ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

Adaptive dissimilarity weighting for prototype-based classification optimizing mixtures of dissimilarities

M. Kaden, D. Nebel, and T. Villmann

University of Applied Sciences Mittweida, Computational Intelligence Group
Technikumplatz 17, 09648 Mittweida, Germany

Abstract. In this paper we propose an adaptive bilinear mixing of dissimilarities for better classification learning. In particular, we focus on prototype based learning like learning vector quantization. In this sense the learning of the mixture can be seen as a kind of dissimilarity learning as counterpart to dissimilarity selection in advance. We demonstrate this approach working for relational as well as median variants of prototype learning for proximity data.

1 Introduction

Data in machine learning are usually compared in terms of dissimilarities or similarities, which have to be properly selected by the user in advance. The concrete choice frequently is a crucial step, because this selection determines the behavior of the data processing model. For example, classification models may distinguish data according to a certain proximity measure whereas another does not display any differences of objects regarding the given classification task. One possibility to diminish this difficulty is to apply parametrized dissimilarity measures, which can be adapted in parallel during the classification learning. Respective approaches became attractive during the last years. In learning vector quantization (LVQ), as one of the most prominent prototype-based classification models, those parametrized approaches are well established starting with the relevance learning of the scaled Euclidean metric parametrized by relevance parameters for each data dimension [1], which was later extended to matrix relevance approach allowing the adaptation of an arbitrary bilinear combination of data dimensions [2]. Generally, those models are referred as adaptive schemes where the adaptation of the applied dissimilarity takes place to improve the classification performance.

If the data objects consist of heterogeneous components, a single dissimilarity measure might not be sufficient to describe the relations between the data. For this situation, in [3] a bilinear mixing of sub-dissimilarities was suggested to describe the data relations in the context of LVQ-learning, each of the sub-dissimilarities only be responsible for partial components of the data vectors. Further, the mixing of the sub-dissimilarities is also subject of optimization during classification learning, such that an adaptive combination scheme is obtained for both mixing and classification. It is assumed in the approach [3] that all sub-dissimilarities are differentiable and the dissimilarities can be calculated at time. Hence, an online stochastic gradient descent learning (SGDL) can be applied when a cost based classifier is used like the generalized LVQ (GLVQ,[4]).

If several dissimilarities are available for the data reflecting different properties of the data, then each of them can be used for a separate classifier learning.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

Afterwards a fusion strategy could be applied to combine the single classifiers. For this purpose, several strategies are known including boosting or ensemble learning. An overview related to prototype-based classification is given in [5].

In this work we combine both approaches for adaptive combination of dissimilarities in prototype based classification learning. Particularly, we take the mixing and weighting of proximities as a scheme for dissimilarity learning, i.e. determination of which dissimilarity or which combination of proximity is most adequate for the classification task to be learned. This can be interpreted as a kind of distance learning as proposed in [6, 7], which may be also used for dissimilarity selection and weighting with respect to the given task.

Unfortunately, the differentiability assumption for the dissimilarity measures as it is required for SGDL in [3] cannot always be made in applications. For example, for many tasks the data objects to be classified are only provided by dissimilarity or similarity matrices describing their relations. For those scenarios, relational and kernels methods are adequate in classification [8, 9]. However, for support vector machines (SVM) feature weighting seems to be difficult [10]. Particularly, the bilinear mixing of kernels, according to the bilinear mixing suggested in [3] for dissimilarities, may lead to an overall proximity measure not longer being a kernels as supposed for SVM.

We show in this paper, how to integrate bilinear mixing and weighting of dissimilarities in prototype-based classification learning, if only proximity data are provided. Thereby, we exemplify the approach using variants of GLVQ as an intuitive and powerful classifier model. However, a transfer to other LVQ-models like Robust Soft LVQ (RSLVQ,[11]), for example, is straight forward.

2 Combination of dissimilarities in GLVQ

The GLVQ is a modification of standard LVQ providing a cost function, which can be minimized by SGDL [4]. Assume N data objects $\mathbf{v}_j$, K prototypes $W = \{\mathbf{w}_1, \ldots, \mathbf{w}_K\}$ and M single dissimilarity measures $d_k(\mathbf{v}, \mathbf{w})$. Adopting the idea from [3] we merge them into the combined bilinear measure

$$\delta(\mathbf{v}, \mathbf{w}) = \mathbf{d}^T(\mathbf{v}, \mathbf{w}) \, \mathbf{\Lambda} \mathbf{d}(\mathbf{v}, \mathbf{w}) \tag{1}$$

where $\mathbf{\Lambda} = \mathbf{\Omega}^T \mathbf{\Omega}$ and $\mathbf{\Omega} \in \mathbb{R}^{l \times M}$ is the mixing matrix and $\mathbf{d}(\mathbf{v}, \mathbf{w}) = (d_1(\mathbf{v}, \mathbf{w}), \ldots, d_M(\mathbf{v}, \mathbf{w}))^T$ the dissimilarity vector. It becomes linear for diagonal $\mathbf{\Lambda}$, i.e. this corresponds to a linear combination of the dissimilarities. For the dissimilarity measure we only assume non-negativeness and reflexivity $d_k(\mathbf{v}, \mathbf{v}) = 0$ according to [12, 13]. Hence, $\delta(\mathbf{v}_i, \mathbf{v}_j)$ is a non-negative bilinear form with $\delta(\mathbf{v}, \mathbf{w}) = (\mathbf{\Omega} \cdot \mathbf{d}(\mathbf{v}, \mathbf{w}))^2$. Then the cost function of the respective GLVQ with combined dissimilarity (CD-GLVQ) becomes

$$E_{CD-GLVQ} = \sum_j f\left(\mu^\delta(\mathbf{v}_j)\right) \tag{2}$$

with f is a monotonically increasing squashing function and

$$\mu^\delta(\mathbf{v}_j) = \frac{\delta^+(\mathbf{v}_j) - \delta^-(\mathbf{v}_j)}{\delta^+(\mathbf{v}_j) + \delta^-(\mathbf{v}_j)} \tag{3}$$

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

is the classifier function. The quantity $\delta^+(\mathbf{v}_j)$ represents the overall dissimilarity of $\mathbf{v}_j$ to the most similar prototype $\mathbf{w}^+$ of the correct class and $\delta^-(\mathbf{v}_j)$ is defined analogously for the incorrect class. Accordingly, we have $\mathbf{d}^\pm(\mathbf{v})$ and $d_k^\pm(\mathbf{v})$. The squashing function frequently is supposed to be the sigmoid function $f_\theta(x) = 1/(1+\exp(-\theta x))$ such that for large values θ the Heaviside function is approximated [14].

If differentiability is assumed for single dissimilarities d_k, the SGDL for the prototypes according to (2) is obtained as

$$\Delta\mathbf{w}^\pm = \mp\alpha_W \cdot f' \cdot \xi^\pm \cdot \left\langle \mathbf{\Lambda} \cdot \mathbf{d}^\pm(\mathbf{v}) \,|\partial_{\mathbf{w}^\pm}\mathbf{d}^\pm(\mathbf{v})\right\rangle \tag{4}$$

for a given data object $\mathbf{v}$ with the scaling factors

$$\xi^\pm = \frac{2\cdot\delta^\mp(\mathbf{v}_j)}{\left(\delta^+(\mathbf{v}_j)+\delta^-(\mathbf{v}_j)\right)^2}, \text{ and } \partial_{\mathbf{w}^\pm}\mathbf{d}^\pm(\mathbf{v}) = \left(\frac{\partial d_1^\pm(\mathbf{v})}{\partial\mathbf{w}^\pm},\ldots,\frac{\partial d_M^\pm(\mathbf{v})}{\partial\mathbf{w}^\pm}\right)^T$$

is the vector of the derivatives of the single dissimilarities. Further, $\langle\cdot|\cdot\rangle$ in (4) denotes the Euclidean inner product. We can update the mixing matrix $\mathbf{\Omega}$ by

$$\Delta\Omega_{kl} = -\alpha_{\mathbf{\Omega}}\cdot f'\cdot\left(\xi^+\cdot d_l^+(\mathbf{v})\cdot\left[\mathbf{\Omega}\cdot\mathbf{d}^+(\mathbf{v})\right]_k - \xi^-\cdot d_l^-(\mathbf{v})\cdot\left[\mathbf{\Omega}\cdot\mathbf{d}^-(\mathbf{v})\right]_k\right) \tag{5}$$

for a given data object $\mathbf{v}$, realizing a SGDL for $\mathbf{\Omega}$ simultaneously applied to the prototype learning (4). Thereby, $[\cdot]_k$ denotes the kth component of a vector.

If only proximity data are available, and single dissimilarities d_k are assumed to be embeddable into the Pseudo-Euclidean space [12], stochastic gradient descent *relational* learning (SGDRL) has to replace the SGDL (4) for prototype learning [8], whereas the mixing matrix update remains unaffected. For relational methods it is supposed that the prototypes $\mathbf{w}_k$ are convex linear combinations of the data, i.e. $\mathbf{w}_l = \langle\gamma_l|\mathbf{v}_j\rangle$ with the non-negative coefficient vector $\gamma_l = (\gamma_{l1},\ldots,\gamma_{lN})^T$ and data dissimilarities $d_k(\mathbf{v}_i,\mathbf{v}_j)$ are given by the symmetric dissimilarity matrix $\mathbf{D}_k \in \mathbb{R}_+^{N\times N}$. Thus the SGDL of the prototypes $\mathbf{w}_l$ take place as a respective update of the coefficient vectors γ_l according to

$$\Delta\gamma^\pm \propto \mp f'\cdot\xi^\pm\cdot\left\langle\mathbf{\Lambda}\cdot\mathbf{d}^\pm(\mathbf{v}_i)\,|\partial_{\gamma^\pm}\mathbf{d}^\pm(\mathbf{v}_i)\right\rangle \tag{6}$$

where

$$\partial_{\gamma^\pm}\mathbf{d}\left(\mathbf{v}_i,\mathbf{w}^\pm\right) = \left(\frac{\partial d_1^\pm(\mathbf{v}_i)}{\partial\gamma^\pm},\ldots,\frac{\partial d_M^\pm(\mathbf{v}_i)}{\partial\gamma^\pm}\right)^T \tag{7}$$

is the vector of the formal derivatives. Particularly, we have

$$\frac{\partial d_k^\pm(\mathbf{v}_i)}{\partial\gamma^\pm} = \frac{\partial\left(\left[\mathbf{D}_k\cdot\gamma^\pm\right]_i - \left(\gamma^\pm\right)^T\cdot\mathbf{D}_k\cdot\gamma^\pm\right)}{\partial\gamma^\pm} \tag{8}$$

with the derivative is calculated according to

$$\frac{\partial\left(\left[\mathbf{D}_k\cdot\gamma_l\right]_i - \left(\gamma_l\right)^T\mathbf{D}_k\cdot\gamma_l\right)}{\partial\gamma_l} = [\![\mathbf{D}_k]\!]_i - \mathbf{D}_k\gamma_l \tag{9}$$

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

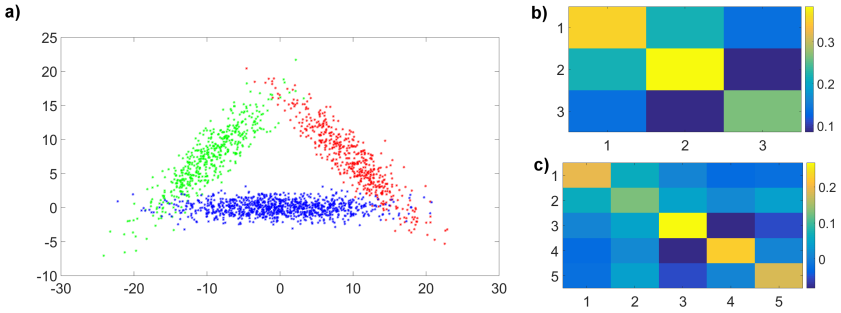

Figure 1: a) Triangle data set (AD) with the three classes ○, ∗ and ◇. b) Averaged mixing matrix of the AD data set learned using MLVQ. c) Averaged mixing matrix of the spectral data set learned using MLVQ.

as shown in [8], whereby $[\![\mathbf{D}_k]\!]_i$ denotes the ith column vector of $\mathbf{D}_k$.
If a subset or all single dissimilarities d_k are neither differentiable nor Pseudo-Euclidean embeddable, for example $\mathbf{D}_k$ is not symmetric, median variants like Median-GLVQ (MLVQ) come into play [15]. Median-GLVQ does not require Pseudo-Euclidean assumption on data and, hence, can be applied also in those cases. Doing so, the cost function $E_{CD-GLVQ}$ from (2) can be adapted by an alternating scheme as a greedy strategy: First, the prototypes are optimized by Median-GLVQ keeping the mixing matrix $\mathbf{\Omega}$ fixed. Afterwards, $E_{CD-GLVQ}$ is minimized with respect to the mixing matrix according to the SGDL given by (5) for fixed prototypes, which remains still applicable [16].

3 Experiments

We illustrate the approach by the application to two datasets. The first one is an artificial example whereas the second one is real world hyperspectral dataset for coffee classification. The artificial data set (AD) consists of two-dimensional data belonging to three overlapping classes arranged like a triangle, see Fig.1. The classes are generated by Gaussians with covariance matrices given as

$$\Sigma_1 = \begin{pmatrix} 20 & -18 \\ -18 & 20 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 50 & 0 \\ 0 & 1 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 20 & 18 \\ -18 & 20 \end{pmatrix}$$

with locations $\mu_1 = (10, 7)$, $\mu_2 = (0, 0)$, and $\mu_3 = (-10, 7)$. The number of data for the classes are $N_1 = 500$, $N_2 = 1000$, $N_3 = 500$. The distances used in this experiment are chosen as the class specific Mahalanobis-distances (d^{Σ_i}).

The coffee dataset (CD) consists of 256-band hyperspectra with equidistant bands in the range between 970nm and 2500nm already described in [17]. The data belong to 5 coffee types with 5000 spectra per class. The dissimilarities used here were the squared Euclidean distance (d^E), the squared Euclidean distance of the spatial derivatives as defined for Sobolev-distances (d^S), several γ-divergences (d^γ) including the Cauchy-Schwarz-divergence for ($\gamma = 1$).

For both dataset we conducted the following experiments: First we applied relational GLVQ (RLVQ,[8]) and MLVQ separately for each of the available

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

dissimilarities (separate runs). Thereafter, we applied again both approaches for a fixed linear combination (fLC) with equal weighting. Both experiments serve as a baseline. To demonstrate the capability of the proposed distance metric learning we applied in the next step an adaptive linear combination (aLC) and a adaptive bilinear mixing (bLM) using the full $\mathbf{\Omega}$-matrix. The aLC was used in combination with both, RLVQ and MLVQ, whereas bLM was applied only with MLVQ. The results for the experiments are given as averages over 15 runs.

The results of the experiments are collected in Tab.1. For the AD, the results of the separate runs were outperformed using the mixtures for MLVQ and RLVQ. Additionally, both adaptive variants show a further improvement in comparison to fLC with a small advantage in case of RLVQ. The dissimilarity weighting vectors are obtained as $\lambda_{MLVQ} = (0.406, 0.372, 0.222)^T$ and $\lambda_{RLVQ} = (0.359, 0.320, 0.321)^T$ for aLC. Thus, the dissimilarity d^{Σ_1} is indicated as most important for class discrimination within the both linear combination. Note, the accuracy of the respective separate run is not highest, i.e. this strong importance of the dissimilarity d^{Σ_2} only appears in combination with the others. The mixing matrix for bLM with MLVQ is depicted in Fig.1b). We observe that the off-diagonal values indicate the importance of bilinear dissimilarity mixing for further improvement.

For CD, the separate runs were again improved by the combined approaches for both MLVQ and RLVQ. However, for RLVQ using the d^S-dissimilarity, no further improvement is achieved using mixtures of dissimilarities. The weighting coefficients according to the dissimilarity vector $\left(d^E, d^S, d^\gamma_{0.25}, d^\gamma_{1.00}, d^\gamma_{1.50}\right)$ are resulted as $\lambda_{MLVQ} = (0.216, 0.238, 0.338, 0.093, 0.115)^T$ and $\lambda_{RLVQ} = (0.125, 0.558, 0.238, 0.013, 0.066)^T$. Both methods indicate the γ-divergences with $\gamma \geq 1$ as less important whereas the other ones are ranked differently. This may be dedicated to the greater flexibility of RLVQ compared to MLVQ. We observe that the separate run of RLVQ with d^S-dissimilarity yielded a similar performance as aLC and, particularly, also outperformed MLVQ, whereas MLVQ with bilinear mixing is comparable to combined RLVQ (the mixing matrix is depicted in Fig.1c)). Hence, MLVQ, which has strong restrictions regarding the prototypes, profits more from the flexibility offered by distance mixing learning than RLVQ. However, a direct interpretation of the weighting is at least difficult.

4 Conclusions

In this paper we describe a method for adaptive dissimilarity weighting in prototype based classification learning. The method is applicable for proximity data and shows good performance. We exemplified the methodology for GLVQ. However, it can be easily transferred to other LVQ-schemes like RSLVQ.

References

[1] B. Hammer and Th. Villmann. Generalized relevance learning vector quantization. *Neural Networks*, 15(8-9):1059–1068, 2002.

[2] P. Schneider, B. Hammer, and M. Biehl. Adaptive relevance matrices in learning vector quantization. *Neural Computation*, 21:3532–3561, 2009.

[3] E. Mwebaze, G. Bearda, M. Biehl, and D. Zühlke. Combining dissimilarity measures for prototype-based classification. In M. Verleysen, editor, *Proceedings of the European Sym-*

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

Triangle		d^{Σ_1}	d^{Σ_2}	d^{Σ_3}	fLC	aLC	bLM
MLVQ	acc	93.5	91.6	93.0	95.0	95.5	96.5
	($\pm std$)	($\pm$1.2)	($\pm$0.6)	($\pm$1.0)	($\pm$0.8)	($\pm$0.9)	($\pm$0.3)
RLVQ	acc	91.6	91.6	90.1	93.7	95.4	
	($\pm std$)	($\pm$1.2)	($\pm$0.6)	($\pm$1.6)	($\pm$0.1)	($\pm$0.4)	

Coffee		d^E	d^S	$d^\gamma_{0.25}$	$d^\gamma_{1.0}$	$d^\gamma_{1.5}$	fLC	aLC	bLM
MLVQ	acc	75.5	72.4	76.8	75.6	75.4	81.8	82.7	83.9
	($\pm std$)	(3.8)	(2.3)	(2.4)	(3.1)	(2.8)	(2.3)	(1.9)	(0.8)
RLVQ	acc	75.7	83.2	76.6	75.0	75.4	82.8	83.2	
	($\pm std$)	(4.0)	(1.8)	(2.8)	(2.5)	(0.2)	(2.2)		

Table 1: Averaged classification test accuracies in % with standard deviations (in %) for experiments described in the text.

posium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2015), pages 31–36, Louvain-La-Neuve, Belgium, 2015. i6doc.com.

[4] 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. Proc. of the 1995 Conf.*, p. 423–9. MIT Press, Cambridge, MA, USA, 1996.

[5] U. Knauer, A. Backhaus, and U. Seiffert. Beyond standard metrics - on the selection and combination of distance metrics for an improved classification of hyperspectral data. In T. Villmann, F.-M. Schleif, M. Kaden, and M. Lange, eds., *Advances in Self-Organizing Maps and Learning Vector Quantization: Proc. of 10th Internat. Workshop WSOM 2014, Mittweida*, p. 167–177, Berlin, 2014. Springer.

[6] L. Wang, M. Sugiyama, C Yang, K. Hatano, and J. Feng. Theory and algorithm for learning with dissimilarity functions. *Neural Computation*, 21(5):1459–1484, 2009.

[7] K.Q. Weinberger and L.K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009.

[8] B. Hammer, D. Hofmann, F.-M. Schleif, and X. Zhu. Learning vector quantization for (dis-)similarities. *Neurocomputing*, 131:43–51, 2014.

[9] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, 2002.

[10] T.-H. Wang, S.-F. Tian, and H.-K. Huang. Feature weighted support vector machine. *Journal of Electronics and Information Technology*, 31(3):514–518, 2009.

[11] S. Seo and K. Obermayer. Soft learning vector quantization. *Neural Computation*, 15:1589–1604, 2003.

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

[13] T. Villmann, M. Kaden, D. Nebel, and A. Bohnsack. A technical note about data similarities, dissimilarities, inner products and data metrics in the context of machine learning. *Machine Learning Reports*, 9(MLR-04-2015):1–16, 2015. ISSN:1865-3960, http://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_04_2015.pdf.

[14] M. Kaden, M. Riedel, W. Hermann, and T. Villmann. Border-sensitive learning in generalized learning vector quantization: an alternative to support vector machines. *Soft Computing*, 19(9):2423–2434, 2015.

[15] D. Nebel, B. Hammer, K. Frohberg, and T. Villmann. Median variants of learning vector quantization for learning of dissimilarity data. *Neurocomputing*, 169:295–305, 2015.

[16] D. Nebel and M. Kaden. Dissimilarity extraction in a median variant of learning vector quantization. *Machine Learning Reports*, 9(MLR-03-2015):33–40, 2015. ISSN:1865-3960, http://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_03_2015.pdf.

[17] Andreas Backhaus, Felix Bollenbeck, and Udo Seiffert. High-throughput quality control of coffee varieties and blends by artificial neural networks and hyperspectral imaging. In *Proc. of the 1st International Congress on Cocoa, Coffee and Tea, CoCoTea 2011*, 2011.