A Comparison of Validation Methods for Learning Vector Quantization and for Support Vector Machines on Two Biomedical Data Sets

David Sommer and Martin Golz**

Department of Computer Science, University of Applied Sciences Schmalkalden

Abstract. We compare two comprehensive classification algorithms, support vector machines (SVM) and several variants of learning vector quantization (LVQ), with respect to different validation methods. The generalization ability is estimated by "multiple-hold-out" (MHO) and by "leave-one-out" (LOO) cross validation method. The $\xi\alpha$-method, a further estimation method, which is only applicable for SVM and is computationally more efficient, is also used.
Calculations on two different biomedical data sets generated of experimental data measured in our own laboratory are presented. The first data set contains 748 feature vectors extracted of posturographic signals which were obtained in investigations of balance control in upright standing of 48 young adults. Two different classes are labelled as "without alcoholic impairment" and "with alcoholic impairment". This classification task aims the detection of small unknown changes in a relative complex signal with high inter-individual variability.
The second data set contains 6432 feature vectors extracted of electroencephalographic and electroocculographic signals recorded during overnight driving simulations of 22 young adults. Short intrusions of sleep during driving, so-called microsleep events, were observed. They form examples of the first class. The second class contains examples of fatigue states, whereas driving is still possible. If microsleep events happen in typical states of brain activity, the recorded signals should contain typical alterations, and therefore discrimination from signals of the second class, which do not refer to such states, should be possible.
Optimal kernel parameters of SVM are found by searching minimal test errors with all three validation methods. Results obtained on both different biomedical data sets show different optimal kernel parameters depending on the validation method. It is shown, that the $\xi\alpha$-method seems to be biased and therefore LOO or MHO method should be preferred.
A comparison of eight different variants of LVQ and six other classification methods using MHO validation yields that SVM performs best for the second and more complex data set and SVM, GRLVQ and OLVQ1 show nearly the same performance for the first data set.

1 Introduction

Support Vector Machines and Learning Vector Quantization are two efficient methods of machine learning which are approved e.g. in handwritten word

** Corresponding author: FH Schmalkalden, mail box 10 04 52, D-98574 Schmalkalden, Germany

recognition, robotic navigation, textual categorization, face recognition and time series prediction [Müller et al. (2001), Osuna et al. (1997), Cao and Tay (2003)]. The aim of this paper is to compare both methods on two real world biomedical data sets using several variants of LVQ and of SVM and of some other classification algorithms. Among them are several methods of automatic relevance detection, e.g. recently introduced GRLVQ [Hammer and Villmann (2002)].

Calculations were done on two fully different biomedical data sets coming from two different disciplines: biomechanics and electrophysiology applied to psychophysiology. The first data set comes out of an investigation of balance control in upright standing of 48 young volunteers. They were investigated without impairment and 40 minutes after consumption of 32 grams of alcohol. Therefore we have two different classes which are labelled as "without alcoholic impairment" and "with alcoholic impairment". Subjects had to stand on a solid plate with elevated arms and turned hands, the so-called supination position [Golz et al. (2004)]. Signals of four force sensors located between plate and ground are combined to calculate the two-dimensional signal of the centre-of-foot-pressure, which is a sensitive measure of postural sway. From both signals the power spectral densities were estimated by Burg's autoregressive modelling method. This two-class problem is nearly weight out and consists of 376 feature vectors of 40 components. This classification task aims the detection of small unknown changes in a relative complex signal with high inter-individual variability.

The second and clearly more extensive and higher-dimensional data set contains power spectral densities of electroencephalograms (EEG) and electrooculograms (EOG) recorded during strong fatigue states and during microsleep events of 16 young car drivers [Sommer and Golz (2003)]. Microsleep events are defined as short intrusions of sleep into ongoing wakefulness during attentional tasks and are coupled to dangerous attention losses. The decision which behavioural event belongs to "microsleep events" and which to "strong fatigue" was made by two independent experts. This was mainly done by visual scoring of video recordings. Subjects had to drive overnight starting at 1:00 a.m. (7 x 40 min) in our driving simulation lab under monotonic conditions. Small segments (duration 6 sec) of EEG and EOG were taken during both events. A comparison of several spectral estimation methods yields that Burg's autoregressive method is outperformed by the simple periodogram method [Sommer and Golz (2003)]. In this paper we therefore report only on results for the second data set using the latter method. The extracted data set contains 5728 feature vectors of 207 components. This classification task also aims the detection of small unknown changes in a relative complex signal with high inter-individual variability. If microsleep events happen in typical states of brain activity, the recorded signals should contain typical alterations, and therefore discrimination from signals of the second class, which do not refer to such states, should be possible.

There exists no expert knowledge to solve both classification tasks. Knowledge extraction in both fields is strongly impaired due to high inter-individual differences in the observed biosignals and due to high noise. Therefore, adaptive and robust methods of machine learning are essential.
Learning Vector Quantization (LVQ) [Kohonen (2001)] is a supervised learning and prototype-vector based classification method which adapts a piecewice linear discriminant function using a relative simple learning rule due to the principle of competitive learning. Activation of neurons is based on distance measures and therefore depends on metrics used. A known disadvantage of LVQ is its high dependence on initialization of the weight matrix [Song and Lee (1996)] which can be decreased by an initial unsupervised phase of training [Golz et al. (1998)]. [Sato (1999)] developed a modification, the so-called Generalized LVQ to decrease variance due to initializations. Other developments are LVQ methods which iteratively adapt a feature weighting during training to improve results and to give back a feature relevance measure. Here we used three representatives, the Distinctive Selection LVQ (DSLVQ), the Relevance LVQ (RLVQ) and the Generalized Relevance LVQ (GRLVQ) (for references we refer to [Hammer and Villmann (2002)].
The Support Vector Machine (SVM) [Vapnik (1995)] is also a supervised learning method and is more computationally expensive than LVQ. In its basic version, SVM can only adapt to linearly separable two-class problems. Advantageously, training is restricted to search for only those input vectors which are crucial for classification. They are called support vectors and are found by solving a quadratic optimization problem. For real world applications the soft-margin SVM [Cortes and Vapnik (1995)] is commonly used which allows a restricted number of training set errors. Another advantage of SVM in comparison to many other classification methods is the uniqueness of the solution found and the resulting independence on initialization and on training sequencing. Important parameters are the slack variable and the type and parameters of the kernel function. Disadvantages of SVM like the relative large memory allocation during training and the relative slow convergence can be removed by optimization of the training algorithm [Joachims (2002)]. This is essential to apply SVM to larger sized problems.

2 Performance Measurement

The performance of a classification algorithm is generally problem dependent. The ability of generalization is a measure of expected correct classifications of unknown patterns of the same underlying distribution function as of the training set. It can be estimated empirically by calculation of test set error rate. Here, we utilize two cross validation method, the "multiple hold-out" (MHO) and the "leave-one-out" (LOO) method [Devroye et al. (1996)]. Both methods require a learning set (training + test set) of statistically independent feature vectors. This is e.g. violated in time series processing when using

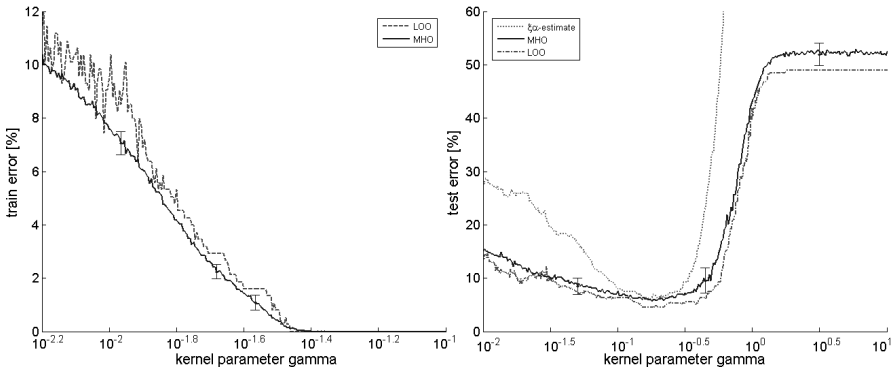

Fig. 1. Semilogarithmic plot of mean training (left) and mean test (right) errors of SVM vs. parameter gamma of Gaussian kernel function applied to posturography data. Estimates of LOO method are shown by left upper graph and by right lower graph, estimates of MHO method are shown by graph with errorbars, and $\xi\alpha$-estimate by right, upper graph. The regularization parameter of C = 10 was separately found to be sufficient.

overlapping segmentation; otherwise too optimistic estimates are resulting.
The acquisition of statistically independent patterns is expensive. In biomedical problems this process often requires an independent scoring process mostly done by experts and requires experimental and organisational effort. As a consequence, relative small data sets on small groups of test subjects are mostly available. Processing of those data sets should be as efficient as possible under the restriction of computational resources [Joachims (2002)]. The MHO validation consumes less computational time than the LOO method. The first method has the ratio of sizes of test and training set as a free selectable parameter for which upper and lower bounds are estimable [Kearns (1996)]. After repeating N times the random partition in test and training set following up by single hold-out estimation one can conclude estimates of adaptivity and ability of generalization by descriptive statistics. We calculate mean and standard deviations of training and test errors. Disadvantageously MHO is biased, because of the limited hypothesis space [Joachims (2002)]. This limitation is minimal in case of LOO because the size of the training set is reduced by only one feature vector. Therefore, this method supplies an almost unbiased estimate of the true classification error. In the special case of the SVM classificator the $\xi\alpha$-estimate was proposed [Joachims (2002)]. This estimator avoids high computational effort.
There is no common criterium for the choice of kernel function [Müller et al. (2001)]. Each function type has few parameters which can be defined empirically. Mostly this is done by variation of parameters and calculation of classification errors or the VC-dimension [Van Gestel et al. (2002), Joachims (2002)]. The slack variable is determined in the same manner. For our data sets we have tested the linear, the polynomial and the Gaussian kernel func-

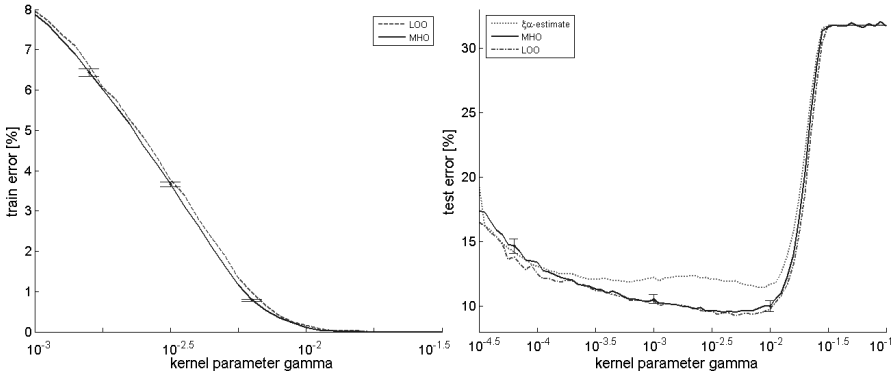

Fig. 2. Semilogarithmic plot of mean training (left) and mean test (right) errors of SVM vs. parameter gamma of Gaussian kernel function applied to microsleep data. Estimates of LOO method are shown by left upper graph and by right lower graph, estimates of MHO method are shown by graph with errorbars, and $\xi\alpha$-estimate by right, upper graph. The regularization parameter of C = 1.5 was separately found to be sufficient.

tion. In the following we refer only to results of Gaussian kernel SVM because they performed best in all cases.
Variation of the parameter gamma, which predefines the influence region of single support vectors, shows even in a semilogarithmic plot a gradually decreasing test error which is abruptly increasing after the optimum (Fig. 1 right). Test errors are in case of SVM efficiently computable by LOO method and are mostly slightly lower than mean errors of MHO method. The same plot, but calculated for training errors (Fig. 1 left), shows an inverse result. Training errors of LOO method are mostly slightly higher than MHO results. The $\xi\alpha$-estimate shows a different dependence on gamma and is in the vicinity of both other estimate only in a small range of gamma. Therefore, the $\xi\alpha$-estimate should not be suitable for selection of parameters.
Results of the second data set (Fig. 2) are similar to the first, though the processes of data generation are fundamentally distinct. A difference is seen in optimal value of gamma and another in optimal value of mean test errors (Fig. 2 right). The optimal mean test error of the microsleep data set is about 9.8% and the standard deviation is clearly lower, which is argued by the clearly higher size of the data set. On this data set the $\xi\alpha$-method is resulting in the same optimal parameter gamma than both other estimations, but is estimating clearly higher errors.

3 Comparison of different classification methods

In the following we want to compare several variants of LVQ, SVM and other classification methods applied to both data sets. In addition to the originally

proposed variants, LVQ1, LVQ2.1, LVQ3, OLVQ1 [Kohonen (2001)], we used four further variants for relevance detection and feature weighting as mentioned above. Furthermore, some unsupervised learning methods which are calibrated by class labels after training. We compare well-known k-Means and Self-Organizing Map to a representative of incremental neural networks, the Growing Cell Structures [Fritzke (1994)]. All three methods find out a trade-off between a quantized adaptation of the probability density function and a minimization of the mean squared error of vector quantization. In all three unsupervised methods we tested also the modification "supervised" (sv) which is using the class label as a further component in input vectors of the training set [Kohonen (2001)]. The term is somewhat misleading because training remains unsupervised. Though this modification has only a small effect on distance calculations during training, the algorithm should be able to adapt better. Therefore, training errors are always lower than without modification "sv". The posturography data set (Tab. 1A) is very well adaptable reflecting in very low training errors, especially for supervised learning methods which perform nearly equally by mean errors of about 1% and lower. The ability of generalization is also nearly equal suggesting by mean test errors of about 4% which is unusually low for real world biosignals. The quickly converging method OLVQ1 arrives at same level than modern methods GR-LVQ and SVM. As expected, in (Tab. 1A) a large difference in test errors to unsupervised learning methods is evident. The modification "sv" allows the algorithm to find a more generalizable discriminant function.
The second and more complex data set (microsleep data) supplies different results (Tab. 1B). Training errors are much higher despite the exception of no errors of SVM. Unsupervised learning methods with modification "sv" perform better than all LVQ variants with respect to training errors. Among all LVQ variants OLVQ1 performs best. Two modified LVQ algorithms for relevance detection perform slightly worse, but better than standard LVQ. The higher complexity is also reflected in test errors. They are between 14% and 16% for all LVQ variants and are best for OLVQ1. Here, SVM shows lowest errors and the best ability to handle higher complexity. The relative improvement (ΔE / E) compared to LVQ variants is about 30%. As expected, unsupervised methods are not able to perform comparably. Interestingly, in case of microsleep data there is no difference in test errors between unsupervised methods with and without modification "sv". This modification shows better adaptivity in all cases shown by lower training errors (Tab. 1), but doesn't improve test errors in more complex data.

4 Conclusions

Both real world two-class problems have been solved with low error rates using prototype-vector based classification methods. The posturography data set has shown very good discriminability indicating high sensitivity of this

classification method	(A) posturography		(B) microsleep	
	E_{TRAIN} [%]	E_{TEST} [%]	E_{TRAIN} [%]	E_{TEST} [%]
LVQ1	0,8 ± 0,7	4,8 ± 2,7	11,9 ± 0,6	15,9 ± 1,2
LVQ2.1	0,8 ± 0,7	4,9 ± 2,6	11,8 ± 0,6	15,8 ± 1,3
LVQ3	0,8 ± 0,7	4,8 ± 2,6	11,9 ± 0,6	15,9 ± 1,2
OLVQ1	0,2 ± 0,3	4,2 ± 2,2	9,2 ± 0,4	13,9 ± 0,9
$OLVQ1_{LOO}$	0,3 ± 0,2	3,9 ± 1,7	9,5 ± 0,3	13,5 ± 0,5
RLVQ	1,2 ± 0,9	6,5 ± 2,8	12,5 ± 1,2	15,5 ± 1,8
DSLVQ	0,3 ± 0,3	4,7 ± 2,2	9,8 ± 0,5	15,1 ± 1,2
GLVQ	0,5 ± 0,2	5,9 ± 2,4	11,5 ± 0,5	15,9 ± 0,9
GRLVQ	0,3 ± 0,3	4,4 ± 2,3	9,6 ± 0,4	14,2 ± 0,5
SVM	0,0 ± 0,0	4,3 ± 2,4	0,0 ± 0,0	10,6 ± 0,4
SVM_{LOO}	0,0 ± 0,0	4,0 ± 0,0	0,0 ± 0,0	9,8 ± 0,0
kM	6,9 ± 1,7	16,8 ± 4,4	16,3 ± 1,2	16,8 ± 1,5
kM sv	0,3 ± 0,9	8,9 ± 3,3	6,3 ± 0,7	16,2 ± 1,4
SOM	8,9 ± 0,6	13,3 ± 1,7	13,2 ± 0,8	17,7 ± 1,6
SOM sv	1,4 ± 0,6	9,1 ± 3,4	8,6 ± 0,7	16,7 ± 1,3
GCS	2,9 ± 1,2	11,2 ± 5,0	12,9 ± 1,5	17,2 ± 2,1
GCS sv	0,3 ± 0,4	8,8 ± 4,4	5,5 ± 1,0	16,9 ± 1,9

Table 1. mean and standard deviations of test and training errors of different classification methods applied to posturography (A) and to microsleep data (B)

measurement technique to small and unknown changes. This result is achievable only by processing spectral domain features. As not reported here, we failed in achieving similar results using alternatively 23 time domain features which were reported of several authors in the posturography literature of the last two decades. As well as processing of all 23 features and as also processing some combinations of them did not lead to similar results as by spectral domain features. This indicates that no simple effects, like changes in amplitude histogram, but dynamical aspects of postural time series are influenced the effect of alcohol intake on posture. OLVQ1, SVM and the recently introduced GRLVQ perform best. The first method is the most simplest and fastest in convergence. Their iterative adaptation rule of step size during training seems to be the key point to outperform other adjacently associated methods, like LVQ1.

In a more complex data set (microsleep) which has much more feature vectors and higher dimensionality than the posturography data set SVM outperforms all other methods. In contrast to all other methods SVM is not dependent on initializations and always finds out the global minimum of the error function [Müller et al. (2001)]. Utilizing LOO method to estimate the ability of generalization is computationally expensive but in case of SVM an efficient calculation using support vectors only can be used. The $\xi\alpha$-estimator is also an efficient method, but as our empirical results on both biomedical data sets indicate, this estimator seems to be biased. Therefore, SVM combined with

LOO validation exposes to be the most recommendable combination. Nevertheless, in some parameter settings the SVM combined with all three mentioned validation methods needs up to 100 times more computational effort than OLVQ1 combined with MHO validation. For extensive scanning of parameters in the whole processing cue, we therefore recommend to apply OLVQ1 / MHO and for subsequent fine tuning we recommend to apply SVM / LOO.

References

CAO, L.J. and TAY, F.E.H. (2003): Support Vector Machine With Adaptive Parameters in Financial Time Series Forecasting. *IEEE Transactions on Neural Networks, 14, 1506–1518.*

CORTES, C. and VAPNIK, V. (1995): Support Vector Networks. *Machine Learning, 20, 273–297.*

DEVROYE, L.; GYORFI, L.; LUGOSI, G. (1996): *A probabilistic theory of pattern recognition.* Springer; New York.

FRITZKE, B. (1994): Growing Cell Structures - A Self-Organizing Network for Unsupervised and Supervised Learning. *Neural Networks, 7, 1441–1460.*

GOLZ, M.; SOMMER, D.; LEMBCKE, T.; KURELLA, B. (1998): Classification of the pre-stimulus-EEG of k-complexes using competitive learning networks. *EUFIT ́98.* Aachen, 1767–1771.

GOLZ, M.; SOMMER, D.; WALTHER, L.; EURICH, C. (2004): Discriminance Analysis of Postural Sway Trajectories with Neural Networks *SCI2004, VII.* Orlando, USA, 151–155.

HAMMER, B. and VILLMANN, T. (2002): Generalized relevance learning vector quantization. *Neural Networks, 15, 1059–1068.*

JOACHIMS, T. (2002): *Learning to Classify Text Using Support Vector Machines.* Kluwer; Boston.

KEARNS, M. (1996): A Bound on the Error of Cross Validation Using the Approximation and Estimation Rates, with Consequences for the Training-Test Split. *Advances in Neural Information Processing Systems, 8, 183–189.*

KOHONEN, T. (2001): *Self-Organizing Maps (third edition).* Springer, NewYork.

MÜLLER, K.-R.; S. MIKA; RÄTSCH, G.; TSUDA, K.; SCHÖLKOPF, B. (2001): An Introduction to Kernel-Based Learning Algorithms. *IEEE Transactions on Neural Networks, 12(2), 181–201.*

OSUNA, E.; FREUND, R.; GIROSI, F.; (1997): Training Support Vector Machines: an Application to Face Detection. *Proceedings of CVPR ́97.* Puerto Rico.

SATO, A. (1999): An Analysis of Initial State Dependence in Generalized LVQ. In: D. Willshaw et al. (Eds.): *(ICANN ́99).* IEEE Press; , 928–933.

SOMMER, D. and GOLZ, M. (2003): Short-Time Prognosis of Microsleep Events by Artificial Neural Networks. *Proc. Medizin und Mobilität.* Berlin, 149–151.

SONG, H. and LEE, S. (1996): LVQ Combined with Simulated Annealing for Optimal Design of Large-set Reference Models. *Neural Networks, 9, 329–336.*

VAN GESTEL, T.; SUYKENS, J.; BAESENS, B.; VIAENE, S.; VANTHIENEN, J.; DEDENE, G.; DE MOOR, B.; VANDEWALLE, J. (2002): Benchmarking least squares support vector machine classifiers. *Machine Learning.*

VAPNIK, V. (1995): *The Nature of Statistical Learning Theory.* Springer, New York.