# Relevance LVQ versus SVM

Barbara Hammer[1], Marc Strickert[1], and Thomas Villmann[2]

[1] Department of Mathematics/Computer Science,
University of Osnabrück, D-49069 Osnabrück, Germany
[2] Clinic for Psychotherapy and Psychosomatic Medicine, University of Leipzig,
Karl-Tauchnitz-Straße 25, D-04107 Leipzig, Germany

**Abstract.** The support vector machine (SVM) constitutes one of the most successful current learning algorithms with excellent classification accuracy in large real-life problems and strong theoretical background. However, a SVM solution is given by a not intuitive classification in terms of extreme values of the training set and the size of a SVM classifier scales with the number of training data. Generalized relevance learning vector quantization (GRLVQ) has recently been introduced as a simple though powerful expansion of basic LVQ. Unlike SVM, it provides a very intuitive classification in terms of prototypical vectors the number of which is independent of the size of the training set. Here, we discuss GRLVQ in comparison to the SVM and point out its beneficial theoretical properties which are similar to SVM whereby providing sparse and intuitive solutions. In addition, the competitive performance of GRLVQ is demonstrated in one experiment from computational biology.

## 1 Introduction

Starting with the pioneering work of Vapnik and Chervonenkis, empirical risk minimization and structural risk minimization have been identified and quantified as the two goals when training a supervised machine learning classifier [15]. Thereby, the structural risk is not necessarily directly connected to the number of parameters; rather the interior capacity of the function class implemented by the respective model measured in terms of the VC dimension, for example, is the relevant quantity. Because of this fact, effective machine learning models can be designed also for very high dimensional data and complex underlying regularities.

The support vector machine constitutes a prime example of a machine learner which directly aims at structural risk minimization during training [3]. The structural risk of the SVM is given by the classification margin. A large margin ensures good mathematical generalization bounds and, in practice, excellent generalization ability of SVMs. Two further benefits of the SVM can be observed: the dual Lagrangian problem of SVM training can be solved optimum in polynomial time, thus SVM training is guaranteed to converge to a global optimum. Input data are implicitly mapped to a high dimensional feature space by a kernel, which offers a natural interface to adapt the model to specific settings and to integrate prior knowledge [5]. However, one drawback of the SVM is the expansion of a solution in terms of the dual Lagrangian variables: given training data $(x^i, y_i) \in \mathbb{R}^n \times \{-1, 1\}$ and a kernel $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, the final SVM classifier into the two classes $-1$ and $1$ is of the form

$$x \mapsto \begin{cases} 1 & \text{if } \sum_i \alpha_i y_i k(x^i, x) \geq \theta \\ -1 & \text{otherwise} \end{cases}$$

where $\alpha_i$ are the dual Lagrangian variables optimized during training, and $\theta$ is the bias. The Lagrangian variables $\alpha_i$ are nonvanishing for support vectors $x^i$. Support vectors are all points which are errors, all points which have a too small margin, and points which are correctly classified but have a minimum margin, i.e. 'extreme' points of the training set. Thus the solution is not formulated in terms of 'typical' points but in terms of 'atypical' training examples. In addition, the number of support vectors is usually given by a fraction of the training set, i.e. the size of the found solution scales linearly with the size of the training set. Thus training is quadratic with respect to the training set size, and classification is linear. Alternatives to the SVM which expand solutions in terms of typical vectors have been proposed, such as Bayes-point machines [10].

Here, we focus on another very intuitive learning model which is based on a different training paradigm: prototype based learning vector quantization (LVQ) as proposed by Kohonen [11]. We discuss a recent extension of simple LVQ which integrates an adaptive metric, and we show that the new model has the same beneficial properties as SVM: it converges to a global optimum, it allows to integrate prior knowledge in a kernelized version, and it can be interpreted as a large margin optimizer for which mathematical dimensionality independent generalization bounds exist. In contrast to SVM, the method provides an expansion in terms of typical vectors, and the solution does not scale with the size of the training set, thus training time is linear in the number of training patterns and the classification effort is constant. In addition, we demonstrate that the method is competitive to SVM in one experiment.

## 2 GRLVQ

Generalized relevance learning vector quantization (GRLVQ) has been proposed in [9] as an extension of simple LVQ. Assume a finite set of training data $(x^i, y_i) \in \mathbb{R}^n \times \{1, \ldots, C\}$ is given, $C$ being the number of different classes. A GRLVQ network represents every class $c$ by a finite set of prototypes in $\mathbb{R}^n$. A prototype is denoted by $w^r$, $c_r$ being its class. A new signal is classified by the winner-takes-all rule $x \mapsto c(x) = c_r$ such that $d^\lambda(x, w^r)$ is minimum, whereby $d^\lambda(x, w^r)$ is the scaled squared Euclidean distance $d^\lambda(x, w^r) = \sum_i \lambda_i (x_i - w_i^r)^2$ with relevance terms $\lambda_i \geq 0$ for each dimension $i$ which add up to 1, i.e. $\sum_i \lambda_i = 1$.

Training adapts the prototypes and the relevance terms according to the given data. The training rule is formally derived as a stochastic gradient descent method on the cost function

$$E_{\mathrm{GRLVQ}} = \sum_i \mathrm{sgd}\left(\frac{d_{r^+} - d_{r^-}}{d_{r^+} + d_{r^-}}\right)$$

where $\mathrm{sgd}(x) = (1 + \exp(-x))^{-1}$ is the logistic function, the sum is over all training vectors $x^i$, $d_{r^+}$ denotes the distance of $x^i$ from the closest prototype $w_{r^+}$ with the same class as $x^i$, and $d_{r^-}$ denotes the distance of $x^i$ from the closest prototype $w_{r^-}$ with a different class label than $x^i$. Note that the nominator $d_{r^+} - d_{r^-}$ is negative if and only if the point $x^i$ is classified correctly. Thus minimization of the cost function aims at maximizing the number of correctly classified points.

The learning rule of GRLVQ can be derived thereof taking the gradients of the summands. The qualitative form of the update rules is the same as for basic LVQ: having presented a point $x^i$, the closest correct prototype is adapted by $\triangle w_{r^+} = \eta \cdot c_1 \cdot (w^{r^+} -$

$x^i$), where $\eta$ is the learning rate. $c_1$ is a factor the precise form of which is of no importance for this article. The closest incorrect prototype is adapted by $\triangle w_{r^-} = -\eta \cdot c_2 \cdot (w^{r^-} - x^i)$, $c_2$ being another factor, and the relevance terms are adapted by $\triangle \lambda_l = -\eta \cdot (c_3 \cdot (w_l^{r^+} - x_l^i)^2 - c_4 \cdot (w_l^{r^-} - x_l^i)^2)$, followed by normalization, $c_3$ and $c_4$ being additional factors. It has been discussed in [9] that this update scheme can be interpreted as a straightforward application of Hebbian learning, as used for basic LVQ. As pointed out in [9], where also the precise update formulas can be found, the additional scaling terms $c_i$ and the adaptation of the metric via relevance terms leads to a stable and efficient alternative of original LVQ which can also deal with very high dimensional and heterogeneous data.

Unlike SVM, a GRLVQ classifier is formulated in terms of prototypes. These are adapted during training such that typical positions of the data set are found; thus they provide a very intuitive classification. In addition the number of prototypes is fixed and it has to be chosen according to the number of modes of the underlying distribution, but it does not depend on the number of concrete training examples.

### 2.1 Local optima

GRLVQ as introduced above constitutes a stochastic gradient descent method on a cost function, thus it might converge to a local optimum of the cost function. This fact can be prevented introducing neighborhood cooperation of the prototypes as proposed in [6]. In the article [6], the cost function of GRLVQ is merged with the cost function of Neural Gas (NG), which constitutes an unsupervised and very reliable clustering algorithm [12]. NG introduces a data optimum neighborhood cooperation of the prototypes such that initialization of the prototypes has almost no effect on the training result and the prototypes spread faithfully among the data points. As demonstrated in [6], the combination of NG with GRLVQ allows to apply GRLVQ also to highly multimodal settings, and the modification reliably converges to global optima of the GRLVQ cost function.

### 2.2 Kernelization

GRLVQ relies on the weighted Euclidean metric. Because of adaptive relevance terms, irrelevant and noisy dimensions can automatically be detected and the method can also deal with high dimensional input data. However, nonlinear transformations of the input data and correlations of the input dimensions are not accounted for and the weighted Euclidean metric might be an inappropriate similarity measure for such situations. GRLVQ (and also the combination with NG) is formulated in an abstract way as cost minimization. Therefore we can achieve greater flexibility of the method by just substituting the Euclidian similarity measure $d^\lambda$ by another problem dependent differentiable similarity measure with possibly adaptive parameters $\lambda$ in the cost function. For every such choice the update formulas of GRLVQ can formally be derived taking the derivatives. Thus prior knowledge can be integrated into GRLVQ using a different similarity measure in the cost function. This new version with similarity measure $\tilde{d}^\lambda$ instead of $d^\lambda$ can be interpreted as a kernelized of standard GRLVQ iff a function $\Phi$ exists such that $\tilde{d}(x,y) = d^\lambda(\Phi(x), \Phi(y))$ holds. The article [13] discusses conditions under which such $\Phi$ can be found. A sufficient condition is, for example, that $\tilde{d}$ is symmetric with $\tilde{d}(0,0) = 0$, and $-\tilde{d}$ being conditionally positive definite. Examples for alternative kernels especially suited e.g. for time series data have been proposed in [7].

### 2.3 Large margin bounds

Generalization bounds for standard LVQ have recently been presented in the article [4]. GRLVQ differs from LVQ in the essential property that the metric is also adapted during training, thus its capacity is larger. Nevertheless, it is possible to derive dimensionality independent large margin bounds for GRLVQ networks: assume the class of GRLVQ classifiers with $p$ prototypes, two classes, adaptive squared Euclidean metric, and weights and inputs restricted to the length $b$ is considered. The empirical loss of a given training set is the number of misclassified points. To derive a large margin bound, this loss is modified in analogy to [1] in the following way: we fix a margin parameter $\rho > 0$. The 'security' or margin of the classification of a point $x^i$ is given by the quantity $-d_{r^+} + d_{r^-}$, the distance of the point from the closest correct prototype compared to the closest incorrect one. If this term is negative, the classification is incorrect. If it is positive, it is correct, but possibly only with a small margin. We define the loss function

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

and the empirical error including the margin, given $m$ training points $x^i$,

$$\hat{E}_m^L := \sum_{i=1}^{m} L(-d_{r^+} + d_{r^-})/m$$

This error collects all misclassification and it also punishes all classifications which are correct, but which have a margin smaller than $\rho$. We are, of course, not interested in this empirical loss, but in the generalization ability of a given GRLVQ network, i.e. the quantity

$$E := P(-d_{r^+} + d_{r^-} \leq 0)$$

for an arbitrary data point $(x, y)$ chosen according to a fixed probability measure $P$ on $\mathbb{R}^n \times \{0, 1\}$. If the training points $(x^i, y_i)$ are chosen independent and identically distributed according to $P$, one can show that the error $E$ deviates from the empirical error $\hat{E}_m^L$ by a term of order

$$\frac{p^2}{\rho \cdot \sqrt{m}} \cdot (b^3 + \sqrt{\ln 1/\delta})$$

with confidence $\delta$ [8]. This bound does not depend on the input dimensionality, thus GRLVQ is well suited also for high dimensional input data. Rather, the larger the margin $\rho$, the better the generalization bound. Remarkably, the margin is the security according to which the classification is done, i.e. the term $-d_{r^+} + d_{r^-}$. This term is included in the nominator of the cost function of GRLVQ. Thus GRLVQ directly aims at margin optimization during training.

## 3 Experiments

Having discussed the analogy of GRLVQ and SVM with respect to large margin generalization bounds, convergence to a global optimum, and flexibility by a different choice

| $\text{GRLVQ}_{\text{EUC}}$ | $\text{GRLVQ}_{\text{LIK}}$ | HMM | $\text{SVM}_{\text{LIK}}$ | $\text{SVM}_{\text{TOP}}$ | $\text{SVM}_{\text{FK}}$ |
|---|---|---|---|---|---|
| 95.6 | 96.5 | 94 | 96.3 | 94.6 | 94.7 |

**Table 1.** Accuracy (in $\%$) of various methods achieved on the IPsplice dataset, the classification accuracy on the test set is reported for the models. The results for alternatives to GRLVQ are taken from [14].

of the similarity measure, we add an experiment where we compare several recent classification results achieved by the SVM with GRLVQ networks. The task is to distinguish pieces of human DNA according to the three classes donor, acceptor, or neither. These classes refer to the borders between coding and non-coding regions of the DNA. The data set is the publicly available IPSplice data set from the UCI repository [2], containing 765 acceptors, 767 donors, and 1654 decoys. The inputs are given by 60 nucleotides around a potential splice site, whereby we encode the $4$ nucleotides T, C, G, A in $\mathbb{R}^3$ Thus the input dimensionality is $180$.

GRLVQ is trained with $8$ prototypes per class, the standard weighted Euclidean metric, and, alternatively, the locality improved metric (LIK) which is designed to take local correlations of time series into account. LIK first adds up the distances of entries in small consecutive windows (in our case windows of radius $3$ to account for potential reading frames). The accumulated distances within windows are taken to the power of $d = 3$, thus local correlations are computed. The result of all windows is afterwards accumulated weighted with adaptive relevance terms $\lambda_i$. Training has been done for $2000$ epochs and the learning rates have been optimized for this data set. Initial neighborhood cooperation of NG has been included to ensure convergence to a global optimum.

Results of a 10-fold crossvalidation of GRLVQ with the standard metric (EUC) and LIK are reported in Table1. The result is compared to the classification accuracy achieved by hidden Markov models (HMM), and the SVM with different kernels, the locality improved kernel, and two kernels derived from a statistical model (TOP and FK) [14]. Note that the accuracy of our method is competitive to the solutions found by SVM, whereby the achieved classifiers are very sparse for our setting (only $8$ prototypes per class) and the training complexity is only linear in the size of the training set.

Interestingly, the relevance profile achieved by GRLVQ mirrors biological knowledge. As depicted in Fig. 1, the relevance terms are maximum in the direct neighborhood of the potential splice site (place $0$), indicating consensus strings at this region. In addition, the left part is emphasized a bit more than the right one, corresponding to a pyrimidine rich region before potential acceptors.

## 4 Conclusions

We have discussed the classifier GRLVQ in comparison to the SVM and we have pointed out its similar theoretical properties: it constitutes a large margin optimizer with convergence to a global optimum and large flexibility due to a general metric. Thereby, GRLVQ provides more intuitive solutions in terms of typical vectors than SVM, and the size of the found classifier does not scale with the size of the training set. These theoretical facts have been accompanied by one experiment. Thus, GRLVQ constitutes a valuable alternative to SVMs.

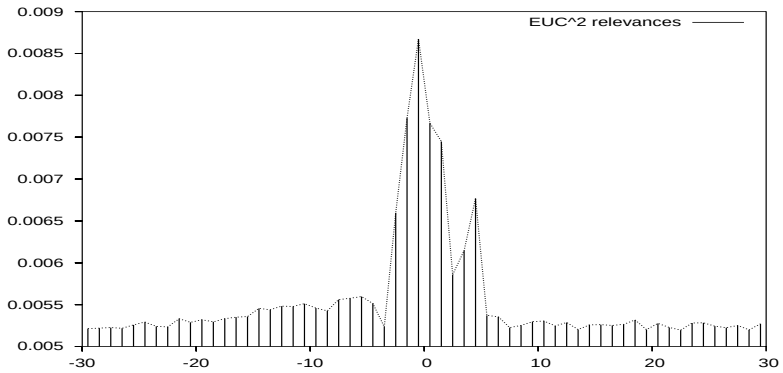


**Fig. 1.** Relevance profile for the weighted Euclidean similarity measure for the IPSplice data set. Each position shows the average over the successive relevance factors for the four nucleotides at one position of the window.

## References


1. P.L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning and Research* 3:463-482, 2002.
2. C.L. Blake and C. J. Merz, UCI Repository of machine learning databases, Irvine, CA: University of California, Department of Information and Computer Science.
3. C. Burges. A tutorial on support vector machines for pattern recognition. *Knowledge Discovery and Data Mining*, 2(2), 1998.
4. K. Crammer, R. Gilad-Bachrach, A. Navot, and A. Tishby. Margin analysis of the LVQ algorithm. In: *NIPS 2002*.
5. T. Gärtner. A survey of kernels for structured data. *SIGKDD Explorations* 5(2):49-58, 2003.
6. B. Hammer, M. Strickert, and T. Villmann. Learning vector quantization for multimodal data. In J.R. Dorronsoro, editor, *ICANN'02*, pages 370-375. Springer, 2002.
7. B. Hammer, M. Strickert, and T. Villmann. Supervised Neural Gas with General Similarity Measure. To appear in Neural Processing Letters.
8. B. Hammer, M. Strickert, and T. Villmann. On the generalization ability of GRLVQ networks. Osnabrücker Schriften zur Mathematik, Preprint, no. 249, 10/2003.
9. B. Hammer and T. Villmann. Generalized relevance learning vector quantization. *Neural Networks*, 15:1059–1068, 2002.
10. R. Herbrich, T. Graepel, and C. Campbell. Bayes point machines. *Journal of Machine Learning Research*, 1:245–279, 2001.
11. T. Kohonen. *Self-Organizing Maps*. Springer, 1995.
12. T. Martinetz, S.G. Berkovich, and K.J. Schulten. 'Neural-gas' networks for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks*, 4(4):558–569, 1993.
13. B. Schölkopf. *The kernel trick for distances*. Technical Report MSR-TR-2000-51. Microsoft Research, Redmond, WA, 2000.
14. S. Sonnenburg, G.Rätsch, A. Jagota, and K.-R. Müller. New methods for splice site recognition. In: J. R. Dorronsoro (ed.), *ICANN'2002*, pages 329–336, Springer, 2002.
15. V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications* 16(2):264-280, 1971.