Engineering and Automated Learning IDEAL 2008, Lecture Notes in Computer Science, vol. 5323, Springer, 2008, pp. 452–458.
[9] A. Banerjee, S. Merugu, I. Dhillon, J. Ghosh, Clustering with Bregman divergences, Journal of Machine Learning Research 6 (2005) 1705–1749.
[10] T. Villmann, B. Hammer, F.-M. Schleif, W. Herrmann, M. Cottrell, Fuzzy classification using information theoretic learning vector quantization, Neurocomputing 71 (2008) 3070–3076.
[11] K. Torrkola, Feature extraction by non-parametric mutual information maximization, Journal of Machine Learning Research 3 (2003) 1415–1438.
[12] K. Bunte, B. Hammer, T. Villmann, M. Biehl, A. Wismüller, Exploratory observation machine (XOM) with Kullback–Leibler divergence for dimensionality reduction and visualization, in: M. Verleysen (Ed.), European Symposium on Artificial Neural Networks (ESANN 2010), d-side publishing, 2010, pp. 87–92.
[13] D.J. Newman, S. Hettich, C.L. Blake, C.J. Merz, UCI repository of machine learning databases. Department of Information and Computer Science, University of California, Irvine, CA, available at: 〈http://www.ics.uci.edu/mlearn/MLRepository.html〉, 1998.
[14] J.R. Aduwo, E. Mwebaze, J.A. Quinn, Automated vision-based diagnosis of cassava mosaic disease, in: P. Perner (Ed.), Industrial Conference on Data Mining-Workshops, IBaI Publishing, 2010, pp. 114–122.
[15] T. Kohonen, Self-Organizing Maps, second edition, Springer, Berlin, 1997.
[16] A.S. Sato, K. Yamada, Generalized learning vector quantization, Advances in Neural Information Processing Systems, vol. 8, 1996, pp. 423–429.
[17] T. Villmann, S. Haase, Mathematical aspects of divergence based vector quantization using frechet-derivatives, Technical Report MLR-03-2009, University of Leipzig/Germany, 2009. ISSN:1865-3960 〈http://www.uni-leipzig.de/~compint/〉.
[18] T. Villmann, S. Haase, F.-M. Schleif, B. Hammer, M. Biehl, The mathematics of divergence based online learning in Vector Quantization, Proceedings of the Fourth International Workshop on Artificial Neural Networks in Pattern Recognition (ANNPR 2010), Springer Lecture Notes in Artificial Intelligence, LNAI, vol. 5998, Springer, 2010, pp. 108–119.
[19] E. Mwebaze, P. Schneider, F.-M. Schleif, S. Haase, T. Villmann, M. Biehl, Divergence based learning vector quantization, in: M. Verleysen (Ed.), Proceedings of the 18th European Symposium on Artificial Neural Networks (ESANN), d-side publishing, 2010, pp. 247–252.
[20] J.C. Principe, J.F. III, D. Xu, Information theoretic learning, in: S. Haykin (Ed.), Unsupervised Adaptive Filtering, Wiley, , 2000.
[21] M. Strickert, U. Seiffert, N. Streenivasulu, W. Weschke, T. Villmann, B. Hammer, Generalized relevance LVQ (GRLVQ) with correlation measures for gene expression analysis, Neurocomputing 69 (7–9) (2006) 651–659.
[22] K.P. Bennett, O.L. Mangasarian, Robust linear programming discrimination of two linearly inseparable sets, Optimization Methods and Software 1 (1992) 23–34.
[23] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, seconnd Edition, Wiley-Interscience, 2000.
[24] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letter 27 (2006) 861–874.
[25] P. Schneider, M. Biehl, B. Hammer, Hyperparameter learning in probabilistic prototype-based models, Neurocomputing 73 (7–9) (2010) 1117–1124.
[26] S. Seo, K. Obermayer, Soft learning vector quantization, Neural Computation 15 (7) (2003) 1589–1604.



**E. Mwebaze** received his Undergraduate and Graduate diplomas in Electrical Engineering and Computer Science, respectively, from Makerere University, Uganda, in 2003 and 2005. He is currently a Ph.D. student at Makerere University and the University of Groningen. His research interests lie in the fields of prototype based classification, specifically Divergence based Vector Quantization and causality learning.



**P. Schneider** received her Diploma in Computer Science at the University of Bielefeld, Germany, in 2005. In 2010, she received a Ph.D. in Computer Science from the University of Groningen, The Netherlands. Her Ph.D. research focused on Machine Learning techniques, especially Learning Vector Quantization, and applications in biology and medicine. Recently, she joined the School for Clinical and Experimental Medicine, University of Brimingham, UK, as postdoctoral research fellow applying Machine Learning methods in the field of Endocrinology.



**F.-M. Schleif** received his Ph.D. in Computer Science from the University of Clausthal, Germany, in 2006. He was working multiple years for Bruker Biosciences in the department of research and development. From 2006–2009 he was a research assistant in the group of computational intelligence at the University of Leipzig working on multiple bioinformatic projects. In 2010 he joined the chair of theoretical computer science at the University of Bielefeld as a Post-Doc working for multiple projects in machine learning and bioinformatics. His areas of expertize include machine learning, signal processing, data analysis and bioinformatics. Several long term research stays have taken him to UK, the Netherlands and Japan. He is co-editor of the Machine Learning Reports and reviewer for multiple journals and conferences in the field of machine learning and computational intelligence.



**J.R. Aduwo** is a Principal Consultant and Head of Information Technology Department at Uganda Management Institute and is currently a Ph.D. student at the Faculty of Computing and Informatics, Makerere University, Uganda. She holds a Master of Science Degree in Computer Science from the National University of Science and Technology—Zimbabwe, Post Graduate Diploma in Computer Science from Makerere University—Kampala, Bachelor of Science Degree from Makerere University Kampala—Uganda, one year's course on Improving Training Quality from the World Bank Institute. She is a member of the Machine Learning Group, Faculty of Computing and Informatics, Makerere University. Her research interests include Machine Learning, data mining, Knowledge Management, Management Information Systems, Business Software Applications, Computer & Internet Security and Computer Ethics. E-mail: jraduwo@umi.ac.ug/jraduwo@yahoo.com.



**J.A. Quinn** is a Senior Lecturer in Computer Science at Makerere University, Kampala, Uganda. He received a B.A. in Computer Science from the University of Cambridge (2000) and a Ph.D. from the University of Edinburgh (2007). He coordinates the Machine Learning Group at Makerere, and his research interests are in pattern recognition and computer vision particularly applied to developing world problems.



**S. Haase** received his Diploma in Applied Mathematics from University of Applied Sciences Mittweida, Germany, in 2009. Currently, he is a Ph.D. candidate in the Computational Intelligence Group of the Faculty Mathematics, Sciences and Computer Science at the University of Applied Sciences Mittweida. His research interest is in Machine Learning, especially theoretical aspects of prototype-based classification methods.



**T. Villmann** received his Ph.D. in Computer Science in 1996 and his venia legendi in the same subject in 2005, both from University of Leipzig. From 1997 to 2009 he led the computational intelligence group of the clinic for psychotherapy at Leipzig University. Since 2009 he is Professor for Technomathematics/Computational Intelligence at the University of Applied Sciences Mittweida, Germany. He is founding member of the German chapter of ENNS. His research focus includes the theory of prototype based clustering and classification, non-standard metrics, information theory and their application in pattern recognition for use in medicine, bioinformatics, satellite remote sensing and others.



**M. Biehl** received his Ph.D. in Physics from the University of Giessen, Germany, in 1992 and the venia legendi in Theoretical Physics from the University of Wuerzburg, Germany, in 1996. In 2003 he was appointed Assistant Professor in Computer Science at the University of Groningen, The Netherlands. He was promoted to Associate Professor with tenure in 2009. His main research interest is in the theory and application of Machine Learning techniques with recent emphasis on prototype based learning and adaptive metrics. He is furthermore active in the modeling and simulation of complex physical systems. Further information, re- and preprints etc. are available at http://www.cs.rug.nl/~biehl.