ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Generalized functional relevance learning vector quantization

M. Kästner[1] B. Hammer[2] M. Biehl[3] T. Villmann[1]

1 - University of Applied Science - Dept. of Mathematics,
Natural and Computer Sciences, 09648 Mittweida, Germany

2 - CITEC - Faculty of Technology, Bielefeld University, 33594 Bielefeld, Germany

3 - Johann Bernoulli Institute for Mathematics and Computer Science,
University of Groningen, The Netherlands

Abstract. Generalized learning vector quantization (GRLVQ) is a prototype based classification algorithm with metric adaptation weighting each data dimensions according to their relevance for the classification task. We present in this paper an extension for functional data, which are usually very high dimensional. This approach supposes the data vectors have to be functional representations. Taking into account, these information the so-called relevance profile are modeled by superposition of simple basis functions depending on only a few parameters. As a consequence, the resulting *functional* GRLVQ has drastically reduced number of parameters to be adapted for relevance learning. We demonstrate the ability of the new algorithms for standard functional data sets using different basis functions, namely Gaussians and Lorentzians.

Keywords: clustering; LVQ; metric adaption; functional relevance learning

1 Introduction

Prototype based classification is an important issue of many data analysis problems. One of the most prominent class of algorithms is the heuristically motivated family of Learning Vector Quantizers (LVQ) as introduced by Kohonen [4]. Sato and Yamada generalized this model such that an energy function reflecting the classification error is optimized by stochastic gradient learning (GLVQ) [7]. A further extension deals with metric adaptation to weight the input dimensions of the data according to their relevance for a given classification task (GRLVQ) [3]. Usually, this so called relevance learning, is based on weighting the Euclidean metric. After adaptation a relevance profile is obtained which consists of a vector weighting each data dimension according to its importance for classification. Yet, in Euclidean metric as well as in their weighted variant, the data dimensions are processed as uncorrelated features, i.e. the sequence of data dimensions does not contribute. This leads to a large number of independently parameters to be optimized in relevance learning. Especially, if the data are really high dimensional, as it is frequently in case of spectral data, time series etc., the relevance optimization may become crucial. Otherwise, the dimensions of data vectors obtained from such functional data carry lateral information and, therefore, should not be ignored for relevance learning.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

In this paper we introduce a functional relevance learning scheme for LVQ taking into account this functional information. For this purpose, the original vectorial relevance profile is now interpreted as a function to be adapted. Thereby, we propose to model the relevance function by superposition of a few parametrized basis functions. In this manner, the number of free parameters to be optimized in relevance learning is drastically reduced in comparison to original relevance learning. This can be seen as a kind of inherent regularization in relevance learning which also leads to greater stability.

2 Learning Vector Quantization by GRLVQ

Learning Vector Quantization as introduced by Kohonen is a heuristically motivated learning scheme. Given is a set of example data $\mathbf{v} \in V \subset \mathbb{R}^D$ with their labels $x_{\mathbf{v}} \in \mathcal{C} = \{1, 2, 3, ..., C\}$, the task is to distribute a set of prototypes $\mathbf{w} \in W \subset \mathbb{R}^D$ such that they represent the data set for classification, i.e. the classification accuracy should be minimized. For this purpose each prototype is also equipped with a class label $y_{\mathbf{w}}$ such that $\mathcal{C}$ is covered by all $y_{\mathbf{w}}$. After LVQ training a data point is assigned to the class of that prototype $\mathbf{w} \in W$ which has minimum distance.

A gradient based LVQ scheme was proposed by Sato and Yamada [7] (GLVQ) using the following energy function:

$$E(W) = \frac{1}{2} \sum_{\mathbf{v} \in V} f(\mu(\mathbf{v})) \text{ with } \mu(\mathbf{v}) = \frac{d^+(\mathbf{v}) - d^-(\mathbf{v})}{d^+(\mathbf{v}) + d^-(\mathbf{v})} \tag{1}$$

as approximation for the non-differentiable classification error. The function $f : \mathbb{R} \to \mathbb{R}$ is monotonically increasing, usually chosen as sigmoidal. Further, $\mu(\mathbf{v})$ is the classifier function with $d^+(\mathbf{v}) = d(\mathbf{v}, \mathbf{w}^+)$ denotes the distance between the data point $\mathbf{v}$ and the nearest prototype $\mathbf{w}^+$, which has the same label like $x_{\mathbf{v}} = y_{\mathbf{w}^+}$. In the following we abbreviate $d^+(\mathbf{v})$ simply by d^+. Further, $d(\mathbf{v}, \mathbf{w})$ is some differentiable dissimilarity measure with respect to $\mathbf{w}$. Analogously d^- is defined as the distance to the best prototype of all other classes.

The stochastic gradient learning for $E(W)$ is performed by

$$\frac{\partial_s E}{\partial \mathbf{w}^+} = \frac{\partial_s E}{\partial d^+} \frac{\partial d^+}{\partial \mathbf{w}^+}, \quad \frac{\partial_s E}{\partial \mathbf{w}^-} = \frac{\partial_s E}{\partial d^-} \frac{\partial d^-}{\partial \mathbf{w}^-} \tag{2}$$

with $\frac{\partial_s}{\partial}$ denotes the stochastic gradient and

$$\frac{\partial_s E}{\partial d^+} = \frac{2d^- \cdot f'(\mu(\mathbf{v}))}{(d^+ + d^-)^2}, \quad \frac{\partial_s E}{\partial d^-} = -\frac{2d^+ \cdot f'(\mu(\mathbf{v}))}{(d^+ + d^-)^2} \ .$$

In case of Euclidean metric usually applied in GLVQ $\partial d(\mathbf{v}, \mathbf{w}) / \partial \mathbf{w} = -(\mathbf{v} - \mathbf{w})$. GRLVQ is obtained if the scaled Euclidean metric is used for $d(\mathbf{v}, \mathbf{w})$:

$$d_\lambda(\mathbf{v}, \mathbf{w}) = \sum_{j=1}^{D} \lambda_j (v_j - w_j)^2 \tag{3}$$

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

with $\lambda_j \geq 0$ and $\|\lambda\|_1 = 1$. Additionally, in GRLVQ the λ_j are adapted again as gradient learning on E:

$$\frac{\partial_s E}{\partial \lambda_j} = \frac{\partial_s E}{\partial d^+} \frac{\partial d^+}{\partial \lambda_j} + \frac{\partial_s E}{\partial d^-} \frac{\partial d^-}{\partial \lambda_j} \quad \text{and} \quad \lambda_j := \lambda_j - \epsilon_\lambda \frac{\partial_s E}{\partial \lambda_j} \,. \tag{4}$$

Thereby, $0 < \epsilon_\lambda < 1$ is the learning rate which has to be chosen such that an adiabatic change compared to prototype learning is guarantied. The vector λ is called relevance profile. Further, note that the data dimensions are treated independently in this relevance learning scheme.

3 Generalized Functional Relevance LVQ

We now consider high dimensional data, which represent functions, i. e. $v_j = v(t_j)$ and, hence, prototypes will represent functions, too. Thus, the vector dimensions are not longer independently, yet, the index carries spatial or time information (depending on the interpretation on $\mathbf{t}$). Consequently, the relevance profile should be interpreted as a function $\lambda_j = \lambda(j)$, too.

Frequently, such functional data like spectra or time series may have data dimensions in the thousands or more. In that case, relevance learning in GRLVQ may become instable or/and slow in convergence due to the independent handling of data dimensions. Therefore, we suggest to reduce the number of free parameters in relevance learning by taking into account the functional character of the relevance profile.

In particular, we propose the approximation of $\lambda(j)$ by a weighted sum of K basis functions $\lambda(j) = \sum_{k=1}^{K} \beta_k \cdot \lambda_k(j)$ with $\beta_k > 0$ and $\sum_k \beta_k = 1$. Common choices for $\lambda_k(j)$ are standard Gaussians or Lorentzians, the latter for more sharply peaked profiles:

$$\lambda_k(j) = \frac{1}{\sigma_k \sqrt{2\pi}} \, e^{\frac{-(j-\Theta_k)^2}{2\sigma_k^2}} \,, \quad \lambda_k(j) = \frac{1}{\pi} \, \frac{\eta_k}{\eta_k^2 + (j-\Theta_k)^2} \,. \tag{5}$$

For both functions Θ_k is the center whereas the width and the height are determined by $\sigma_k \geq 0$ and $\eta_k \geq 0$, respectively. Now, relevance learning consists in adaptation of the weights and the parameters of these basis functions according to gradient learning on E. For example for the Gaussians, we preserve:

$$\frac{\partial_s E}{\partial \beta_k} = \frac{\partial_s E}{\partial d^+} \frac{\partial d^+}{\partial \beta_k} + \frac{\partial_s E}{\partial d^-} \frac{\partial d^-}{\partial \beta_k} \tag{6}$$

with

$$\frac{\partial d(\mathbf{v}, \mathbf{w})}{\partial \beta_k} = \frac{1}{\sigma_k \sqrt{2\pi}} \sum_{j=1}^{D} e^{\frac{-(j-\Theta_k)^2}{2\sigma_k^2}} (v_j - w_j)^2 \tag{7}$$

for β_k adaptation. The parameter Θ_k, σ_k and η_k are handled analogously:

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

In particular, for the Gaussians we obtain

$$\frac{\partial d(\mathbf{v},\mathbf{w})}{\partial \sigma_k} = \frac{\beta_k}{\sigma_k^2\sqrt{2\pi}} \sum_{j=1}^{D} \left(\frac{(j-\Theta_k)^2}{\sigma_k^2} - 1 \right) e^{\frac{-(j-\Theta_k)^2}{2\sigma_k^2}} (v_j - w_j)^2 \tag{8}$$

$$\frac{\partial d(\mathbf{v},\mathbf{w})}{\partial \Theta_k} = \frac{\beta_k}{\sigma_k^3\sqrt{2\pi}} \sum_{j=1}^{D} (j-\Theta_k) e^{\frac{-(j-\Theta_k)^2}{2\sigma_k^2}} (v_j - w_j)^2 \ , \tag{9}$$

whereas for the Lorentzians we get:

$$\frac{\partial d(\mathbf{v},\mathbf{w})}{\partial \beta_k} = \frac{1}{\pi} \sum_{j=1}^{D} \frac{\eta_k}{\eta_k^2 + (j-\Theta_k)^2} (v_j - w_j)^2 \tag{10}$$

$$\frac{\partial d(\mathbf{v},\mathbf{w})}{\partial \Theta_k} = \frac{\beta_k}{\pi} \sum_{j=1}^{D} \frac{2\eta_k (j-\Theta_k)}{(\eta_k^2 + (j-\Theta_k)^2)^2} (v_j - w_j)^2 \tag{11}$$

$$\frac{\partial d(\mathbf{v},\mathbf{w})}{\partial \eta_k} = \frac{\beta_k}{\pi} \sum_{j=1}^{D} \frac{(j-\Theta_k)^2 - \eta_k^2}{(\eta_k^2 + (j-\Theta_k)^2)^2} (v_j - w_j)^2 \ . \tag{12}$$

In that way the total number of free parameters becomes $3 * K$ for both models, which should be drastically smaller than D for reasonable K.

To avoid instabilities in learning which may lead to the fact that the center of basis functions become more or less equal, the following penalty term is added with a weighting factor $\alpha_r > 0$ to the energy function (1):

$$E(W) = \frac{1}{2} \sum_{\mathbf{v} \in V} f(\mu(\mathbf{v})) + \ \alpha_r \ P \quad \text{with} \qquad P = \sum_{i=1}^{K} \sum_{\substack{j=1 \\ j \neq i}}^{K} e^{-\frac{(\Theta_i - \Theta_j)^2}{2\zeta_i \zeta_j}} \tag{13}$$

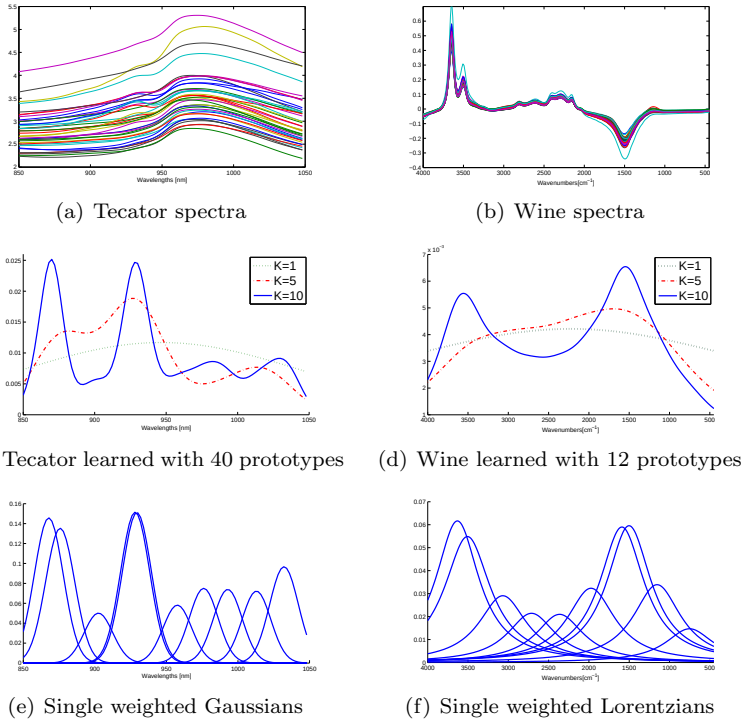
This term can be interpreted as a repulsion between the centers depend on the range of influence of the basis functions, e. g. $\zeta_k = \sigma_k$ and $\zeta_k = \eta_k$, respectively. This leads to an additional term for the gradient learning of Θ_k and σ_k for Gaussians:

$$\frac{\partial P}{\partial \Theta_k} = \sum_{\substack{i=1 \\ i \neq k}}^{K} \frac{2(\Theta_i - \Theta_k)}{\sigma_i \sigma_k} \ e^{-\frac{(\Theta_i - \Theta_k)^2}{2\sigma_i \sigma_k}} \tag{14}$$

$$\frac{\partial P}{\partial \sigma_k} = \sum_{\substack{i=1 \\ i \neq k}}^{K} \frac{(\Theta_i - \Theta_k)^2}{\sigma_i \sigma_k^2} \ e^{-\frac{(\Theta_i - \Theta_k)^2}{2\sigma_i \sigma_k}} \tag{15}$$

and analogously with η_k for Lorentzians. Hence, a minimum spreading of the basis function centers is guarantied.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.



(a) Tecator spectra (b) Wine spectra

(c) Tecator learned with 40 prototypes (d) Wine learned with 12 prototypes

(e) Single weighted Gaussians (f) Single weighted Lorentzians

Fig. 1: Examples of both data sets and the relevance profiles with a different number of basis functions. Last row: Distribution of the adapted weighted basis functions (Gaussians/Lorentzians) for Tecator (left) and Wine data (right) for the case $K = 10$ basis functions.

4 Experiments

We tested the GFRLVQ for classification on two well known real world data sets obtain from StatLib and UCI, Tecator (100 dimensions) and Wine (256 dimensions), respectively. Both data sets are spectra and, therefore, functional.

The Tecator data set consists of 215 spectra obtained for several meat probes, available on [1], see Fig.1a. The spectral ranges is between $850 - 1050$ nm wavelength. The data are split randomly into 144 training and 71 test data and labeled according to the two fat levels (low/high).

The Wine data set contains 121 absorbing infrared spectra of wine between 4000 and 400 cm^{-1} devided into 91 training and 30 test data [2], see Fig.1b. The data are classified according to their two alcohol levels (low/high) as given in [5].

According to the general shape of the data, we applied GFRLVQ with Gaussian basis functions for Tecator and Lorentzians for Wine, because the latter one is more sharply peaked. For both data sets we varied the value $K \in \{1, 5, 10\}$ (number of basis functions) and the number of prototypes. Hence, the relevance

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

	Tecator			Wine set		
$K \setminus \|W\|$	10	20	40	4	8	12
1	70.8% 70.5%	84.1% 70.5%	90.0% 85.3%	90.1% 73.3%	91.2% 80.0%	93.4% 83.3%
5	71.1% 71.6%	81.7% 75.8%	90.0% 83.2%	91.2% 76.7%	89.0% 73.3%	93.4% 83.3%
10	75.0% 76.8%	83.0% 80.0%	90.8% 84.2%	89.0% 80.0%	90.1% 80.0%	93.4% 86.7%
GRLVQ	71.7% 70.5%	87.5% 77.9%	94.2% 83.2%	93.4% 83.3%	91.2% 86.7%	93.4% 80.0%

Table 1: Correct classification rate of the training (1st value) and test (2nd value) sets with a different number of basis functions K and prototypes $|W|$

parameters are drastically reduced from 100 and 256 for Tecator and Wine, respectively. The results are depicted in the Tab.1. The achieved accuracy is also comparable to standard GRLVQ (see Tab.1) or other approaches, see [5], but with considerably lower number of parameters to be adapted. The obtained relevance profiles are depicted in Fig.1c,d which are also in good agreement with earlier publications [6], [5]. As one can expect, the shape of the relevance profile becomes richer with increasing K-value. The achieved distribution and the adapted shape of the weighted basis functions are depicted in Fig.1e,f for both data sets exemplary for the case $K = 10$. One can observed that height, width as well as the centers of the weighted basis functions were properly adapted.

5 Conclusion

We presented a functional extension of standard GRLVQ stressing the functional behavior of the data. In this way the number of parameters to be adapted for relevance learning is significantly reduced in case of such data, which are usually very high dimensionally. This can also be seen as a kind of inherent regularization in GFRLVQ compared to GRLVQ which leads to faster convergence preserving almost the accuracy.

References

[1] Submitted by Hans Henrik Thodberg. Tecator meat sample dataset, available on: http://lib.stat.cmu.edu/datasets/tecator.

[2] Wine data set provided by Prof. Marc Meurens. Available on http://www.ucl.ac.be/mlg/index.php?page=databases. meurens@bnut.ucl.ac.be.

[3] Barbara Hammer and Thomas Villmann. Generalized relevance learning vector quantization. *Neural Networks*, 15:1059–1068, 2002.

[4] T. Kohonen. *Learning vector quantization.* The handbook of brain theory and neural networks, Cambridge, 1995. MA: MIT Press.

[5] C. Krier, M. Verleysen, F. Rossi, and D. François. Supervised variable clustering for classification of nir spectra. In *Proceedings of XVIth European Symposium on Artificial Neural Networks (ESANN 2009)*, pages 263–268, Bruges, Belgique, April 2009.

[6] M. Strickert, U. Seiffert, N. Sreenivasulu, W. Weschke, T. Villmann, and B. Hammer. Generalized relevance lvq (grlvq) with correlation measures for gene expression analysis. *Neurocomput.*, 69:651–659, March 2006.

[7] A. S. Sato & K. Yamada. *Generalized learning vector quantization*, volume 7. Advances in neural information processing systems, Cambridge, 1995.