ESANN'2006 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2006, d-side publi., ISBN 2-930307-06-4.

Margin based Active Learning for LVQ Networks

F.-M. Schleif[1] and B. Hammer[2] and Th. Villmann[3]

1- Bruker Daltonik GmbH & Univ. of Leipzig, Dept. of Math. and Comp. Sc., Leipzig, Germany
2- Clausthal Univ. of Technology, Dept. of Comp. Sc., Clausthal-Zellerfeld, Germany
3- Univ. of Leipzig, Clinic for Psychotherapy, Leipzig, Germany

Abstract. In this article, we extend a local prototype-based learning model by active learning, which gives the learner the capability to select training samples and thereby increase speed and accuracy of the model. Our algorithm is based on the idea of selecting a query on the borderline of the actual classification. This can be done by considering margins in an extension of learning vector quantization based on an appropriate cost function. The performance of the query algorithm is demonstrated on real life data.

1 Introduction

In supervised learning, we frequently are interested in training a classifier such that the underlying (unknown) target distribution is well estimated. Whereas traditional approaches usually adapt the model according to all available and randomly sampled training data, the field of active learning restricts to only few actively selected samples. This method avoids the shortcoming of traditional approaches that the average amount of new information per sample decreases during learning and that additional data from some regions are basically redundant. Further, it accounts for the phenomenon which is increasingly common e.g. in bioinformatics or web search that unlabeled data are abundant whereas reliable labeling is costly. Different variants of active and query based learning have been proposed quite early for neural models [2, 7].

In query algorithms proposed so far, samples are chosen according to some heuristic e.g. [2] or in a principled way by optimizing an objective function such as the expected information gain of a query e.g. [7], or the model uncertainty e.g. [3]. A common feature of these query algorithms, however, is that they have been applied to global learning algorithms. Only a few approaches incorporate active strategies into local learning such as [10] where a heuristic query strategy for simple vector quantization is proposed. In this paper we include active learning into a recent margin-based, potentially kernelized learning vector quantization approach, which combines the good generalization ability of margin optimization with the intuitivity of prototype-based local learners where subunits compete for predominance in a region of influence [9].

Now, we briefly review the basic of this kernel-extension of LVQ and its accompanying learning theoretical generalization bounds, therefrom we derive a margin based active learning strategy. We demonstrate the benefit of this mode by comparing the classification performance of the algorithm on randomly selected training data and active strategies for several data set stemming from clinical proteomics.

2 Generalized relevance learning vector quantization

Standard LVQ and variants as proposed by KOHONEN constitute popular simple and intuitive prototype based methods, but they are purely heuristically motivated local learners [11]. They suffer from the problem of instabilities for overlapping classes. Further they strongly dependent on the initialization of prototypes, and a restriction to classification scenarios in Euclidean space. Generalized relevance learning vector quantization (GRLVQ) has been introduced by the authors to cope with these problems [9]. It is based on a cost function such that neighborhood incorporation, integration of relevance learning, and kernelization of the approach become possible which gives Supervised Relevance Neural Gas denoted as SRNG [9]. The method can be accompanied by a large margin generalization bound [8], which is directly connected to the cost function of the algorithm and which opens the way towards active learning strategies, as we will discuss in this article.

We first introduce the basic algorithm. Input vectors are denoted by $\mathbf{v}$ and their corresponding class labels by $c_{\mathbf{v}}$, $\mathcal{L}$ is the set of labels (classes). Let $V \subseteq \mathbb{R}^{D_V}$ be a set of inputs $\mathbf{v}$. The model uses a fixed number of representative prototypes (weight vectors, codebook vectors) for each class. Let $\mathbf{W} = \{\mathbf{w_r}\}$ be the set of all codebook vectors and $c_{\mathbf{r}}$ be the class label of $\mathbf{w_r}$. Furthermore, let $\mathbf{W}_c = \{\mathbf{w_r} | c_{\mathbf{r}} = c\}$ be the subset of prototypes assigned to class $c \in \mathcal{L}$. The task of vector quantization is realized by the map Ψ as a winner-take-all rule, i.e. a stimulus vector $\mathbf{v} \in V$ is mapped onto that neuron $\mathbf{s} \in A$ the pointer $\mathbf{w}_s$ of which is closest to the presented stimulus vector $\mathbf{v}$,

$$\Psi^{\lambda}_{V \to \mathcal{A}} : \mathbf{v} \mapsto \mathbf{s}(\mathbf{v}) = \mathrm{argmin}_{\mathbf{r} \in A} d^{\lambda}(\mathbf{v}, \mathbf{w_r}) \tag{1}$$

with $d^{\lambda}(\mathbf{v}, \mathbf{w})$ being an arbitrary differentiable similarity measure, which may depend on a parameter vector λ. The subset of the input space

$$\Omega^{\lambda}_{\mathbf{r}} = \left\{\mathbf{v} \in V : \mathbf{r} = \Psi^{\lambda}_{V \to A}(\mathbf{v})\right\} \tag{2}$$

which is mapped to a particular neuron $\mathbf{r}$ according to (1), forms the (masked) receptive field of that neuron. If the class information of the weight vector is used, the boundaries $\partial\Omega^{\lambda}_{\mathbf{r}}$ generate the decision boundaries for classes. A training algorithm should adapt the prototypes such that for each class $c \in \mathcal{L}$, the corresponding codebook vectors $\mathbf{W}_c$ represent the class as accurately as possible.

To achieve this goal, GRLVQ optimizes the following cost function, which is related to the number of misclassifications of the prototypes, via a stochastic gradient descent:

$$\mathrm{Cost}_{\mathrm{GRLVQ}} = \sum_{\mathbf{v}} f(\mu_{\lambda}(\mathbf{v})) \quad \text{with} \quad \mu_{\lambda}(\mathbf{v}) = \frac{d^{\lambda}_{\mathbf{r}_+} - d^{\lambda}_{\mathbf{r}_-}}{d^{\lambda}_{\mathbf{r}_+} + d^{\lambda}_{\mathbf{r}_-}} \tag{3}$$

where $f(x) = (1 + \exp(-x))^{-1}$ is the standard logistic function, $d^{\lambda}_{\mathbf{r}_+}$ is the similarity of the input vector $\mathbf{v}$ to the nearest codebook vector labeled with $c_{\mathbf{r}_+} = c_{\mathbf{v}}$, say $\mathbf{w}_{\mathbf{r}_+}$, and $d^{\lambda}_{\mathbf{r}_-}$ is the similarity measure to the best matching prototype labeled with $c_{\mathbf{r}_-} \neq c_{\mathbf{v}}$, say $\mathbf{w}_{\mathbf{r}_-}$. Note that the term $f(\mu_{\lambda}(\mathbf{v}))$ scales the differences of the closest two competing prototypes to $(-1, 1)$, negative values correspond to correct classifications. As shown in [13], this cost function shows robust behavior whereas original LVQ2.1 yields divergence. Our active learning approach holds for each kind of such G(R)LVQ-type learning.

3 Margin based active learning

The first dimensionality independent large margin generalization bound of LVQ classifiers has been provided in [6]. For GRLVQ-type learning, a further analysis is possible, which accounts for the specific cost function and the fact that the similarity measure is adaptive during training. Assume, for the moment, that the squared Euclidean metric is used, and a two-class problem with labels $\{-1, 1\}$ is given[1]. We further assume that data are chosen i.i.d. according to the data distribution $P(V)$ and the class labels are determined by an unknown function f. Generalization bounds limit the error, i.e. the probability that the learned classifier does not classify given data correctly:

$$E_P(\Psi) = P(c_{\mathbf{v}} \neq \Psi^{\lambda}_{V\to\mathcal{A}}(\mathbf{v})) \tag{4}$$

Given a classifier Ψ and a sample $(\mathbf{v}, c_{\mathbf{v}})$, we define the margin as

$$M_\Psi(\mathbf{v}, c_{\mathbf{v}}) = -d^{\lambda}_{\mathbf{r}+} + d^{\lambda}_{\mathbf{r}-}, \tag{5}$$

i.e. the difference of the distance of the data point from the closest correct and the closest wrong prototype. (To be precise, we refer to the absolute value as the margin.) For a fixed parameter $\rho \in (0, 1)$, the loss function is defined as

$$L : \mathbb{R} \to \mathbb{R}, t \mapsto \begin{cases} 1 & \text{if } t \leq 0 \\ 1 - t/\rho & \text{if } 0 < t \leq \rho \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

The term

$$\hat{E}^L_m(\Psi) = \sum_{\mathbf{v}\in V} L(M_\Psi(\mathbf{v}, c_{\mathbf{v}}))/|V| \tag{7}$$

denotes the empirical error on the training data. It counts the data points which are classified wrong and, in addition, punishes all data points with too small margin.

Using techniques from [1], we can limit the difference of the error (4) and the empirical error (7) by

$$P\left(E_P(\Psi) > \hat{E}^L_m(\Psi) + K \cdot \frac{\ln|V|}{\rho \cdot \sqrt{|V|}} \cdot \sqrt{\ln(1/\delta)} \cdot |\mathbf{W}|^2 B^3\right) \leq \delta \tag{8}$$

with probability at least $\delta \in (0, 1)$, whereby K is a universal constant, and B limits the norm of the samples. This bound holds for every prototype based learning algorithm with diagonal Euclidean metric and *adaptive* relevance parameters as long as the absolute sum of the relevance parameters λ is restricted by 1, whereby the parameters may even be adapted locally for each prototype vector. The basic observation to prove this bound consists in the possibility to express a bound for the deviation of the empirical error $\hat{E}_m(\Psi)$ and the error $E_P(\Psi)$ by the Gaussian complexity of the function class defined by the model, as shown in [1]. The Gaussian complexity of GRLVQ networks with local adaptive diagonal metric can be easily estimated, because the classifier can be expressed as a Boolean formula in simple terms, which compare the distance of a given sample for only two prototypes. This, is described by a quadratic form, for which the Gaussian complexity is known [1]. The exact proof of this fact can be found in [8]. It should be mentioned, that the margin (5) occurs as nominator in the cost function of GRLVQ. Hence GRLVQ and SRNG maximize its margin during training according to this cost function.

This generalization bound in terms of the margin proposes an elegant scheme to transfer margin based active learning to local learners. Margin based sample selection

[1] These constraints are technical to derive the generalization bounds which have already been derived by two of the authors in [8], the active learning strategies work also for > 2 classes and alternative metrics

has been presented e.g. in the context of SVM in [5, 12]. Obviously, the generalization ability of the GRLVQ algorithm does only depend on the points with too small margin (5). Thus, only the extremal margin values need to be limited and a restriction of the respective update to extremal pairs of prototypes would suffice. This argument proposes schemes for active data selection if a fixed and static pattern set is available: We fix a monotonically decreasing non-negative function $L^c : \mathbb{R} \to \mathbb{R}$ and actively select training points from a given sample, in analogy to e.g. [12], based on the probability $L^c(M_\Phi(\mathbf{v}, c_\mathbf{v}))$ for sample $\mathbf{v}$. Thereby, different realizations are relevant:

1. $L^c(t) = 1$ if $t \leq \rho$, otherwise, it is 0. That means, all samples with margin smaller than ρ are selected (Threshold strategy).

2. $L^c(t) = 1$ for $t > 0$ and $L^c(t) \sim |t|^\alpha$, otherwise, i.e. the size of the margin determines the probability of $\mathbf{v}$ being chosen annealed by a parameter α (Probabilistic strategy).

Both strategies focus on the samples which are not yet sufficiently represented in the model. Therefore they directly aim at an improvement of the generalization bound (8). Strategy (1) allows an adaptation of the margin parameter ρ during training in accordance to the confidence of the model in analogy to the recent proposal [12] for SVM. For each codebook vector $\mathbf{w_r} \in W$ we introduce a new parameter α_r measuring the mean distance of data points in its receptive field (2) to the current prototype $\mathbf{w_r}$. This parameter can be easily computed during training as a moving average with no extra costs[2]. We choose ρ_r locally as $\rho_r = 2 \cdot \alpha_r$. Thus, points which margin compares favorable to the size of the receptive fields are already represented with sufficient security and, hence, they are abandoned. For strategy (2), a confidence depending on the distance to the closest correct prototype and the overall classification accuracy can be introduced in a similar way. Doing this the normalized margin is taken as a probability measure for data selection.

We would like to mention that, so far, we have restricted active selection strategies to samples where all labels are known beforehand, because the closest correct and wrong prototype have to be determined in (5). This setting allows to improve the training speed and performance of batch training. If data are initially unlabeled and queries can be asked for a subset of the data, we can extend these strategies in an obvious way towards this setting: in this case, the margin (5) is given by the closest two prototypes which possess a different class label, whereby the (unknown) class label of the sample point has no influence. $L^c(t)$ is substituted by $L^c(|t|)$.

4 Experiments and Results

We now compare the SRNG with randomly selected samples with the SRNG using the proposed query strategies. The first data set is the *Wisconsin Breast Cancer*-Data set (WDBC) as given by UCI [4]. It consist of 569 measurements with 30 features in 2 classes. It is processed with 50% of the samples for training and the remaining samples for test. The other two data sets are taken from proteomic studies named as Proteom$_1$ and Proteom$_2$. The Proteom$_1$ data set consists of 199 samples in three classes with 250 dimensions. The data set Proteom$_2$ consists of 737 measurements with two classes and 148 dimensions. For classification, we use 6 prototypes for the WDBC data,9 prototypes for the Proteom$_1$ dataset and 10 for Proteom$_2$. The classification results are

[2] The extra computational time to determine the active learning control variables is negligible

	SRNG		$\text{SRNG}_{\text{active strategy 1}}$			$\text{SRNG}_{\text{active strategy 2}}$		
	Rec.	Pred.	Rec.	Pred.	Rel. #Q	Rec.	Pred.	Rel. #Q
WDBC	91%	90%	92%	92%	30%	91%	90%	55%
Proteom_1	76%	74%	80%	71%	38%	80%	71%	70%
Proteom_2	73%	70%	70%	72%	52%	69%	72%	81%

Table 1: Recognition (Rec.) vs. Prediction (Pred.) rates for the SRNG algorithm using different query strategies. For the two active Learning strategies (Threshold - 1; Probabilistic - 2) the relative number of queries is denoted by (Rel. #Q).

given in Tab. 1. Features of all data sets have been normalized. First we upper bounded the data set by 1.0 and subsequently data are transformed such that we end with zero mean and variance 1.0. For the active learning constraints the margin has been scaled by the number of input dimensions.

We applied the SRNG algorithm using the different queries strategies as introduced above. The results for recognition and prediction rates are shown in Tab. 1[3].

For the WDBC dataset and the Proteom_2 data set we found small improvements in the prediction accuracy using the active strategy 1. The Proteom_1 data set showed a small over-fitting behavior using the new query strategies and a small decrease in the overall prediction accuracy. Both new query strategies were capable to significantly decrease the necessary number of queries by keeping at least reliable prediction accuracies with respect to a random query approach.

5 Conclusion

Margin based active learning strategies for GRLVQ/SRNG networks have been studied. We compared two alternative query strategies incorporating the margin criterion of the GLVQ networks with a random query selection. Both active learning strategies show reliable or partially better results in their generalization ability with respect to the random approach. Thereby we found a signification faster convergence with a much lower number of necessary queries. For the *threshold strategy* we found that it shows an overall stable behavior with good prediction rate and a significantly decrease in processing time. Due to the automatically adapted threshold parameter the strategy is quite simple but depends on a sufficiently well estimation of the local data distribution. By scaling the threshold parameter an application specific choice between prediction accuracy and speed can be obtained. The *probabilistic strategy* has been found to get similar results with respect to the prediction accuracy but the number of queries is quite dependent of the annealing strategy, simulated less restrictive constraints showed a faster convergence but over-fitting on smaller training data sets. Especially, for larger data sets the proposed active learning strategies show great benefits in speed and prediction. Especially for the considered mass spectrometric cancer data sets an overall well performance improvement has been observed. This is interesting from a practical point of view, since the technical equipment for measuring e.g. a large number of mass spectrometric data becomes more and more available.

[3] The relative number of queries is calculated with respect to the maximal number of queries possible up to convergence of SRNG using the corresponding query strategy.

References

[1] P. Bartlett and S. Mendelsohn. Rademacher and gaussian complexities: risk bounds and structural results. *Journal of Machine Learning and Research*, 3:463–482, 2002.

[2] E.B. Baum. Neural net algorithms that learn in polynomial time from examples and queries. *IEEE Transactions on Neural Networks*, 2:5–19, 1991.

[3] L. M. Belue, K. W. Bauer Jr., and D. W. Ruck. Selecting optimal experiments for multiple output multilayer perceptrons. *Neural Computation*, 9:161–183, 1997.

[4] C. Blake and C. Merz. UCI repository of machine learning databases., 1998. available at: http://www.ics.uci.edu/ mlearn/MLRepository.html.

[5] C. Cambell, N. Cristianini, and A. Smola. Query learning with large margin classifiers. In *International Conference in Machine Learning*, pages 111–118, 2000.

[6] K. Crammer, R. Gilad-Bachrach, A.Navot, and A.Tishby. Margin analysis of the lvq algorithm. In *Proc. NIPS 2002*, http://www-2.cs.cmu.edu/Groups/NIPS/NIPS2002/NIPS2002preproceedings/index.html, 2002.

[7] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby. Information, prediction and uery by committee. In *Advances in Neural Information Processing Systems 1993*, pages 483–490, 1993.

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

[9] Barbara Hammer, Marc Strickert, and Thomas Villmann. Supervised neural gas with general similarity measure. *Neural Processing Letters*, 21(1):21–44, February 2005.

[10] M. Hasenjäger and H. Ritter. Active learning with local models. *Neural Processing Letters*, 7:107–117, 1998.

[11] Teuvo Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995. (2nd Ext. Ed. 1997).

[12] P. Mitra, C.A. Murthy, and S.K. Pal. A probabilistic active support vector learning algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(3):412–418, 2004.

[13] A. Sato and K. Yamada. A formulation of learning vector quantization using a new misclassification measure. In A. K. Jain, S. Venkatesh, and B. C. Lovell, editors, *Proceedings. Fourteenth International Conference on Pattern Recognition*, volume 1, pages 322–5. IEEE Computer Society, Los Alamitos, CA, USA, 1998.