ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

Median-LVQ for Classification of Dissimilarity Data based on ROC-Optimization

D. Nebel* and T. Villmann

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

Abstract. In this article we consider a median variant of the learning vector quantization (LVQ) classifier for classification of dissimilarity data. However, beside the median aspect, we propose to optimize the receiver-operating characteristics (ROC) instead of the classification accuracy. In particular, we present a probabilistic LVQ model with an adaptation scheme based on a generalized Expectation-Maximization-procedure, which allows a maximization of the area under the ROC-curve for those dissimilarity data. The basic idea behind is the utilization of ordered pairs as a structured input for learning. The new scheme can be seen as a supplement to the recently introduced LVQ-scheme for ROC-optimization of vector data.

1 Introduction

Learning vector quantization (LVQ) as introduced by T. KOHONEN is a popular approach for classification of vector data [1]. The basic idea of this approach is to represent the data classes by prototype vectors. Many variants of the basic Hebbian learning scheme were developed, an up-to-date overview can be found in [2]. Yet, the main ingredients, the optimization of the classification accuracy as well as the differentiable data dissimilarity measure in data space for the stochastic gradient method were kept in most of these variants.

Recently, the focus was shifted to more advanced classification goals like optimization of sensitivity, specificity or the F_β-measure, which are based on the evaluation of the confusion matrix. These statistical quality measures are more adequate for class-imbalanced training data [3]. Sensitivity and F_β-measure are closely related to the Receiver-Operating-Characteristics (ROC) [4], which is an important tool for performance comparison of binary classifiers [5]. Moreover, ROC-curves allow an user specific configuration of the classifier in dependence on required values of sensitivity etc. ROC-curves are usually compared by their area under the curve (AUROC), which should be maximum [6, 7]. LVQ-like optimization of the area under the ROC-curve (AUROC) for vector data was recently proposed [8], but it is still depending on the differentiability of the underlying dissimilarity measure.

Thus, the topic of LVQ-extensions for classification of dissimilarity or relational data emerges, as such variants are already known for unsupervised vector quantization [9, 10, 11, 12]. For relational approaches, the prototypes may be assumed as linear combination of the data such that gradient methods can be adapted [13]. For general dissimilarity data prototypes are restricted to be data samples. The latter strategy is known as median-learning. First attempts for median LVQ-variants optimizing the classification accuracy were provided in [14, 15]. In the present publication, we extend these ideas to the AUROC optimization based on the LVQ-paradigm.

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

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

2 Notations and Basic Concepts

In the following we suppose data objects $\mathbb{X} = \{x_i\}_{i=1,\ldots,N}$. We assume a binary classification problem with the classes $C = \{\oplus, \ominus\}$. Let $c(\cdot)$ be the formal class label function, which assigns to each data object the class label $y_i = c(x_i)$. The matrix D with entries $d_{ij} \geq 0$ provides the object dissimilarities. Furthermore, we define the set $X = \{(x_i, x_j) | y_i = \oplus \wedge y_j = \ominus\}$ of all *ordered pairs* of data objects generated from $\mathbb{X}$ with the *cardinality* denoted by $|X|$.

The ROC was developed as a graphical tool for comparison of binary classifiers with respect to their performance [5]. These performances are measured in terms of the true positive rate (recall/sensitivity) $\rho = \frac{TP}{TP+FN}$ and the false positive rate $\varphi = \frac{FP}{FP+TN}$ calculated according to the confusion matrix 2.

		true		
		$\oplus$	$\ominus$	
predicted	$\oplus$	TP	FP	$\tilde{N}_+$
	$\ominus$	FN	TN	$\tilde{N}_-$
		N_+	N_-	N

Table 1: Confusion matrix

If a parametrized classifier is considered, the resulting pairs of these values may be plotted into two-dimensional diagram - the so-called ROC-curve. The area A_{ROC} under this ROC-curve (AUROC) is a performance measure for this parametrized classifier. The higher the AUROC-value, the better the classifier. Assuming a binary classifier with a *classifier function* $\mu(\kappa|x_i)$ to predict the class κ for a sample x_i. Then the area A_{ROC} has a probabilistic interpretation:

$$A_{ROC} = P(\mu(\oplus|x_i) > \mu(\oplus|x_j)) \tag{1}$$

for a randomly chosen *ordered* pair $(x_i, x_j) \in X$ [5], which yields due to the underlying rank statistics [16, 17]. If we define an so-called *ordering function*

$$O(x_i, x_j) = H(\mu(\oplus|x_i) - \mu(\oplus|x_j)) \tag{2}$$

where H is the Heaviside function

$$H(x) = \begin{cases} 0 & if \quad x \leq 0 \\ 1 & else \end{cases}, \tag{3}$$

the probability P in (1) can be estimated by

$$\hat{A}_{ROC} = \frac{1}{|X|} \sum_{(x_i,x_j)\in X} O(x_i, x_j) \tag{4}$$

as proposed in [8].

3 AUROC-Optimizing Median LVQ Classifier

In the following we develop a binary classifier for AUROC-optimization based on the LVQ-prototype principle. However, instead of gradient learning we will develop a *generalized Expectation-Maximization* optimization scheme (gEM) based on a probabilistic interpretation of LVQ-prototypes. Thus we suppose M prototypes $\theta_k \in \Theta$, i.e. the

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

cardinality of Θ is M. Analogously as for data, $c_j = c(\theta_j)$ returns the predefined class label of the prototype. Further, M^+ denotes the number of prototypes assigned to the class $\oplus$. Because of the median paradigm, the prototypes are restricted to be data points itself. Based on the median LVQ-variant for accuracy maximization developed in [14], the classifier function can be defined as

$$\mu(\kappa|x_i) = \frac{d_\kappa(x_i) - d^-(x_i)}{d_\kappa(x_i) + d^-(x_i)} \tag{5}$$

where $d_\kappa(x_i)$ denotes the smallest dissimilarity of all prototypes corresponding to class κ and $d^-(x_i)$ is the smallest dissimilarity to all prototypes assigned to the opposite class. We introduce

$$O_\sigma(x_i, x_j) = f_\sigma(\mu(\oplus|x_i) - \mu(\oplus|x_j)) \tag{6}$$

as a smooth approximation of ordering function $O(x_i, x_j)$ from (2) with $f_\sigma(z) = 1/(1 + z/\sigma)$ being a sigmoid function. The parameter σ determines the slope. In the limit $\sigma \searrow 0$ we get $O_\sigma \to O$. For the probabilistic model to be developed in the next steps, we define positive, bounded functions $g((x_i, x_j), \Theta) = O_\sigma(x_i, x_j) + \varepsilon$ paying attention to the circumstance of ordered pairs required for AUROC-optimization. The small constant $\varepsilon > 0$ avoids numerical instabilities in the following. We define

$$K(\mathbb{X}) = \ln\left(\sum_{i,j} g((x_i, x_j), \Theta)\right) \tag{7}$$

as the logarithmic cost function (LCF) to be maximized instead of (1). This trick allows to apply a generalized Expectation-Maximization-scheme for optimization as we will explain in the following. For this purpose, we introduce the formal probability

$$p((x_i, x_j)) = \frac{g((x_i, x_j), \Theta)}{\sum_{i,j} g((x_i, x_j), \Theta)}$$

for an ordered data object pair (x_i, x_j). Thus, assuming arbitrary non-negative real numbers $\gamma_{i,j}$ fulfilling the restriction $\sum_{i,j} \gamma_{i,j} = 1$, which can be also interpreted as formal probability values, we can decompose the LCF $K(\mathbb{X})$ into

$$K(\mathbb{X}) = \mathcal{L}(\gamma, \Theta) + \mathcal{K}(\gamma||p) \tag{8}$$

with the formal Kullback-Leibler-divergence (KLD,[18]) and the loss term

$$\mathcal{L}(\gamma, \Theta) = \sum_{i,j} \gamma_{i,j} \ln\left(\frac{g((x_i, x_j), \Theta)}{\gamma_{i,j}}\right) \tag{9}$$

as shown in [19]. Hence, $\mathcal{L}(\gamma, \Theta)$ is a lower bound for the LCF $K(\mathbb{X})$ due to the non-negativeness of the KLD $\mathcal{K}(\gamma||p)$. Using this property we obtain the following maximizing strategy for the LCF $K(\mathbb{X})$:

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

1. **Expectation-step (E-step)**: set

$$\begin{aligned} \gamma_{i,j} &:= p\left((x_i, x_j)\right) \\ &\Rightarrow \\ \mathcal{K}(\gamma||p) &= 0 \\ &\Rightarrow \\ K(\mathbb{X}) &= \mathcal{L}(\gamma, \Theta) \end{aligned}$$

   Note that the cost function value $K(\mathbb{X})$ does not change in this E-step, because $K(\mathbb{X})$ is independent from the parameters $\gamma_{i,j}$.

2. ***generalized* Maximization-step (gM-step)**: take the parameters $\gamma_{i,j}$ as fixed and find new prototypes Θ^{new}, such that:

$$\mathcal{L}(\gamma, \Theta^{new}) \geq \mathcal{L}(\gamma, \Theta^{old})$$

3. **Convergence criterion**: if $\Theta^{new} = \Theta^{old}$ stop. Else goto 1.

We remark that the new prototypes Θ^{new} maybe found by any search procedure. Thus it is not required to apply a gradient learning scheme. If we apply a sophisticated discrete search, with prototypes restricted to be selected from the data objects, a median-like optimization scheme is obtained. Further, because the new prototypes Θ^{new} have not to be maximizing the function $\mathcal{L}$ in each iteration step, it is not a precise maximization step and, therefore, we denote it as a generalized M-step (gM-step) and the overall procedure a *generalized* EM-optimization (gEM). Because of the lack of space, we refer to [19] for convergence details. We refer to our approach as *median ROC-LVQ* (mROC-LVQ).

4 Numerical Experiments

We tested the mROC-LVQ for several data sets. The results are compared with those obtained by median-LVQ (m-LVQ), which is described in [14]. The following datasets are investigated:

Insect This dataset is taken from [20]. There are 69 high-dimensional sequences of joint angles for stick insect locomotions. The whole-body kinematics was captured for two walking conditions: a straight walk (class A, 36% of the data), and a climbing task (class B, 64% of the data). The dissimilarity matrix was calculated according to a dynamic time warping procedure. For details we refer to [20].

Aural Sonar The data set is obtained from [21]. It consists of 100 sonar signals out of two classes (target of interest/clutter) with 50 samples for each. The dissimilarities between the signals were determined by an ad hoc classification of humans.

Voting This data set is also from [21]. The dataset contains 435 samples distinguished into 2 classes, representing categorical data records. Each record represents the votes of an U.S. House of Representatives Congressmen to 16 different problems. Thus, each data point includes 16 features, the voting results (yes or no). The goal is to classify democrats and republicans. The class ratio within the dataset is 267 to 168. The samples are compared based on the value difference metric.

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

Wisconsin Breast Cancer (WDBC) The dataset contains vectors computed from digitized images of fine needle aspirates (FNA) of breast masses [22]. The vectors describe characteristics of the cell nuclei present in the image. The data vectors are compared by the Euclidean distance. Overall, there are 212 malignant samples and 357 benign.
PIMA indians diabetes The UCI PIMA dataset collects records of 268 diabetic and 500 non-diabetic humans [22]. The record attributes are : 1. Number of times pregnant 2. Plasma glucose concentration after 2 hours in an oral glucose tolerance test 3. Diastolic blood pressure (mm Hg) 4. Triceps skin fold thickness (mm) 5. 2-hour serum insulin (mu U/ml) 6. Body mass index (weight in kg/(height in m)^2) 7. Diabetes pedigree function 8. Age (years). The dissimilarity between the data records is estimated by the Euclidean distance.
For each of the above datasets Π the following testing scenario was applied. The dataset was 100 times partitioned randomly into three parts of equal size. For each partition Π_k the obtained splits $\pi_{k,i}$, $i = 1 \ldots 3$, have itself the same class distribution as the whole dataset. A simulation run for a partition Π_k takes the three splits as training set, test set and validation set, respectively. Thus we get 6 possible combinations. For each combination several classifier runs with varying smoothing parameter σ for f_σ are trained and tested by the respective sets. The best classifier according to test set is chosen and validated using the validation set. The performance for the validation set yields the result for the considered configuration. In summary, we obtain 600 results for a selected dataset Π, the average of which gives the final performance $\mu_{alg}(\Pi)$ for the dataset Π and classifier alg. The performances are compared by the Welch-test with the zero-hypothesis that $\mu_{mROC-LVQ}(\Pi) \leq \mu_{m-LVQ}(\Pi)$ with error probability $\alpha = 0.01$. All simulations use only one prototype per class. The simulation results are collected in Tab.2.

dataset Π	$\mu_{m-LVQ}(\Pi)$	$\mu_{mROC-LVQ}(\Pi)$	significance
Insect	0.919 (0.0555)	0.926 (0.0520)	$*$
Pima	0.758 (0.0382)	0.751 (0.0445)	$-$
WDBC	0.964 (0.0100)	0.965 (0.0096)	$*$
Aural Sonar	0.881 (0.0633)	0.895 (0.0558)	$*$
Voting	0.978 (0.0133)	0.986 (0.0086)	$*$

Table 2: Averaged validation accuracies μ and variances for mROC-LVQ compared to m-LVQ together with significance level of the difference according to the Welch-test statistics with error probability $\alpha = 0.01$ ('$*$' - significant, '$-$' - non-significant).

We observe that direct optimization of the area A_{ROC} by mROC-LVQ outperforms m-LVQ in most cases, as expected. The only exception is the PIMA-dataset. However, this loss is not significant according to accompanied Welsh-test statistics.

5 Conclusion

In the present paper we developed a median variant of learning vector quantization to optimize the area under the curve of the receiver-operating-characteristics in classification tasks for dissimilarity data. The approach is a probabilistic model based on the LVQ-prototype principle. The optimization procedure follows a generalized EM-scheme adopted from median-LVQ. However, beside mathematical details, the most

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

important difference is the use of ordered data pairs and the ordering functions motivated by the probabilistic interpretation of the area under the ROC-curve. The paper outlines the mathematical justification. Further, several exemplary applications on well-known data sets verify the expected progress.

References

[1] Teuvo Kohonen. Learning vector quantization for pattern recognition. Report TKK-F-A601, Helsinki University of Technology, Espoo, Finland, 1986.

[2] 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*, 39(2):79–105, 2014.

[3] M. Kaden, W. Hermann, and T. Villmann. Optimization of general statistical accuracy measures for classification based on learning vector quantization. In M. Verleysen, editor, *Proc. of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2014)*, pages 47–52, Louvain-La-Neuve, Belgium, 2014. i6doc.com.

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

[5] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861–874, 2006.

[6] A.P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1149–1155, 1997.

[7] J. Huang and C. X. Ling. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3):299–310, 2005.

[8] M. Biehl, M. Kaden, P. Stürmer, and T. Villmann. ROC-optimization and statistical quality measures in learning vector quantization classifiers. *Machine Learning Reports*, 8(MLR-01-2014):23–34, 2014. ISSN:1865-3960, http://www.techfak.uni-bielefeld.de/˜fschleif/mlr/mlr_01_2014.pdf.

[9] M. Cottrell, B. Hammer, A. Hasenfuß, and T. Villmann. Batch and median neural gas. *Neural Networks*, 19:762–771, 2006.

[10] B. Hammer and A. Hasenfuss. Topographic mapping of large dissimilarity data sets. *Neural Computation*, 22(9):2229–2284, 2010.

[11] R.J. Hathaway, J.W. Davenport, and J.C. Bezdek. Relational duals of the c-means clustering algorithms. *Pattern recognition*, 22(3):205–212, 1989.

[12] R.J. Hathaway and J.C. Bezdek. NERF c-means: Non-Euclidean relational fuzzy clustering. *Pattern Recognition*, 27(3):429–437, 1994.

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

[14] 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.

[15] D. Nebel, B. Hammer, and T. Villmann. Supervised generative models for learning dissimilarity data. In M. Verleysen, editor, *Proc. of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2014)*, pages 35–40, Louvain-La-Neuve, Belgium, 2014. i6doc.com.

[16] H.B. Mann and D.R. Whitney. On a test whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statististics*, 18:50–60, 1947.

[17] F. Wilcoxon. Andividual comparisons by ranking methods. *Biometrics*, 1:80–83, 1945.

[18] S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[19] D. Nebel and T. Villmann. Median variants of LVQ for optimization of statistical quality measures for classification of dissimilarity data. *Machine Learning Reports*, 8(MLR-03-2014):1–25, 2014. ISSN:1865-3960, http://www.techfak.uni-bielefeld.de/˜fschleif/mlr/mlr_03_2014.pdf.

[20] F.-M. Schleif, B. Mokbel, A. Gisbrecht, L. Theunissen, V. Dürr, and B. Hammer. Learning relevant time points for time-series data in the life sciences. In *Artificial Neural Networks and Machine Learning–ICANN 2012*, pages 531–539. Springer, 2012.

[21] Y. Chen, E.K. Garcia, M.R. Gupta, A. Rahimi, and L. Cazzanti. Similarity-based classification: Concepts and algorithms. *The Journal of Machine Learning Research*, 10:747–776, 2009.

[22] C.L. Blake and C.J. Merz. UCI repository of machine learning databases. Irvine, CA: University of California, Dep. of Information and Computer Science, available at http://www.ics.edu/mlearn/MLRepository.html, 1998.