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/.

Learning Matrix Quantization and Variants of Relevance Learning

K. Domaschke[1], M. Kaden[2,*], M. Lange[2], and T. Villmann[2]

1 - Life Science Inkubator Dresden, Dresden - Germany

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

Abstract. We propose an extension of the learning vector quantization framework for matrix data. Data in matrix form occur in several areas like gray-scale images, time dependent spectra or fMRI data. If the matrix data are vectorized, important spatial information may be lost. Thus, processing matrix data in matrix form seems to be more appropriate. However, it requires matrix dissimilarities for data comparison. Here Schatten-p-norms come into play. We show that they can be used in a natural way replacing the vector dissimilarities in the learning framework. Moreover, we transfer the concept of vectorial relevance learning also to this new matrix variant. We apply the resulting learning matrix quantization approach to the classification of time-dependent fluorescence spectra as an exemplary real world application.

1 Introduction

Classification of complex data is still a challenging task in machine learning. Many methods were developed for processing vectorial data ranging from prototype based classifiers like Support Vector Machines (SVM, [17]) or the family of Learning Vector Quantizers (LVQ, [12]) to classification trees [5]. LVQ provides an intuitive and robust algorithmic approach based on Euclidean distance learning, frequently achieving promising results. One of the key extensions of LVQ is relevance learning, which weights the data dimensions performance improvement [8]. It can be further improved taking the data dimension correlations into account [18]. These methods can also be applied to data in matrix form (*matrix data*), if those data are simply vectorized. However, vectorization may destroy spatial relations within a matrix, such that important information is lost.

For this reason, we propose an extension of the LVQ framework for matrix data using dissimilarities explicitly based on matrix norms. A standard norm applied for distances between matrices is the Schatten-p-norm, as a natural counterpart of the l_p-norms for vectorial data. In the following we explain the modification of the LVQ required to process matrix data and denote this new variant as *Learning Matrix Quantization* (LMQ). The automatic weighting of the dimensions the classification performance improvement is a vitally extension of the LVQ known as *relevance learning* [8]. In this contribution we also discuss concepts of relevance learning in LMQ. An exemplary real world application illustrates the usefulness of the LMQ with relevance learning.

2 Learning Vector Quantization based on l_p-norms

LVQ was introduced by KOHONEN as an intuitive prototype based learning classifier for vector data heuristically approximating a Bayes-classifier [11]. The generalized learn-

*M. Kaden 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/.

ing vector quantization (GLVQ) model [15] is a cost function based modification of the intuitive LVQ approximating the classification error as the objective to be minimized. The prototypes of the GLVQ model are the set $W = \{\mathbf{w}_k \in \mathbb{R}^n, k = 1 \dots M\}$. Each data vector $\mathbf{v} \in V \subseteq \mathbb{R}^n$ of the training data belongs to a class $x_\mathbf{v} \in \mathcal{C} = \{1, \dots, C\}$. The prototypes are labeled by $y_\mathbf{w} \in \mathcal{C}$ such that there is at least one prototype per class. Thereby, the dissimilarities between data points $\mathbf{v}$ and an arbitrary prototype $\mathbf{w}$ are judged in terms of a measure d based on the l_p-norm $\|\mathbf{v} - \mathbf{w}\|_p = \sqrt[p]{\sum_{i=1}^{n} |v_i - w_i|^p}$. In particular, we consider $d_p(\mathbf{v}, \mathbf{w}) = \left(\|\mathbf{v} - \mathbf{w}\|_p\right)^p$. Further $d_p^+(\mathbf{v}) = d_p(\mathbf{v}, \mathbf{w}^+)$ 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 $d_p^-(\mathbf{v}) = d_p(\mathbf{v}, \mathbf{w}^-)$ is the dissimilarity degree for the best matching prototype $\mathbf{w}^-$ with a class label $y_{\mathbf{w}^-}$ different from $x_\mathbf{v}$. In standard GLVQ, the squared Euclidean distance $d_2(\mathbf{v}, \mathbf{w}) = (\|\mathbf{v} - \mathbf{w}\|_2)^2$ is used. GLVQ maximizes the hypothesis margin $d_p^+(\mathbf{v}) - d_p^-(\mathbf{v})$ [2, 7]. 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 squashing function usually chosen as a sigmoid or the identity function and

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

is the classifier function with $\mu(\mathbf{v}) \in [-1, 1]$. Learning in GLVQ is performed by the *stochastic* gradient descent learning for the cost function E_{GLVQ}. For the more general l_p-norm, the prototype updates of $\mathbf{w}^+$ and $\mathbf{w}^-$ become

$$\Delta\mathbf{w}^\pm \propto \mp \frac{\partial f}{\partial \mu(\mathbf{v})} \cdot \frac{\partial \mu(\mathbf{v})}{\partial d_p^\pm(\mathbf{v})} \cdot \frac{\partial d_p^\pm(\mathbf{v})}{\partial \mathbf{w}^\pm} \tag{3}$$

with $\frac{\partial d_p^\pm}{\partial \mathbf{w}^\pm}$ are the formal derivatives of $d_p(\mathbf{v}, \mathbf{w})$ [14]. Several extensions were proposed to adapt this basic scheme according to specific classification tasks. A recent overview can be found in [10]. One of the most successful modifications is relevance learning in GLVQ (GRLVQ, [3, 8]). In the GRLVQ, the dissimilarity measure $d_p(\mathbf{v}, \mathbf{w})$ is replaced by

$$d_{p,\mathbf{r}}(\mathbf{v}, \mathbf{w}) = \|\mathbf{r} \circ (\mathbf{v} - \mathbf{w})\|_p^p \tag{4}$$

with the relevance vector $\mathbf{r}$ consisting of the relevances r_i with normalization $\sum_{i=1}^{n} |r_i|^p = 1$. Here, $\mathbf{r} \circ \mathbf{x}$ denotes the Hadamard product. The *relevances r_i weight each data dimension independently* to improve the classifier performance and can also be adapted by stochastic gradient learning according to

$$\Delta r_i \propto -\frac{\partial f}{\partial \mu(\mathbf{v})} \cdot \left(\frac{\partial \mu(\mathbf{v})}{\partial d_{p,\mathbf{r}}^+(\mathbf{v})} \cdot \frac{\partial d_{p,\mathbf{r}}^+(\mathbf{v})}{\partial r_i} - \frac{\partial \mu(\mathbf{v})}{\partial d_{p,\mathbf{r}}^-(\mathbf{v})} \cdot \frac{\partial d_{p,\mathbf{r}}^-(\mathbf{v})}{\partial r_i} \right) . \tag{5}$$

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/.

The GRLVQ can be further generalized when using

$$d_{p,\mathbf{\Omega}}\left(\mathbf{v},\mathbf{w}\right)=\left\|\mathbf{\Omega}\left(\mathbf{v}-\mathbf{w}\right)\right\|_p^p \tag{6}$$

where $\mathbf{\Omega}\in\mathbb{R}^{m\times n}$ is a mapping matrix. Then $\mathbf{\Lambda}=\mathbf{\Omega}^T\mathbf{\Omega}$ can be interpreted after stochastic gradient learning as a classification correlation matrix combining those data dimensions, which supports the class separabilities [18].

3 Learning Matrix Quantization based on Schatten-p-norms

Matrix data are the obvious extension of vector data. However, those data are frequently processed applying a vectorization scheme and then utilizing of vector methods. Thus, structural information can be destroyed by the vectorization. Therefore, we propose to modify LVQ for matrix data $\mathbf{V}\in V_{m,n}\subseteq\mathbb{R}^{m\times n}$. In this case, the prototype set W consists of matrices $\mathbf{W}_k\in\mathbb{R}^{m\times n}$. In the next step, we have to replace the l_p-norm $\left\|\mathbf{v}\right\|_p$ determining the dissimilarity measure $d_p\left(\mathbf{v},\mathbf{w}\right)$ by a respective matrix norm. Mathematically, the set $\mathbb{R}^{m\times n}$ of matrices is a vector space, which can easily be equipped with a respective matrix norm fulfilling the usual norm axioms. Examples are the maximum norm defined as the maximum absolute value of the matrix entries or the Ky-Fan-Norm as the sum of the first K singular values of the matrix [9]. More sophisticated norms are those, which additionally are consistent with the matrix multiplication, i.e. satisfying the Cauchy-Schwarz-inequality $\left\|\mathbf{X}\cdot\mathbf{Y}\right\|\leq\left\|\mathbf{X}\right\|\cdot\left\|\mathbf{Y}\right\|$ [6]. One prominent example for these so-called *sub-multiplicative* norms is the Schatten-p-norm

$$s_p\left(\mathbf{A}\right)=\sqrt[p]{tr(\left|\mathbf{A}\right|^p)} \tag{7}$$

where $tr(\bullet)$ is the trace operator [16]. For $p=2$, the Schatten-p-norm reduces to the Frobenius norm $s_2\left(\mathbf{A}\right)=\sqrt{tr(\mathbf{A}\mathbf{A}^T)}$. Schatten-$p$-norms are closely related to l_p-norms according to $s_p\left(\mathbf{A}\right)=\left\|\sigma\left(\mathbf{A}\right)\right\|_p$, where $\sigma\left(\mathbf{A}\right)$ denotes the vector of the singular values of $\mathbf{A}$. Based on (7), we take

$$\delta_p\left(\mathbf{V},\mathbf{W}\right)=\left(s_p\left(\mathbf{V}-\mathbf{W}\right)\right)^p \tag{8}$$

as a dissimilarity measure comparable to $d_p\left(\mathbf{v},\mathbf{w}\right)$, which can be plugged into the cost function (1) reading now as

$$E_{GLMQ}\left(W\right)=\frac{1}{2}\sum_{\mathbf{V}\in V_{m,n}}f\left(\mu\left(\mathbf{V}\right)\right)\ . \tag{9}$$

Applying the same formalism as for GLVQ, we obtain the formal update rules

$$\Delta\mathbf{W}^{\pm}\propto\mp\frac{\partial f}{\partial\mu\left(\mathbf{V}\right)}\cdot\frac{\partial\mu\left(\mathbf{V}\right)}{\partial\delta_p^{\pm}\left(\mathbf{V}\right)}\cdot\frac{\partial\delta_p^{\pm}\left(\mathbf{V}\right)}{\partial\mathbf{W}^{\pm}} \tag{10}$$

for $\mathbf{W}^{\pm}$. For $p=2$, we simply get $\frac{\partial\delta_p^{\pm}\left(\mathbf{V}\right)}{\partial\mathbf{W}^{\pm}}=-2\mathbf{W}^{\pm}$. Accordingly, the algorithm is denoted as *Generalized Learning Matrix Quantization* (GLMQ).

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/.

4 Relevance Learning in GLMQ

In the following we will discuss variants of relevance learning for LMQ. The obvious counterpart to the vector variant (4) for Schatten-p-norms would be

$$\delta_{p,\mathbf{R}\circ}(\mathbf{V},\mathbf{W}) = \left(s_p\left(\mathbf{R}\circ(\mathbf{V}-\mathbf{W})\right)\right)^p \tag{11}$$

with the *relevance matrix* $\mathbf{R}$ *weighting independently each entry of a data matrix* $\mathbf{V}$ by the Hadamard product. For $p = 2$, this approach yields $\delta_{2,\mathbf{R}\circ}(\mathbf{V},\mathbf{W}) = tr\left((\mathbf{R}\circ(\mathbf{V}-\mathbf{W}))(\mathbf{R}\circ(\mathbf{V}-\mathbf{W}))^T\right)$. The formal relevance update as the stochastic gradient of (9) becomes

$$\Delta\mathbf{R} \propto -\frac{\partial f}{\partial\mu(\mathbf{V})}\cdot\left(\frac{\partial\mu(\mathbf{V})}{\partial\delta^+_{p,\mathbf{R}\circ}(\mathbf{V})}\cdot\frac{\partial\delta^+_{p,\mathbf{R}\circ}(\mathbf{V})}{\partial\mathbf{R}} - \frac{\partial\mu(\mathbf{V})}{\partial\delta^-_{p,\mathbf{R}\circ}(\mathbf{V})}\cdot\frac{\partial\delta^-_{p,\mathbf{R}\circ}(\mathbf{V})}{\partial\mathbf{R}}\right) \tag{12}$$

where $\frac{\partial\delta_{2,\mathbf{R}\circ}(\mathbf{V})}{\partial\mathbf{R}} = 2\mathbf{R}\circ(\mathbf{V}-\mathbf{W})\circ(\mathbf{V}-\mathbf{W})$ is obtained for $p = 2$. It is accompanied by the respective prototype derivatives involving the term $\frac{\partial\delta_{2,\mathbf{R}\circ}(\mathbf{V})}{\partial\mathbf{W}^\pm} = -2\mathbf{R}\circ\mathbf{R}\circ(\mathbf{V}-\mathbf{W}^\pm)$ for $p = 2$. We denote this variant as Hadamard-Relevance-Learning (HRL). We notice that the HRL learning applying the Frobenius-norm can be transfered to the vectorial counterpart GRLVQ with the Euclidean norm.
Alternatively to the Hadamard product based relevance weighting introduced in (11) we can think about the weighting

$$\delta_{p,\mathbf{R}}(\mathbf{V},\mathbf{W}) = \left(s_p\left(\mathbf{R}\cdot(\mathbf{V}-\mathbf{W})\right)\right)^p \tag{13}$$

using the ordinary matrix multiplication. Here, a *weighted linear relevance mixing* is applied, taking partially linear combinations of matrix entries into account. The relevance update is structurally equivalent as in (12) paying attention to the new derivative $\frac{\partial\delta_{p,\mathbf{R}}(\mathbf{V})}{\partial\mathbf{R}}$. For $p = 2$ we get $\frac{\partial\delta_{2,\mathbf{R}}(\mathbf{V})}{\partial\mathbf{R}} = 2\mathbf{R}\cdot(\mathbf{V}-\mathbf{W})\cdot(\mathbf{V}-\mathbf{W})^T$. The respective prototype update involves the term $\frac{\partial\delta_{2,\mathbf{R}}(\mathbf{V})}{\partial\mathbf{W}^\pm} = -2\mathbf{R}^T\cdot\mathbf{R}\cdot(\mathbf{V}-\mathbf{W}^\pm)$ and the method is referred here as *Multiplicative Relevance Learning* (MRL). At this point, applying the MRL with the Frobenius norm the spatial matrix information is taken into account.

5 Application in Time Resolved Laser induced Fluorescence Spectroscopy

In the last 20 years it turned out that one important way to distinguish between different substances or to characterize a composite sample is the *Time Resolved Laser induced Fluorescence Spectroscopy* (TRLFS). This method selectively excites the molecules of a desired substance to a specified state of higher energy by a laser beam. In the relaxing phase the molecules temporally dissipate this energy thermally or optically in terms of auto-fluorescence. The latter leads to time-dependent emission and can be measured as a signal reflecting the substance specific spatio-temporal characteristics of this process. The obtained two-dimensional signal (matrix) shows the fluorescence intensity [a.u.] dependent on the emission energy [nm] and the time [ns] after fluorescence initialization. These spectra are called Time Resolved Fluorescence Spectra (TRFS). Depending on the experiment in use, the emission spectra of different biological compounds can

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/.

	GLVQ	GRLVQ	GLMQ	GRLMQ (HRL)	GRLMQ (MRL)
acc. in %	77.1	79.7	81.7	81.0	85.2
std. dev.	0.118	0.114	0.110	0.108	0.099

Table 1: Classification test results for the TRFLS dataset averaged over 10 repetitions of 10-fold cross-validatations.

be quite similar and difficult to distinguish, so the important information is the time dependence [13, p.578]. For the problem to be investigated in this study, the given data includes only TRFS of a single biological compound at two different excitation energies (energy level/class). The underlying assumption is that the substrate response specifically and this behavior is reflected in the measured signal matrices [4]. However, the specificity of the signal may be overlayed by noise. Therefore, the time-dependent logarithmic signals are integrated with respect to time to reduce the noise. However, according to this vectorization the time-dependent information is lost, which may contain substantial information.

Thus we obtain for an experiment both vector and matrix data for GLVQ and GLMQ analysis, respectively, for comparison. In particular, we have 60 data for each energy class with a matrix resolution of 100×20 for emission energy and time, respectively. For both algorithms we used only one prototype per class. All test accuracies presented as the results are obtained as average of 10 random repetitions of 10-fold cross-validations. Further, we applied relevance learning for comparison, whereby for GLMQ we considered both introduced variants. The results are collected in Tab. 1. We observe that the vector variant GLVQ is slightly improved by relevance learning, as expected. However, a similar improvement is obtained by GLMQ without relevance learning. If relevance learning is included, a further improvement may be achieved. The improvement amount depends on the kind of the relevance learning method. While application of the HRL, which is mathematically comparable to GRLVQ, does not delivers an improvement, the progress becomes moderate if MRL is used. This can be dedicated to the fact that MRL takes more structured matrix relations into account than HRL.

6 Conclusion and Future Work

In the contribution we propose the Learning Matrix Quantization framework as extension of GLVQ for matrix data. We emphasize the fact that otherwise vectorization of matrix data may destroy structured information coded in the matrix. Further, we discuss possibilities of relevance learning for GLMQ. So far we considered Hadamard-relevance and (left) multipicative learning. The latter one delivers better results, which may dedicated to the partially considered correlations within the data matrices. An obvious option for future research would be to consider the right multiplicative variant or, as a more sophisticated possibility, combining both. This leads to the $\mathbf{QR}$-norms as introduced in fMRI-analysis to relate spatio-temporal dependencies [1], but causing substantially increased numerical complexity. Another, way could be to apply also the Kronecker-matrix-multiplication offering another kind of structural combination of matrix entries [6, 9]. Finally, the vectorial matrix learning by GMLVQ would lead to tensor-based relevance learning for matrix data as adequate counterpart.

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/.

References

[1] G. Allen, L. Grosenick, and J. Taylor. A generalized least squares matrix decomposition. *Journal of the American Statistical Association, Theory & Methods*, 109(505):145–159, 2012.

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

[3] T. Bojer, B. Hammer, D. Schunk, and T. von Toschanowitz K. Relevance determination in learning vector quantization. In *9th European Symposium on Artificial Neural Networks. ESANN'2001. Proceedings. D-Facto, Evere, Belgium*, pages 271–6, 2001.

[4] P. Clark, Z.-P. Liu, J. Zhang, and L. Gierasch. Intrinsic tryptophans of CRABPI as probes of structure and folding. *Protein Science*, 5(6):1108–1117, 1996.

[5] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, 1973.

[6] G. Golub and C. V. Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. John Hopkins University Press, Baltimore, Maryland, 4th edition, 2013.

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

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

[9] R. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2013.

[10] 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.

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

[12] T. Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995. (Second Extended Edition 1997).

[13] J. Lakowicz. *Principles of Fluorescence Spectroscopy*. Springer, 3rd edition, 2006.

[14] M. Lange and T. Villmann. Derivatives of l_p-norms and their approximations. *Machine Learning Reports*, 7(MLR-04-2013):43–59, 2013. ISSN:1865-3960, http://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_04_2013.pdf.

[15] 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.

[16] R. Schatten. *A Theory of Cross-Spaces*, volume 26 of *Annals of Mathematics Studies*. Princeton University Press, 1950.

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

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