Prototype Based Classification Using Information Theoretic Learning

Th. Villmann[1], B. Hammer[2], F.-M. Schleif[3,4],
T. Geweniger[3,5], T. Fischer[3], and M. Cottrell[6]

[1] University Leipzig, Medical Department, Germany
[2] Clausthal University of Technology, Inst. of Computer Science, Germany
[3] University Leipzig, Inst. of Computer Science, Germany
[4] BRUKER DALTONIK Leipzig, Germany
[5] University of Applied Science Mittweida, Dep. of Computer Science, Germany
[6] University Paris I Sorbonne-Panthéon, SAMOS, France

Abstract. In this article we extend the (recently published) unsupervised information theoretic vector quantization approach based on the Cauchy–Schwarz-divergence for matching data and prototype densities to supervised learning and classification. In particular, first we generalize the unsupervised method to more general metrics instead of the Euclidean, as it was used in the original algorithm. Thereafter, we extend the model to a supervised learning method resulting in a fuzzy classification algorithm. Thereby, we allow fuzzy labels for both, data and prototypes. Finally, we transfer the idea of relevance learning for metric adaptation known from learning vector quantization to the new approach.

1 Introduction

Prototype based unsupervised vector quantization is an important task in pattern recognition. One basic advantage is the easy mapping scheme and the intuitive understanding by the concept of representative prototypes. Several methods have been established ranging from statistical approaches to neural vector quantizers [1],[2],[3]. Thereby, close connections to information theoretic learning can be drawn [4],[5],[6],[7],[8]. Based on the fundamental work of Zador, distance based vector quantization can be related to magnification in prototype base vector quantization which describes the relation between data and prototype density as a power law [9]. It can be used to design control strategies such that maximum mutual information between data and prototype density is obtained [10],[11]. However, the goal is achieved by a side effect but not directly optimized because of that distance based vector quantization methods try to minimize variants of the description error [9]. Yet, vector quantization directly optimizing information theoretic approaches become more and more important [5],[12],[8]. Two basic principles are widely used: maximization of the mutual information and minimization of the divergence, which are for uniformly distributed data

I. King et al. (Eds.): ICONIP 2006, Part II, LNCS 4233, pp. 40–49, 2006.

equivalent. Thereby, several entropies and divergence measures exist. Shannon-entropy and Kullback-Leibler-divergence were the earliest and provided the way for the other [13],[14]. One famous entropy class is the set of Rényi's α-entropies H_α, which are a generalization of the Shannon-entropy and show interesting properties [15]. In particular, the quadratic H_2-entropy is of special interest because of its convenient properties for numerical computation. J. PRINCIPE and colleagues have been shown that, based on the Cauchy-Schwarz-inequality, a divergence measure can be derived, which, together with a consistently chosen Parzen-estimator for the densities, gives a numerically well behaved approach of information optimization based prototype based vector quantization [16].

In this contribution, we extend first this approach to more general data metrices keeping the prototype based principle. In this way a broader range of application becomes possible, for instance data equipped with only available pairwise similarity measure. Further, we allow that the similarity measure may be parametrized to obtain greater flexibility. Doing so, we are able to optimize the metric and, hence, the model with respect to these parameters, too. This strategy is known in supervised learning vector quantization as *relevance learning*. The main contribution is, that we extend the original approach to a supervised learning scheme, e.g., we transfer the ideas from the unsupervised information theoretic vector quantization to an information theoretic *learning* vector quantization approach, which is a *classification* scheme. Thereby, we allow the labels of both data and prototypes to be fuzzy resulting in a prototype based fuzzy classifier, which is an improvement in comparison to standard learning vector quantization approaches, which usually provide crisp decisions and are not able to handle fuzzy labels for data.

The paper is organized as follows: First we review the approach of information theoretic vector quantization introduced by J. PRINCIPE and colleagues, but in the more general variant of arbitrary metric. Subsequently, we explain the new model for supervised fuzzy classification scheme based on the unsupervised method and show, how relevance learning can be integrated. Numerical considerations demonstrate the abilities of the new classifying system.

2 Information Theoretic Based Vector Quantization Using the Hölder-Inequality

In the following we shortly review the derivation of a numerically well behaved divergence measure. It differs in some properties from the well-known Kullback-Leibler-divergence. However, it vanishes for identical probability densities and, therefore, it can be used in density matching optimization task like prototype based vector quantization.

Shannon's definition of entropy was extended by Rényi to a more general approach. For a given density $P(\mathbf{v})$ with data points $\mathbf{v} \in \mathbb{R}^n$, the class of differential Rényi-entropies[1] is defined as [15],[17]:

[1] We will ommit the attribute '*differential*' in the following.

$$H_\alpha(\rho) = \frac{1}{1-\alpha} \log\left(\int P^\alpha(\mathbf{v})\, d\mathbf{v}\right) \tag{1}$$

$$= \frac{1}{1-\alpha} \log V_\alpha(P) \tag{2}$$

for $\alpha > 0$ and $\alpha \neq 1$. The value V_α is denoted as *information potential*. The existing limit for $\alpha \to 1$ is the Shannon entropy

$$H(\rho) = -\int P(\mathbf{v}) \log(P(\mathbf{v}))\, d\mathbf{v} \tag{3}$$

For comparison of probability density functions divergence measure are a common method. Based on Shannon entropy the Kullback-Leibler-divergence is defined as

$$KL(\rho, P) = \int \rho(\mathbf{v}) \log\left(\frac{\rho(\mathbf{v})}{P(\mathbf{v})}\right) d\mathbf{v} \tag{4}$$

for given densities ρ and P. It can be generalized according to the H_α-entropies to

$$KL_\alpha(\rho, P) = \frac{1}{\alpha - 1} \log\left(\int \rho(\mathbf{v}) \cdot \left(\frac{\rho(\mathbf{v})}{P(\mathbf{v})}\right)^{\alpha-1} d\mathbf{v}\right). \tag{5}$$

Again, in the limit $\alpha \to 1$, $KL_\alpha(\rho, P) \to KL(\rho, P)$ holds. Both divergences are non-symmetric and vanish iff $\rho \equiv P$.

For investigation in practical applications of entropy computation one has to estimate and the probabilities and to replace the integral by sample mean. Thereby the most common method for density estimation is Parzen's windowing:

$$\hat{\rho}(\mathbf{v}) = \frac{1}{M \cdot \boldsymbol{\sigma}^2} \sum_{k=1}^{M} K\left(\frac{\xi(\mathbf{v} - \mathbf{w}_k)}{\boldsymbol{\sigma}^2}\right) \tag{6}$$

whereby K is a *kernel function*. $\xi(\mathbf{v} - \mathbf{w}_k)$ is assumed to be an arbitrary difference based distance measure and $\mathbf{w}_k \in \mathbb{R}^n$ are the kernel locations. In the following we will use Gauss-kernels G. Usually, both steps, Parzen estimation and sample mean, cause numerical errors. However, the sample mean error can be eliminated: Using Rényi's quadratic entropy and the properties of kernels the information potential V_2 can be estimated by

$$V_2 = \frac{1}{M^2 \cdot \boldsymbol{\sigma}^4} \sum_{k=1}^{M} \sum_{j=1}^{M} \int G\left(\frac{\xi(\mathbf{v} - \mathbf{w}_k)}{\boldsymbol{\sigma}^2}\right) \cdot G\left(\frac{\xi(\mathbf{v} - \mathbf{w}_j)}{\boldsymbol{\sigma}^2}\right) d\mathbf{v} \tag{7}$$

$$= \frac{1}{M^2 \cdot \boldsymbol{\sigma}^4} \sum_{k=1}^{M} \sum_{j=1}^{M} G\left(\frac{\xi(\mathbf{w}_k - \mathbf{w}_j)}{2\boldsymbol{\sigma}^2}\right) \tag{8}$$

without carrying out the integration in practice.

Unfortunately this approach can not be easily transferred to the quadratic divergence measure KL_2 because it is not quadratic according to all involved

densities. Therefore, PRINCIPE suggested to use a divergence measure derived from the Cauchy-Schwarz-inequality. To do this, we first remark that the general information potential V_α in (1) defines a norm $\|\cdot\|_\alpha = (V_\alpha(\cdot))^{\frac{1}{\alpha}}$ for α-integrable functions. In particular in Hilbert-spaces the Hölder-inequality holds

$$\frac{\|\rho\|_\alpha \cdot \|P\|_{1-\alpha}}{\|\rho \cdot P\|_1} \geq 1 \tag{9}$$

with the equality iff $\rho \equiv P$ except a zero-measure set. For $\alpha = 2$ this is the Cauchy-Schwarz-inequality, which can be used for a divergence definition [8]:

$$D_{CS}(\rho, P) = \frac{1}{2}\log\left(\int \rho^2(\mathbf{v})\, d\mathbf{v} \cdot \int P^2(\mathbf{v})\, d\mathbf{v}\right) - \log\left(\int P(\mathbf{v}) \cdot \rho(\mathbf{v})\, d\mathbf{v}\right) \tag{10}$$

$$= \frac{1}{2}\log\left(V_2(\rho) \cdot V_2(P)\right) - \log Cr(P, \rho) \tag{11}$$

whereby Cr is called the *cross-information potential* and D_{CS} is denoted as *Cauchy-Schwarz-divergence*. Yet, the divergence D_{CS} does not fulfill all properties of the Kullback-Leibler-divergence KL but keeping the main issue that D_{CS} vanishes for $\rho \equiv P$ (in prob.) [18]. Now we can use the entropy estimator for $V_2(\rho)$ and $V_2(P)$ according to (8) and apply the same kernel property to the cross-information potential:

$$Cr(\rho, P) = \int P(\mathbf{v}) \cdot \rho(\mathbf{v})\, d\mathbf{v} \tag{12}$$

$$= \frac{1}{N \cdot M \cdot \boldsymbol{\sigma}^4} \sum_{k=1}^{M} \sum_{j=1}^{N} G\left(\frac{\xi(\mathbf{v}_j - \mathbf{w}_k)}{2\sigma^2}\right) \tag{13}$$

whereby, again, the integration is not to be carried out in practice and, hence, does not lead to numerical errors.

In (unsupervised) vector quantization the data density P is given (by samples), whereas the density ρ is the density of prototypes $\mathbf{w}_k$, which is subject of change. In information optimum vector quantization the adaptation should lead to minimization of D_{CS}.

3 Prototype Based Classification Using Cauchy-Schwarz Divergence

In the following we will extend the above outlined approach to the task of prototype based classification. Although many classification methods are known, prototype based classification is a very intuitive method. Most widely used methods are the learning vector quantization algorithms (LVQ) introduced by KOHONEN [2]. However, the adaptation dynamic does not follow a gradient of any cost function. Heuristically, the misclassification error is reduced. However, for overlapping classes the heuristic causes instabilities. Several modifications are known to overcome this problem [19],[20],[21].

From information theoretic learning point of view, an algorithm maximizing the mutual information using Re was introduced by Torkkola denoted as IT-LVQ [22]. However, compared to other classification approaches, this algorithm does not show convincing performance [23].

A remaining problem is that all these methods do not return fuzzy valued classification decisions as well as are not able do handle fuzzy classified data. Here we propose to use a Cauchy-Schwarz-divergence based cost function, which also can be applied to fuzzy labeled data.

Let $\mathbf{x}(\mathbf{v})$ be the fuzzy valued class label for data point $\mathbf{v} \in \mathbb{R}^n$ and $\mathbf{y}_i$ for prototypes $\mathbf{w}_i \in \mathbb{R}^n$. Assuming, N_c is the number of possible classes, the fuzzy labels are realized as $\mathbf{x}(\mathbf{v}), \mathbf{y}_i \in \mathbb{R}^{N_c}$ with components $x_k(\mathbf{v}), y_i^k \in [0,1]$ with the normalization conditions $\sum_{k=1}^{N_c} x_k(\mathbf{v}) = 1$ and $\sum_{k=1}^{N_c} y_i^k = 1$. Let $P_{\mathbf{X}}(c)$ and $\rho_{\mathbf{Y}}(c)$ be the label density of data labels $\mathbf{X}$ and prototype labels $\mathbf{Y}$ for a given class c, respectively. We define as cost function to be minimized

$$C(\mathbf{Y}, \mathbf{X}) = \sum_{c=1}^{N_c} \varpi_c \cdot 2 \cdot D_{CS}(\rho_{\mathbf{Y}}(\mathbf{v}, c), P_{\mathbf{X}}(\mathbf{v}, c)). \tag{14}$$

with given weighting factors ϖ_c determining the importance of a class. Because of all $P_{\mathbf{X}}(c)$ are determined by given data, minimization of $D_{CS}(\rho_{\mathbf{Y}}(\mathbf{v}, c), P_{\mathbf{X}}(\mathbf{v}, c))$ is equivalent to minimization of

$$\hat{C}(\mathbf{Y}, \mathbf{X}) = \sum_{c=1}^{N_c} \varpi_c \cdot \hat{C}_c(\mathbf{Y}, \mathbf{X}) \tag{15}$$

with class dependent cost functions

$$\hat{C}_c(\mathbf{Y}, \mathbf{X}) = (\log(V_2(\rho_{\mathbf{Y}}(\mathbf{v}, c))) - 2 \log Cr(\rho_{\mathbf{Y}}(\mathbf{v}, c), P_{\mathbf{X}}(\mathbf{v}, c))). \tag{16}$$

Information theoretic learning vector quantization now is taken as optimizing the prototype locations $\mathbf{w}_k$ together with their class responsibilities (labels) $\mathbf{y}^k$ according to minimization of $\hat{C}(\mathbf{Y}, \mathbf{X})$.

To do so, we assume for simplicity that the variance in each data dimension is equal σ^2, the general case is straight forward. We introduce the class (label) dependent Parzen estimates

$$\hat{P}_{\mathbf{X}}(\mathbf{v}, c) = \frac{1}{N} \sum_{i=1}^{N} x_c(\mathbf{v}_i) \cdot G\left(\frac{\xi(\mathbf{v} - \mathbf{v}_i)}{\boldsymbol{\sigma}^2}\right) \tag{17}$$

and

$$\hat{\rho}_{\mathbf{Y}}(\mathbf{v}, c) = \frac{1}{M} \sum_{i=1}^{M} y_i^c \cdot G\left(\frac{\xi(\mathbf{v} - \mathbf{w}_i)}{\boldsymbol{\sigma}^2}\right). \tag{18}$$

We further assume for the moment that all ϖ_c are fixed and equal. Then the class dependent cost functions $\hat{C}_c(\mathbf{Y}, \mathbf{X})$ can be written as

$$\hat{C}_c(\mathbf{Y},\mathbf{X}) \approx \frac{1}{2M}\sum_{i=1}^{M} y_i^c \log\left(\frac{1}{M}\sum_{j=1}^{M} y_j^c G\left(\frac{\xi(\mathbf{w}_i-\mathbf{w}_j)}{2\boldsymbol{\sigma}^2}\right)\right) \tag{19}$$

$$-\frac{1}{M}\sum_{i=1}^{M} y_i^c \log\left(\frac{1}{N}\sum_{j=1}^{N} x_c(\mathbf{v}_j)\cdot G\left(\frac{\xi(\mathbf{w}_i-\mathbf{v}_j)}{2\boldsymbol{\sigma}^2}\right)\right) \tag{20}$$

which yields the class dependent derivatives

$$\frac{\partial \hat{C}_c(\mathbf{Y},\mathbf{X})}{\partial \mathbf{w}_k} = -\frac{1}{4\sigma^2}\left[\begin{array}{c} \frac{\sum_{i=1}^{M} y_i^c y_k^c G\left(\frac{\xi(\mathbf{w}_i,\mathbf{w}_k)}{2\boldsymbol{\sigma}^2}\right)\frac{\partial\xi(\mathbf{w}_i,\mathbf{w}_k)}{\partial\mathbf{w}_k}}{\sum_{i=1}^{M}\sum_{j=1}^{M} y_i^c y_j^c G\left(\frac{\xi(\mathbf{w}_i,\mathbf{w}_j)}{2\boldsymbol{\sigma}^2}\right)} \\ -\frac{\sum_{j=1}^{N} y_k^c x_c(\mathbf{v}_j) G\left(\frac{\xi(\mathbf{v}_j,\mathbf{w}_k)}{2\boldsymbol{\sigma}^2}\right)\frac{\partial\xi(\mathbf{v}_j,\mathbf{w}_k)}{\partial\mathbf{w}_k}}{\sum_{i=1}^{M}\sum_{j=1}^{N} y_i^c x_c(\mathbf{v}_j) G\left(\frac{\xi(\mathbf{v}_j,\mathbf{w}_i)}{2\boldsymbol{\sigma}^2}\right)} \end{array}\right] \tag{21}$$

and

$$\frac{\partial \hat{C}(\mathbf{Y},\mathbf{X})}{\partial y_c^k} = \varpi_c \cdot \frac{\partial \hat{C}_c(\mathbf{Y},\mathbf{X})}{\partial y_c^k} \tag{22}$$

with

$$\frac{\partial \hat{C}_c(\mathbf{Y},\mathbf{X})}{\partial y_c^k} = \frac{\sum_{j=1}^{M} y_j^c G\left(\frac{\xi(\mathbf{w}_j,\mathbf{w}_k)}{2\boldsymbol{\sigma}^2}\right)}{\sum_{i=1}^{M}\sum_{j=1}^{M} y_i^c y_j^c G\left(\frac{\xi(\mathbf{w}_i,\mathbf{w}_j)}{2\boldsymbol{\sigma}^2}\right)} - \frac{2\sum_{j=1}^{N} x(\mathbf{v}_j)\, G\left(\frac{\xi(\mathbf{v}_j,\mathbf{w}_k)}{2\boldsymbol{\sigma}^2}\right)}{\sum_{i=1}^{M}\sum_{j=1}^{N} y_i^c x(\mathbf{v}_j)\, G\left(\frac{\xi(\mathbf{v}_j,\mathbf{w}_i)}{2\boldsymbol{\sigma}^2}\right)}. \tag{23}$$

Both gradients (21) and (23) determine the parallel stochastic gradient descent for minimization of $\hat{C}(\mathbf{Y},\mathbf{X})$ depending on the used distance measure ξ. In case of $\xi(\mathbf{v}-\mathbf{w})$ being the quadratic Euclidean distance, we simply have $\frac{\partial\xi(\mathbf{v}-\mathbf{w})}{\partial\mathbf{w}} = 2(\mathbf{v}-\mathbf{w})$.

We denote the resulting adaptation algorithm

$$\begin{aligned} \triangle\mathbf{w}_k &= -\epsilon\frac{\partial\hat{C}(\mathbf{Y},\mathbf{X})}{\partial\mathbf{w}_k} \\ \triangle y_c^k &= -\tilde{\epsilon}\frac{\partial\hat{C}(\mathbf{Y},\mathbf{X})}{\partial y_c^k} \end{aligned} \tag{24}$$

as *Learning Vector Quantization based on Cauchy-Schwarz-Divergence – LVQ-CSD*

4 Applications

In a first toy example we applied the LVQ-CSD using the quadratic Euclidean distance for ξ to classify data obtained from two two-dimensional overlapping Gaussian distribution, each of them defining a data class. The overall number of data was $N = 600$ equally splitted into test and train data. We used 10 prototypes with randomly initialized positions and fuzzy labels.

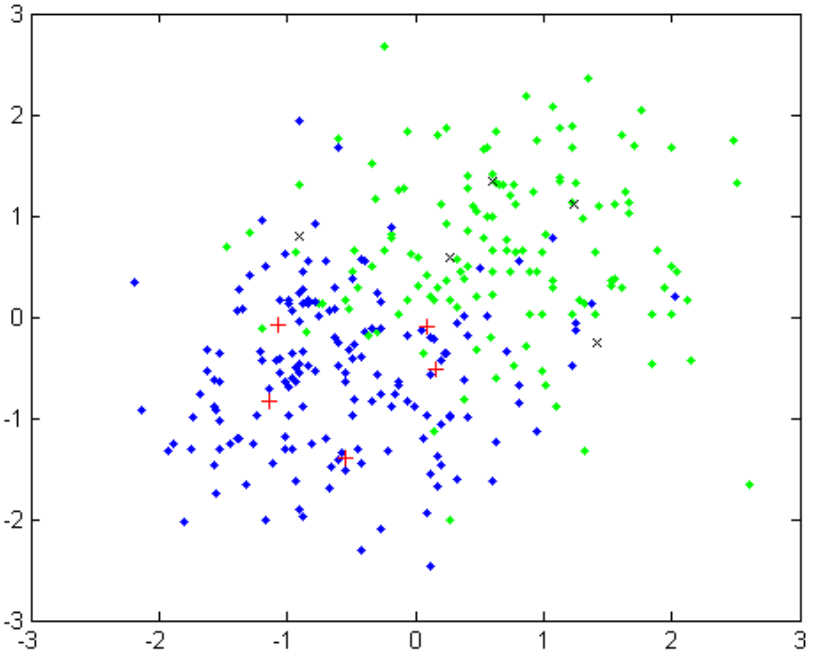

Fig. 1. Visualization of learned prototypes for LVQ-CSD in case of overlapping Gaussians, defining two classes (green, blue). The positions of prototypes are indicated by red '+' and black '×' according to their fuzzy label based majority vote for the blue and green classes, respectively.

One crucial point using Parzen estimators is the adequate choice of the kernel size $\boldsymbol{\sigma}^2$. Silverman's rule gives a rough estimation [24]. Otherwise, as pointed out in [16], $\boldsymbol{\sigma}^2$ also determines the role of cooperativeness range of prototypes in data space during adaptation, which should be larger in the beginning and smaller in the convergence phase for fine tuning. Combining both features we choose for a certain training step t

$$\boldsymbol{\sigma}(t) = \frac{3 \cdot \gamma \cdot \boldsymbol{\sigma}(0)}{1 + \delta \cdot \boldsymbol{\sigma}(0) \cdot t} \tag{25}$$

with $\gamma = 1.06 \cdot n^{-\frac{1}{5}}$ the Silverman-factor ([24]) and $\delta = 5/T$ and T being the total number of training steps. n is the data dimension and $\boldsymbol{\sigma}(0) = \boldsymbol{\sigma}$ is the original data variance.

The resulting classification accuracy (majority vote) for LVQ-CSD for the simple toy example is 93.1%, see Fig. 1. This result is comparable good to the lower accuracy obtained by standard LVQ2.1 [2], which yields 77.5%. Further, for LVQ-CSD prototypes located at overlapping border region, have balanced label vectors whereas prototypes in the center of the class regions show clear label preferences.

Table 1. Test rates for the different algorithms on the WBDC data set. For LVQ-CSD the majority vote was applied for accuracy determination.

	LVQ-CSD	LVQ2.1	GLVQ	SNG	IT-LVQ
toy sample	93.1%	77.5%	91.3%	94.9%	63.3%
PIMA	75.3%	65.3%	74.2%	78.2%	65.8%
WINE	95.5%	93.1%	98.3%	98.3%	61.9%
IONOSPHERE	69.0%	64.1%	81.4%	82.6%	56.2%

In a second mor challenging application, we investigated the behavior of the new algorithm in case of data sets from the UCI repository [25]. The data dimensions are 9, 13 and 34 for the PIMA-, the WINE- and the IONOSPHERE data, respectively. The allover number of data are 768, 178 and 351, respectively. The first and the third task are 2-class problems whereas the second one is a three-class problem. We splitted the data set for training and test randomly such that about 66% are for training.

We compare the LVQ-CSD with LVQ2.1 [2], GLVQ [20], and IT-LVQ [26] covering different principles of learning vector quantization: distance based heuristic, distance based classifier function and mutual information optimization, respectively. Because one can interpret the kernel size $\boldsymbol{\sigma}$ as a range of cooperativeness, we also added a comparison with supervised neural gas (SNG), which is an extension of GLVQ incorporating neighborhood cooperativeness [27]. The number of prototypes were chosen as 10% of train data for all algorithm, again in comparison to the earlier studies [23]. The results are depicted in Tab. 1. Except the IT-LVQ and LVQ2.1, all algorithms show comparable results with small advantages for GLVQ and, in particular SNG. LVQ-CSD shows good performance. It clearly outperforms standard LVQ2.1 and the IT-LVQ, which is based on mutual information maximization. The weak result for IONOSPHERE data set could be adressed to the well known problem arising for all Parzen estimation approaches: For high-dimensional space Parzen estimators may become insensitive because of the properties of the Euclidean norm in high-dimensional spaces: this is that according to the Euclidean distance measure most of the data lie in a thin sphere of the data space [28]. The effect could be the reason for the bad performance. Hoewever, here we have to make further investigations.

5 Conclusion and Future Work

Based on the information theoretic approach of unsupervised vector quantization by density matching using Cauchy-Schwarz-divergence, we developed a new supervised learning vector quantization algorithm, which is able to handle fuzzy labels for data as well as for prototypes. In first simulations the algorithm shows valuable results. We formulated the algorithm for general difference based distance measures $\xi(\mathbf{v} - \mathbf{w})$. However, up to now we only used the Euclidean distance. Yet, it is possible to use more complicate difference based distance measures. In particular, parametrized measures $\xi_{\boldsymbol{\lambda}}$ are of interest with parameter

vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{N_\lambda})$, $\lambda_i \geq 0$ and $\sum \lambda_i = 1$. Then the parametrization can be optimized for a given classification task, too. This method is known as *relevance learning* in learning vector quantization [29],[27]. For this purpose, simply the additional gradient descent $\frac{\partial \hat{C}(\mathbf{Y},\mathbf{X})}{\partial \lambda_j}$ has to be taken into account. Obviously, this idea can be transferred also to Cauchy-Schwarz-divergence as cost function of the unsupervised information theoretic vector quantization, which also would allow an adapted metric for improved performance. The analyze of these extensions in practical applications is subject of current research.

References

1. Simon Haykin, *Neural Networks - A Comprehensive Foundation*, IEEE Press, New York, 1994.
2. Teuvo Kohonen, *Self-Organizing Maps*, vol. 30 of *Springer Series in Information Sciences*, Springer, Berlin, Heidelberg, 1995, (Second Extended Edition 1997).
3. Erkki Oja and Jouko Lampinen, "Unsupervised learning for feature extraction", in *Computational Intelligence Imitating Life*, Jacek M. Zurada, Robert J. Marks II, and Charles J. Robinson, Eds., pp. 13–22. IEEE Press, 1994.
4. R. Brause, *Neuronale Netze*, B. G. Teubner, Stuttgart, 2nd. edition, 1995.
5. G. Deco and D. Obradovic, *An Information-Theoretic Approach to Neural Computing*, Springer, Heidelberg, New York, Berlin, 1997.
6. A. K. Jain, R. P.W. Duin, and J. Mao, "Statistical pattern recognition: A review", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 4–37, 2000.
7. J.N. Kapur, *Measures of Information and their Application*, Wiley, New Delhi, 1994.
8. J. C. Principe, J.W. Fischer III, and D. Xu, "Information theoretic learning", in *Unsupervised Adaptive Filtering*, S. Haykin, Ed. Wiley, New York, NY, 2000.
9. P. L. Zador, "Asymptotic quantization error of continuous signals and the quantization dimension", *IEEE Transaction on Information Theory*, , no. 28, pp. 149–159, 1982.
10. Marc M. Van Hulle, *Faithful Representations and Topographic Maps*, Wiley Series and Adaptive Learning Systems for Signal Processing, Communications, and Control. Wiley & Sons, New York, 2000.
11. T. Villmann and J.-C. Claussen, "Magnification control in self-organizing maps and neural gas", *Neural Computation*, vol. 18, no. 2, pp. 446–469, February 2006.
12. Marc M. Van Hulle, "Joint entropy maximization in kernel-based topographic maps", *Neural Computation*, vol. 14, no. 8, pp. 1887–1906, 2002.
13. S. Kullback and R.A. Leibler, "On information and sufficiency", *Annals of Mathematical Statistics*, vol. 22, pp. 79–86, 1951.
14. C.E. Shannon, "A mathematical theory of communication", *Bell System Technical Journal*, vol. 27, pp. 379–432, 1948.
15. A. Renyi, "On measures of entropy and information", in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. 1961, University of California Press.
16. T. Lehn-Schiler, A. Hegde, D. Erdogmus, and J.C. Principe, "Vector quantization using information theoretic concepts", *Natural Computing*, vol. 4, no. 1, pp. 39–51, 2005.

17. A. Renyi, *Probability Theory*, North-Holland Publishing Company, Amsterdam, 1970.
18. R. Jenssen, *An Information Theoretic Approach to Machine Learning*, PhD thesis, University of Troms, Department of Physics, 2005.
19. S. Seo and K. Obermayer, "Soft learning vector quantization", *Neural Computation*, vol. 15, pp. 1589–1604, 2003.
20. A. Sato and K. Yamada, "Generalized learning vector quantization", in *Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference*, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Eds., pp. 423–9. MIT Press, Cambridge, MA, USA, 1996.
21. S. Seo, M. Bode, and K. Obermayer, "Soft nearest prototype classification", *IEEE Transaction on Neural Networks*, vol. 14, pp. 390–398, 2003.
22. K. Torkkola, "Feature extraction by non-parametric mutual information maximization", *Journal of Machine Learning Research*, vol. 3, pp. 1415–1438, 2003.
23. T. Villmann, F.-M. Schleif, and B. Hammer, "Comparison of relevance learning vector quantization with other metric adaptive classification methods", *Neural Networks*, vol. 19, pp. in press, 2006.
24. B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, 1986.
25. C.L. Blake and C.J. Merz, "UCI repository of machine learning databases", Irvine, CA: University of California, Department of Information and Computer Science, available at: http://www.ics.uci.edu/ mlearn/MLRepository.html, 1998.
26. K. Torkkola and W.M. Campbell, "Mutual information in learning feature transformations", in *Proc. Of International Conference on Machine Learning ICML'2000*, Stanford, CA, 2000.
27. B. Hammer, M. Strickert, and Th. Villmann, "Supervised neural gas with general similarity measure", *Neural Processing Letters*, vol. 21, no. 1, pp. 21–44, 2005.
28. M. Verleysen and D. François, ", in *Computational Intelligence and Bioinspired Systems, Proceedings of the 8th International Work-Conference on Artificial Neural Networks 2005 (IWANN), Barcelona*, J. Cabestany, A. Prieto, and F. S. Hernández, Eds.
29. B. Hammer and Th. Villmann, "Generalized relevance learning vector quantization", *Neural Networks*, vol. 15, no. 8-9, pp. 1059–1068, 2002.