ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Optimization of General Statistical Accuracy Measures for Classification Based on Learning Vector Quantization

M. Kaden[1]*, W. Hermann[2], and T. Villmann[1]

1- University of Appl. Sciences Mittweida - Dept. of Mathematics
Mittweida, Saxonia - Germany

2- Paracelsus Hospital Zwickau - Dep. of Neurology
Zwickau, Germany

Abstract. We propose a framework for classification learning based on generalized learning vector quantization using statistical quality measures as cost function. Statistical measures like the F-measure or the Matthews correlation coefficient reflect better the performance for two-class classification problems than the simple accuracy, in particular if the data classes are imbalanced. For this purpose, we introduce soft approximations of those quantities contained in the confusion matrix, which are the basis for the calculation of the quality measures.

1 Introduction

Classification of data is one of the most frequent task in machine learning and statistical data analysis. Many methods and approaches were developed ranging from prototype based classifiers like Support Vector Machines (SVMs, [1]) or the family of Learning Vector Quantizers (LVQs, [2]) to classification trees. These approaches as well as classical statistical approaches like linear discriminant analysis (LDA) typically try to minimize the *classification error* or at least approximations thereof. Accordingly, classifiers are compared in most cases by the equivalent classification accuracy.

Yet, the performance evaluation of a classifier only based on the accuracy is not the full truth. In case of *imbalanced data* the classification accuracy might be very high although an underrepresented class is poorly recognized. This classifier learning problem frequently occurs in medicine, when only a few patient data are available in comparison to the number of data of volunteers [3, 4, 5]. For this reason, other statistical assessment measures like *precision* and *recall* are more favored. Both values are based on the appraisal of the confusion matrix (CM). The direct evaluation of the entries of the CM is also important if the different types of misclassifications (false negatives / false positives) cause different costs [6]. We denote this scenario as an *asymmetric* classification task (ACT).

Several classification quality indices based on confusion matrix are known in statistical data analysis emphasizing different aspect. Well-known are the F-measure, the χ^2-statistics or the Jaccard-Index [7]. Thus, a direct optimization of these quantities by a learning classifier model would be desirable and was recently proposed for the in F-measure [8, 9]. Yet, the underlying learning models neither allow an easy interpretation nor optimization of other indices of the confusion matrix.

In this paper we present an classifier approach for optimization of those statistical measures, which is based on the generalized learning vector quantization (GLVQ) model [10]. The GLVQ classifier is a cost function based modification of the intuitive learning vector quantization model introduced by KOHONEN [11]. Whereas the latter one

*M. Kaden is supported by a grant of the European Social Fund, Saxony (ESF).

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

heuristically approximates a Bayes-classifier, GLVQ takes an approximation of the classification accuracy as objective. Generally, LVQ models are easy to interpret according to its paradigm as prototype based classifiers. We modify the GLVQ approach in such a manner that arbitrary statistical evaluation measures based on the classification confusion matrix can be optimized. The main ingredient for this modification is the utilization of the recently proposed border sensitive learning in GLVQ [12, 13].

2 Classification Accuracy Maximization by GLVQ

A cost function based variant of LVQ was proposed by SATO&YAMADA (*Generalized* LVQ - GLVQ, [10]). For this model we suppose data vectors $\mathbf{v} \in V \subseteq \mathbb{R}^n$. The prototypes of the GLVQ model are the set $W = \{\mathbf{w}_k \in \mathbb{R}^n, k = 1 \dots M\}$. Each data vector $\mathbf{v}$ of the training data belongs to a class $x_\mathbf{v} \in \mathcal{C} = \{1, \dots, C\}$. The prototypes are labeled by $y_k \in \mathcal{C}$. Further $d^+(\mathbf{v}) = d(\mathbf{v}, \mathbf{w}^+)$ denotes the dissimilarity between the data vector $\mathbf{v}$ and the closest prototype $\mathbf{w}^+$ with the same class label $y_{\mathbf{w}^+} = x_\mathbf{v}$, and $d^-(\mathbf{v}) = d(\mathbf{v}, \mathbf{w}^-)$ is the dissimilarity degree for the best matching prototype $\mathbf{w}^-$ with a class label $y_{\mathbf{w}^-}$ different from $x_\mathbf{v}$. Whereas the original LVQ heuristically optimizes the Bayes decision [11], GLVQ maximizes the hypothesis margin $m(\mathbf{v}) = d^+(\mathbf{v}) - d^-(\mathbf{v})$ [14, 15]. The respective cost function minimized by GLVQ is

$$E_{GLVQ}(W) = \frac{1}{2} \sum_{\mathbf{v} \in V} f(\mu(\mathbf{v})) \tag{1}$$

where f is a monotonically increasing transfer or squashing function usually chosen as sigmoid or the identity function and

$$\mu(\mathbf{v}) = \frac{d^+(\mathbf{v}) - d^-(\mathbf{v})}{d^+(\mathbf{v}) + d^-(\mathbf{v})} \tag{2}$$

is the classifier function. We remark that $\mu(\mathbf{v}) \in [-1, 1]$. The dissimilarity measure $d(\mathbf{v}, \mathbf{w}_k)$ is not necessarily required to be a metric [16] but is assumed to be differentiable with respect to $\mathbf{w}_k$ for stochastic gradient learning.
Learning in GLVQ of $\mathbf{w}^+$ and $\mathbf{w}^-$ is usually performed by the *stochastic* gradient descent with respect to the cost function E_{GLVQ} for a given data vector $\mathbf{v}$. Recent approaches include relational and median learning [17, 18, 19].
The recall for a given data point $\mathbf{v}$ is realized via a winner take all rule: Let

$$s(\mathbf{v}) = argmin_{k=1}^{M} (d(\mathbf{v}, \mathbf{w}_k)) \tag{3}$$

be the index of the matching unit. The respective prototype label $y_{s(\mathbf{v})}$ is the predicted class of the classifier.

3 Classification Accuracy and Statistical Measures in GLVQ

We observe that the classifier function $\mu(\mathbf{v})$ from (2) becomes negative if the data point $\mathbf{v}$ is correctly classified, i.e. if $x_\mathbf{v} = y_{s(\mathbf{v})}$ is valid. Further, the transfer function f in (1) is frequently chosen as the sigmoid function

$$f_\theta(x) = \frac{1}{1 + \exp\left(-\frac{x}{\theta}\right)} \tag{4}$$

with the parameter θ determining the slope [20]. In the limit $\theta \to 0$ the sigmoid f_θ becomes the Heaviside function $H(x)$, such that the cost function E_{GLVQ} approximately counts the *misclassifications* in the GLVQ for this case. The respective variant

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

labels	true			
		C_+	C_-	
predicted	C_+	TP	FP	$\widehat{N}_+$
	C_-	FN	TN	$\widehat{N}_-$
		N_+	N_-	N

Table 1: Confusion matrix: TP - true positives, FP - false positives, TN - true negatives, FN - false negatives, $N_\pm$- number of positive/negative data, $\widehat{N}_+$ - number of predicted positive/negative data.

of GLVQ is known as border-sensitive GLVQ (BS-GLVQ,[13]). Hence, $E_{BS-GLVQ}$ is implicitly based on the classification accuracy evaluation.
Yet, accuracy is not always an appropriate to evaluate a classifier, in particular, if the data are imbalanced [7]. For example, assigning each object to the larger class achieves a high proportion of correct predictions, but is not a useful classification at all. In statistical analysis contingency table evaluations are well-known to deal with this problem more properly. In case of two-class problems with classes C_+ and C_- the table contains the confusion matrix Tab. 3.
Several measures were developed to judge the classification quality based on the confusion matrix emphasizing different aspects. The quantities *precision* π and *recall* ρ defined by

$$\pi = \frac{TP}{TP + FP} = \frac{TP}{\widehat{N}_+} \text{ and } \rho = \frac{TP}{TP + FN} = \frac{TP}{N_+} \tag{5}$$

respectively, are used in the widely applied F_β-measure

$$F_\beta = \frac{\left(1 + \beta^2\right) \cdot \pi \cdot \rho}{\beta^2 \cdot \pi + \rho} \tag{6}$$

developed by C.J. VAN RIJSBERGEN [21]. For the common choice $\beta = 1$ it is the fraction of the harmonic and the arithmetic mean of precision and recall, i.e. β controls the influence of both values. Further, the *specificity* ς and the *negative prediction rate*

$$\varsigma = \frac{TN}{TN + FP} = \frac{TN}{N_-} \text{ and } \xi = \frac{TN}{TN + FN} = \frac{TN}{\widehat{N}_-} \tag{7}$$

are frequently considered in medical applications. These values can be combined in the *weighted accuracy*

$$wAC_\Sigma = \alpha_1 \rho + \alpha_2 \pi + \alpha_3 \varsigma + \alpha_4 \xi$$

with the signature $\Sigma = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$. Another measure considering all four quantities of the confusion matrix is the *Matthews correlation coefficient*

$$MMC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}, \tag{8}$$

which is equivalent to the χ^2-statistics for a 2×2 contingency table [22, 7].
In the following we propose a framework to integrate them into the GLVQ. The idea behind is to keep the basic ingredients of GLVQ, which are *prototype based classification*, *gradient descent learning*, and the *dissimilarity based classifier function* $\mu(\mathbf{v})$.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

At this point we restrict ourselves to the two-class scenario $\{C_+, C_-\}$ of a positive class C_+ with class label '$\oplus$' and a negative class C_- with class label '$\ominus$'. Following the observation that the transfer function f_θ (4) approximates the Heaviside function H, we consider a modified classifier function $\hat{\mu}(\mathbf{v}) = f_\theta(-\mu(\mathbf{v}))$ with $\hat{\mu}(\mathbf{v}) \approx 1$ iff the data point $\mathbf{v}$ is correctly classified and $\hat{\mu}(\mathbf{v}) \approx 0$ otherwise. Now we can express all quantities of the confusion matrix in terms of the new classifier function $\hat{\mu}(\mathbf{v})$:

$$TP = \sum_{j=1}^{N} \delta_{\oplus, x_{\mathbf{v}_j}} \cdot \hat{\mu}(\mathbf{v}_j), \quad FP = \sum_{j=1}^{N} \delta_{\ominus, x_{\mathbf{v}_j}} \cdot (1 - \hat{\mu}(\mathbf{v}_j))$$

$$FN = \sum_{j=1}^{N} \delta_{\oplus, x_{\mathbf{v}_j}} \cdot (1 - \hat{\mu}(\mathbf{v}_j)) \text{ and } TN = \sum_{j=1}^{N} \delta_{\ominus, x_{\mathbf{v}_j}} \cdot \hat{\mu}(\mathbf{v}_j)$$

with $\delta_{\oplus, x_{\mathbf{v}_j}}$ is the Kronecker symbol and $\delta_{\ominus, x_{\mathbf{v}_j}} = 1 - \delta_{\oplus, x_{\mathbf{v}_j}}$. Obviously, all these quantities are differentiable with respect to $\hat{\mu}(\mathbf{v}_j)$ and, hence, also with respect to the prototypes $\mathbf{w}_k$. Now, we suppose a general statistical measure $S(TP, FP, FN, TN)$ to be minimized, which is *continuous and differentiable* with respect to TP, FP, FN, and TN. Thus, it is also differentiable with respect to the prototypes via the chain rule for differentiation, e.g. $\frac{\partial S}{\partial \mathbf{w}_k} = \frac{\partial S}{\partial TN} \cdot \frac{\partial TN}{\partial \mathbf{w}_k}$. Therefor, they can easily plugged into GLVQ serving as a new cost function and not violating the stochastic gradient learning. Hence, the GLVQ can be used in a statistical framework. Clearly, the above mentioned measures F_β, MMC, and wAC_Σ belong to this function class and, therefore, can be plugged into the GLVQ scheme.

4 Simulations

Due to the lack of space, we only report here the results for one real world data set. Other simulation results can be found in [23]. We consider a dataset of neurophysiological data of Wilson disease (WD) patients and probands. WD is an autosomal-recessive disorder copper metabolism in the liver such that suffering patients develop neurophysiological impairments. Thus, in the initial non-neurologic phase, impairments are negligible or at least not defacing, whereas later on (neurologic phase) the disturbances become severe [24]. Yet, there is a smooth transition between both phases. To judge the neurological impairments a ^{18}*F-Fluorodesoxyglucose-Positron-Emission-Tomography* ([^{18}F]FDG-PET,[3]) was applied delivering a neurological impairment profile for each patient/proband. It consists of a 11-dimensional vector with the normalized glucose consumption in different brain regions (*frontal lobe,parietal lobe, temporal lobe, occipital lobe, ant. cingulum, post cingulum, putamen, caput nuclei caudati, cerebellum, midbrain, thalamic area*). A detailed description can be found in [25]. Additionally, a clinical diagnosis suggests an assignment to the neurologic/non-neurologic type [3]. We used this dataset to learn the classification decision based on the neurophysiological impairment profile. The dataset contains 15 proband samples 16 non-neurologic and 34 neurologic samples (N). Probands and non-neurologic patients form the non-neurologic group (NN). All results are obtained from 8-fold cross-validation.

First, we conducted standard GLVQ learning as a baseline for comparison. Thereafter, we applied the modified GLVQ with the F_β from (6) for several β-values, see Tab. (2). For all experiments we calculated also F_β-values. We observe the expected behavior, i.e. the best F_β-values are achieved if the respective cost function was optimized. Moreover, standard GLVQ does not yield as good F_β-results as the modified GLVQ. Further,

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

		GLVQ		F_β-GLVQ		F_β-GLVQ		F_β-GLVQ	
				$\beta^2 = 0.5$		$\beta^2 = 1$		$\beta^2 = 2$	
confusion matrix		true		true		true		true	
		N	NN	N	NN	N	NN	N	NN
~~prediction~~	N	90.9%	28.0%	88.5%	7.2%	93.6%	11.1%	96.7%	15.4%
	NN	9.1%	72.0%	11.5%	92.8%	6.4%	88.9%	3.2%	84.6%
F_β-measure ($\beta^2 = 0.5$)		0.790		**0.907**		0.901		0.887	
F_β-measure ($\beta^2 = 1$)		0.816		0.902		**0.910**		0.906	
F_β-measure ($\beta^2 = 2$)		0.845		0.896		0.918		**0.926**	
precision		0.741		**0.918**		0.885		0.852	
recall		0.909		0.885		0.936		**0.968**	

Table 2: Classification results for the Wilsons disease data set for different types of GLVQ using one prototype per class.

it is well-known that F_β emphasizes high precision for small β-values, whereas recall is more weighted with increasing β-values. This behavior is nicely observable for the experiments. Thus, the experiments show the ability of the modified GLVQ to optimize statistical measure based on the confusion matrix entries.
From a medical point of view, the simulations show that the neurologic phase in WD can clearly distinguished from the neurologic state only considering the [^{18}F]FDG-PET profiles.

5 Conclusion

In the present paper we propose a modified GLVQ using statistical measures for the underlying cost function. The statistical measures are assumed to be continuously depending on the entries of the confusion matrix and differentiable. Then, the key idea is to use the smooth approximations of the quantities of the confusion matrix when the statistical measure is taken to replace the original accuracy based cost function in GLVQ. In this way, the basic principles of GLVQ-like prototype based classification and gradient descent learning are kept.
Thus, the new approach is an alternative to recently proposed classifier systems based on SVM and multilayer perceptron optimizing the F_1-objective [26, 9]. Further, the general formulation allows the utilization of other statistical measures like specificity, precision or recall, to reflect different aspects in classification learning, which are important for imbalanced class data and asymmetric classification tasks.
We presented the framework for a two-class scenario so far. Extensions to more classes could be greedy strategies like hierarchical or weighted one-versus-all classification schemes as suggested in [27]. This, however, remains topic for future research as well as the integration of such statistical measurements into fuzzy classification schemes.

References

[1] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, 2002.
[2] Teuvo Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995. (Second Extended Edition 1997).
[3] W. Hermann, H. Barthel, S. Hesse, F. Grahmann, H.-J. Kühn, A. Wagner, and Th. Villmann. Comparison of clinical types of Wilson's disease and glucose metabolism in extrapyramidal motor brain regions. *Journal of Neurology*, 249(7):896–901, 2002.
[4] W. Hermann, P. Günther, A. Wagner, and T. Villmann. Klassifikation des Morbus Wilson auf der Basis neurophysiologischer Parameter. *Der Nervenarzt*, 76:733–739, 2005.
[5] T. Villmann, G. Blaser, A. Körner, and C. Albani. Relevanzlernen und statistische Diskriminanzverfahren zur ICD-10 Klassifizierung von SCL90-Patienten-Profilen bei Therapiebeginn. In G. Plöt-

tner, editor, *Aktuelle Entwicklungen in der Psychotherapieforschung*, pages 99–118. Leipziger Universitätsverlag, Leipzig, Germany, 2004.

[6] M. Kästner, W. Hermann, and T. Villmann. Integration of structural expert knowledge about classes for classification using the fuzzy supervised neural gas. In M. Verleysen, editor, *Proc. of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2012)*, pages 209–214, Louvain-La-Neuve, Belgium, 2012. i6doc.com.

[7] L. Sachs. *Angewandte Statistik*. Springer Verlag, 7-th edition, 1992.

[8] K. Dembczyński, W. Waegeman, W. Cheng, and E. Hüllermeier. An exact algorithm for F-measure maximization. In *Proc. of the Conference on Neural Information Processing Systems (NIPS)*, 2011.

[9] J. Pastor-Pellicer, F. Zamora-Martínez, S. España-Boquera, and M.J. Castro-Bleda. F-measure as the error function to train neural networks. In I. Rojas, G. Joya, and J. Gabestany, editors, *Advances in Computational Intelligence - 12th International Work-Conference on Artificial Neural Networks, IWANN, Puerto de la Cruz, Tenerife, Spain*, volume 1 of *LNCS*, pages 376–384, 2013.

[10] A. Sato and K. Yamada. Generalized learning vector quantization. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference*, pages 423–9. MIT Press, Cambridge, MA, USA, 1996.

[11] Teuvo Kohonen. Learning Vector Quantization. *Neural Networks*, 1(Supplement 1):303, 1988.

[12] M. Kästner, M. Riedel, M. Strickert, W. Hermann, and T. Villmann. Border-sensitive learning in kernelized learning vector quantization. In I. Rojas, G. Joya, and J. Cabestany, editors, *Proc. of the 12th International Workshop on Artificial Neural Networks (IWANN)*, volume 7902 of *LNCS*, pages 357–366, Berlin, 2013. Springer.

[13] M. Kästner, M. Riedel, M. Strickert, and T. Villmann. Class border sensitive generalized learning vector quantization - an alternative to support vector machines. *Machine Learning Reports*, 6(MLR-04-2012):40–56, 2012. ISSN:1865-3960, http://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_04_2012.pdf.

[14] M. Biehl, B. Hammer, P. Schneider, and T. Villmann. Metric learning for prototype-based classification. In M. Bianchini, M. Maggini, F. Scarselli, and L.C. Jain, editors, *Innovations in Neural Information Paradigms and Applications*, volume 247 of *Studies in Computational Intelligence*, pages 183–199. Springer, Berlin, 2009.

[15] B. Hammer, M. Strickert, and Th. Villmann. Relevance LVQ versus SVM. In L. Rutkowski, J. Siekmann, R. Tadeusiewicz, and L.A. Zadeh, editors, *Artificial Intelligence and Soft Computing (ICAISC 2004)*, Lecture Notes in Artificial Intelligence 3070, pages 592–597. Springer Verlag, Berlin-Heidelberg, 2004.

[16] E. Pekalska and R.P.W. Duin. *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications*. World Scientific, 2006.

[17] B. Hammer, D. Hofmann, F.-M. Schleif, and X. Zhu. Learning vector quantization for (dis-)similarities. *Neurocomputing*, page in press, 2013.

[18] D. Nebel and T. Villmann. A median variant of generalized learning vector quantization. In M. Lee, A. Hirose, Z.-G. Hou, and R.M. Kil, editors, *Proceedings of International Conference on Neural Information Processing (ICONIP)*, volume II of *LNCS*, pages 19–26, Berlin, 2013. Springer-Verlag.

[19] X. Zhu, F.-M. Schleif, and B. Hammer. Semi-supervised vector quantization for proximity data. In M. Verleysen, ed., *Proc. of Europ. Symp. on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2013)*, pages 89–94, Louvain-La-Neuve, Belgium, 2013. i6doc.com.

[20] A.W. Witoelar, A. Gosh, J.J. de Vries, B. Hammer, and M. Biehl. Window-based example selection in learning vector quantization. *Neural Computation*, 22(11):2924–2961, 2010.

[21] C.J. Rijsbergen. *Information Retrieval*. Butterworths, London, 2nd edition edition, 1979.

[22] B.W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta*, 405:442–451, 1975.

[23] M. Kaden, M. Lange, D. Nebel, M. Riedel, T. Geweniger, and T. Villmann. Aspects in classification learning - Review of recent developments in Learning Vector Quantization. *Foundations of Computing and Decision Sciences*, page accepted, 2014.

[24] W. Hermann, Th. Villmann, and A. Wagner. Elektrophysiologisches Schädigungsprofil von Patienten mit einem Morbus Wilson'. *Der Nervenarzt*, 74(10):881–887, 2003.

[25] H. Barthel, Th. Villmann, W. Hermann, S. Hesse, H.-J. Kühn, A. Wagner, and R. Kluge. Different patterns of brain glucose consumption in Wilsons disease. *Zeitschr. f. Gastroenterologie*, 39:241, 2001.

[26] D.R. Musicant, V. Kumar, and A. Ozgur. Optimizing F-measure with support vector machines. In *PROCEEDINGS OF THE SIXTEENTH INTERNATIONAL FLORIDA ARTIFICIAL INTELLIGENCE RESEARCH SOCIETY CONFERENCE*, pages 356–360. Haller AAAI Press, 2003.

[27] P. Baldi, S. Brunak, Y. Chauvin, and C. Andersen H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, 2000.