ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence
and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

Divergence based Learning Vector Quantization

E. Mwebaze[1,2], P. Schneider[2], F.-M. Schleif[3], S. Haase[4], T. Villmann[4], M. Biehl[2]
1 – Faculty of Computing & IT, Makerere Univ., P.O. Box 7062, Kampala, Uganda
2 – Johann Bernoulli Inst. for Mathematics and Computer Science, Univ. of Groningen
P. O. Box 407, 9700AK Groningen, The Netherlands
3 – Computational Intelligence Group, University of Leipzig
Semmelweisstr. 10, 04103 Leipzig, Germany
4 – Department of MPI, University of Applied Sciences,
Technikumplatz 17, 09648 Mittweida, Germany

Abstract. We suggest the use of alternative distance measures for similarity based classification in Learning Vector Quantization. Divergences can be employed whenever the data consists of non-negative normalized features, which is the case for, e.g., spectral data or histograms. As examples, we derive gradient based training algorithms in the framework of Generalized Learning Vector Quantization based on the so-called Cauchy-Schwarz divergence and a non-symmetric Renyi divergence. As a first test we apply the methods to two different biomedical data sets and compare with the use of standard Euclidean distance.

1 Introduction

Learning Vector Quantization (LVQ) provides a widely used family of algorithms for distance based classification. LVQ systems are very flexible, easy to implement, and applicable to 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.

The choice of an appropriate distance measure is crucial for the success of LVQ training and classification. Most practical prescriptions make use of Minkowski, e.g. Euclidean metrics or adaptive versions thereof as in relevance learning [1, 2, 3, 4].

Here we will consider alternatives which are applicable in the presence of non-negative, normalized feature vectors. Restricting the prototypes accordingly, divergences can be employed as dissimilarity or distance measures. We will discuss two specific examples which belong to different families of divergences.

Information theoretic distance measures have been discussed in the context of various machine learning frameworks, previously. This includes prototype based clustering and classification, see [5, 6, 7, 8] 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 employ divergences to quantify the distance between individual feature vectors and prototype vectors, both of which are interpreted as probability distributions. Moreover, we derive gradient based update schemes which exploit the differentiability of the divergences.

After detailing the framework, we discuss two examples of potentially useful divergence measures, one of which is non-symmetric. The mathematical aspects have been presented in greater detail in a recent technical report [9]. As a proof of concept and in order to obtain first insights, we employ the corresponding divergence based GLVQ schemes to two example data sets and compare with the standard scheme which is based on the Euclidean measure.

2 Introduction of divergence based GLVQ

In the following we assume that a set $\{\mathbf{x}^\mu, y^\mu\}_{\mu=1}^P$ of example data is given. Here $\mathbf{x}^\mu \in I\!\!R^N$ and the labels $y^\mu \in \{1, 2, \ldots C\}$ correspond to one of the classes. An LVQ system $W = \{(\mathbf{w}_j, c(\mathbf{w}_j)\}_{j=1}^M$ comprises a number M of N-dim. prototype vectors $\mathbf{w}_j$ which carry labels $c(\mathbf{w}_j) \in \{1, 2, \ldots C\}$.

Given a distance measure $d(\mathbf{x}, \mathbf{w})$, the LVQ classifier employs a *Winner-Takes-All* scheme: an arbitrary input $\mathbf{x}$ is assigned to the class $c(\mathbf{w}_L)$ of the closest prototype with $d(\mathbf{x}, \mathbf{w}_L) \leq d(\mathbf{x}, \mathbf{w}_j)\ \forall j$.

LVQ training can follow heuristic ideas as in Kohonen's original LVQ1 [10]. A variety of modifications has been suggested, a prominent example being the cost function based Generalized Learning Vector Quantization (GLVQ) [11]. We will employ the latter as an example framework in which we introduce and discuss divergence based LVQ. We would like to point out, however, that differentiable measures can be employed more generally in 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

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

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 Φ in Eq. (1) is restricted to the interval $[-1, +1]$. While Φ is in general a non-linear (e.g. sigmoidal) function, we consider here the simple case $\Phi(x) = x$.

In stochastic gradient descent, a randomly selected example $\mathbf{x}$ is presented and the corresponding winners $\mathbf{w}_J, \mathbf{w}_K$ are updated incrementally by

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

where $d_L(\mathbf{x}) = d(\mathbf{x}, \mathbf{w}_L)$ and ∇_L denotes the gradient with respect to $\mathbf{w}_L$. The so-called learning rate η controls the step size of the algorithm.

Practical prescriptions are obtained by inserting a specific dissimilarity $d(\mathbf{x}, \mathbf{w})$ and its gradient. Obviously, the same measure should be used in the working phase of the LVQ system for nearest prototype identification. Meaningful dissimilarities should satisfy the conditions $d(\mathbf{x}, \mathbf{w}) \geq 0$ for all possible vectors $\mathbf{x}, \mathbf{w}$ and $d(\mathbf{x}, \mathbf{w}) = 0$ for $\mathbf{w} = \mathbf{x}$. Note that in the LVQ framework it is not necessary to require metric properties such as symmetry. Both, in training and working phase only distances between data and prototype vectors have to be evaluated, distances between two prototypes or two feature vectors are never used.

In the following we assume that the data consists of feature vectors of non-negative components $x_j \geq 0$ which are normalized to $\sum_{j=1}^{N} x_j = 1$. In other words, the x_j could be interpreted as probabilities. This interpretation could be only formal but is natural in many cases, for instance if the data comprises vectors $\mathbf{x}$ which represent histograms or spectra. A prominent example for the former is the characterization of images by normalized gray value or color histograms. 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. Assuming normalized non-negative data suggests immediately the consideration of prototype vectors which satisfy the same constraints: $w_j \geq 0$ and $\sum_{j=1}^{N} w_j = 1$.

Under the above assumptions, information theory provides a multitude of potentially useful dissimilarity measures. Different classes of divergences and their potential use in prototype based training are discussed in [9]. Here, we consider a couple of specific examples and compare with a standard choice:

a) Squared Euclidean distance

$$d_{eu}(\mathbf{x},\mathbf{w}) \;=\; \frac{1}{2}(\mathbf{x}-\mathbf{w})^2, \qquad \frac{\partial d_{eu}(\mathbf{x},\mathbf{w})}{\partial\, w_k} \;=\; -(x_k - w_k) \tag{3}$$

b) Cauchy-Schwarz divergence (as introduced in [12]):

$$d_{cs}(\mathbf{x},\mathbf{w}) \;=\; \frac{1}{2}\log\left[\mathbf{x}^2\mathbf{w}^2\right] - \log \mathbf{x}^T\mathbf{w}, \quad \frac{\partial d_{cs}(\mathbf{x},\mathbf{w})}{\partial\, w_k} \;=\; \frac{w_k}{\mathbf{w}^2} - \frac{x_k}{\mathbf{x}^T\mathbf{w}} \tag{4}$$

c) Renyi divergence **($\boldsymbol{\alpha} = \mathbf{2}$)** [13]

$$d_{re}(\mathbf{x},\mathbf{w}) \;=\; \log \sum_j x_j^2 /\, w_j, \qquad \frac{\partial d_{re}(\mathbf{x},\mathbf{w})}{\partial\, w_k} \;=\; \frac{-x_k^2/w_k^2}{\sum_j x_j^2/w_j} \tag{5}$$

Please note that the general definition of Renyi divergence contains a parameter α [13]. Here we focus on the particularly convenient case $\alpha = 2$ which yields the definition (5).

While the Euclidean (3) and Cauchy-Schwarz (4) dissimilarities are symmetric, the Renyi divergence is not: $d_{re}(\mathbf{x},\mathbf{w}) \neq d_{re}(\mathbf{w},\mathbf{x})$. We have chosen $d_{re}(\mathbf{x},\mathbf{w})$ as given in (5), because it avoids difficulties which would result from zero feature values $x_j = 0$ in $d_{re}(\mathbf{w},\mathbf{x})$. In principle, the same problems arise for small or zero prototype components in (5) but, in contrast to the data, we can impose additional constraints on the w_j, e.g. of the type $w_j \geq c > 0$. Note that the gradient according to the CS divergence, Eq. (4), is robust with respect to single small feature values or prototype components.

3 Computer Experiments

Two different clinical data sets are used to show the capabilities of the algorithm: the Wisconsin Breast Cancer data set (WBC) from the UCI data repository [14] and the lung cancer data set (LC) taken from [15]. Disregarding 16 vectors containing missing values, the WBC set provides 683 examples in 9 dimensions. The data contains labels corresponding to *malignant* (239 examples) and *benign* (444 examples). For a more detailed description of this data set we refer to [14].

The LC data set contains 100 mass spectra with 22304 features each. It has been down-sampled to 3696 features with no significant loss of information. The data are provided with label information for two classes with 50 examples each, labeled as *cancer* and *control,* respectively. Details about generation and preprocessing of the data can be found in [16] as well as in [17].

We compare the performance of GLVQ based on Euclidean distances with the variants employing the Cauchy-Schwarz (CS). For the WBC data we compare with the Renyi divergence as well. To this end we split the data randomly in training (90% of the data) and test set (10%). Results reported in the following were obtained as averages over 100 randomized splits. Training is performed at constant learning rates. In order to facilitate a fair comparison, we have chosen the optimal learning rate from a range of values for each variant with respect to the achieved performance after 200 training epochs. In the WBC set we employ the learning rates $\eta = 10^{-4}$ (GLVQeu), $\eta = 10^{-6}$ (GLVQcs), and $\eta = 10^{-5}$ (GLVQre). For the LC set, we employ $\eta = 2.5 \times 10^{-3}$ for, both, GLVQeu and GLVQcs. Our first results correspond to the use of one prototype per class, only. Their initial positions are obtained as the mean of 50% randomly selected examples from each class.

After training we introduce a bias θ to the LVQ system: an input vector $\mathbf{x}$ is assigned to class 1 if $d(\mathbf{x}, \mathbf{w}_1) < d(\mathbf{x}, \mathbf{w}_2) + \theta$ where $\mathbf{w}_i$ is the closest prototype representing class i. By varying θ, the full Receiver Operating Characteristics (ROC) of the classifier can be obtained, the results presented in Figure 1 correspond to a threshold-average over the 100 validation runs [18].

The average overall test accuracies after training (for bias $\theta = 0$) and the Area under Curve (AUC) corresponding to the test set ROC curves in Fig. 1 and training set ROC (not shown) are summarized in the follwing tables for the WBC and LC data sets, respectively:

WBC	training acc.	test acc.	AUC (training)	AUC (test)
GLVQeu	85.00 (0.040)	84.46 (0.041)	0.924	0.918
GLVQcs	86.35 (0.003)	85.33 (0.007)	0.923	0.916
GLVQre	84.44 (0.059)	84.17 (0.059)	0.916	0.910

LC	training acc.	test acc.	AUC (training)	AUC (test)
GLVQeu	77.99 (0.006)	75.70 (0.004)	0.809	0.787
GLVQcs	74.06 (0.005)	69.70 (0.009)	0.825	0.796

Here the numbers in parantheses give the standard deviation observed over the 100 validation runs. In general we do not observe drastic differences in the performance. For the high-dimensional LC data set, the use of the CS divergence appears to yield slightly better AUC in the ROC characteristics. Note, however, that at this stage of the investigation we mainly want to demonstrate that the use of divergences in LVQ is feasible.

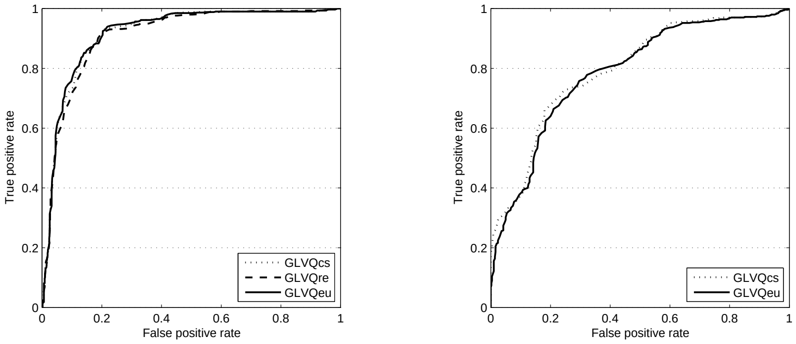

Fig. 1: ROC curves for the WBC data set (left panel) and LC data set (right panel). For both data sets, results are shown on average over 100 randomized data set compositions and for the GLVQ variants based on Euclidean (GLVQeu) and Cauchy-Schwarz (GLVQcs) measure. For WBC we have performed additional experiments using the Renyi ($\alpha = 2$) divergence (GLVQre).

4 Conclusion

We have presented a framework for the derivation of a novel class of LVQ classifiers which are based on information theoretic divergences and their derivatives. The specific examples considered comprise the symmetric Cauchy-Schwarz divergence and a non-symmetric Renyi divergence. We would like to point out, however, that a large variety of differentiable dissimilarities can be employed analogously, a prominent example being the Kullback-Leibler divergence.

The aim of this contribution was to demonstrate the potential usefulness of the approach. To this end, we considered two example data sets from a biomedical context. For the specific data sets considered here we observe very little differences in performance quality when suitable learning rates are chosen. Future applications will have to show to what extent the use of divergences can be advantageous over standard choices. We expect that they will be most useful for data sets where feature vectors are naturally interpreted as probabilities. Besides problems from biology and medicine, we expect favorable performance in, for instance, histogram based classification problems in image processing.

Besides more extensive comparisons in practical applications, future research will also address the extension to generalized divergences which can be used for unnormalized non-negative measures. This step will allow to incorporate relevance learning into the framework and bears the promise to yield very powerful LVQ training schemes.

Acknowledgment: We would like to thank Dr. Markus Kostrzewa, Bruker Daltonik GmbH, Bremen/Germany for providing the LC data set and helpful support in the analysis and interpretation of the data. Frank-M. Schleif was funded by the Federal Ministry of Education and Research, Germany, FZ: 0313833. Ernest Mwebaze was supported by the NUFFIC Project NPT-UGA-

238: Strengthening ICT Training and Research Capacity in Uganda.

References

[1] T. Bojer, B. Hammer, D. Schunk, and K. Tluk von Toschanowitz. Relevance determination in Learning Vector Quantization. In M. Verleysen, editor, *Proc. of Europ. Symp. on Art. Neural Networks (ESANN)*, pages 271–276. d-side, 2001.

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

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

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

[5] E. Jang, C. Fyfe, and H. Ko. Bregman divergences and the self organising map. In C. Fyfe, D. Kim, S.-Y. Lee, and H.Yin, editors, *Intelligent Data Engineering and Automated Learning IDEAL 2008*, pages 452–458. Springer Lecture Notes in Computer Science 5323, 2008.

[6] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

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

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

[9] T. Villmann and S. Haase. Mathematical aspects of divergence based vector quantization using frechet-derivatives. Technical Report MLR-03-2009, Univ. Leipzig/Germany, 2009. ISSN:1865-3960 http://www.uni-leipzig.de/˜compint/.

[10] T. Kohonen. *Self-Organizing Maps.* Springer, Berlin, 2nd edition, 1997.

[11] A. S. Sato and K. Yamada. Generalized learning vector quantization. In *Advances in Neural Information Processing Systems*, volume 8, pages 423–429, 1996.

[12] J.C. Principe, J.F. III, and D. Xu. Information theoretic learning. In S. Haykin, editor, *Unsupervised Adaptive Filtering.* Wiley, 2000.

[13] A. Renyi. *Probability Theory.* North-Holland, Amsterdam, 1970.

[14] D.J. Newman, S. Hettich, 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.

[15] M. Kostrzewa. Lung cancer data, Bruker Daltonik GmbH Bremen. Bruker Daltonik GmbH, Dept. of Bioanalytics, Dr. Markus Kostrzewa, Fahrenheitstrasse 4, D-28359 Bremen, Germany, 2005. personal communication, available on request.

[16] J.K. Boelke, M. Gerhard, F.-M. Schleif, J. Decker, M. Kuhn, T. Elssner, W. Pusch, and M. Kostrzewa. *ClinProTools 2.0 User Documentation*, 2005. Available in the ClinProt - ClinProTools 2.0 Software package.

[17] F.-M. Schleif, T. Villmann, and B. Hammer. Classification in clinical proteomics. *International Journal of Approximate Reasoning*, 47:4–16, 2008.

[18] T. Fawcett. An introduction to ROC analysis. *Patt. Rec. Lett.*, 27:861–874, 2006.