ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Enforcing Feature Sparseness while Preserving Reliable Classification for Prototype-Based Models

M. Kaden[1], J. Voigt[1], S. Saralajew[2], and T. Villmann[1,3] *

1- Saxon Institute for Computational Intelligence and Machine Learning (SICIM), Mittweida University of Applied Sciences, Mittweida, Germany

2- NEC Laboratories Europe GmbH, Heidelberg, Germany

3- Technical University Bergakademie Freiberg, Freiberg, Germany

Abstract. Machine learning classifiers adjust implicitly or explicitly the importance of the data features to solve a given classification task. This feature weighting often does not imply feature sparseness, which, however, may be important for interpretability and model evaluation. This contribution proposes how to force feature sparseness in combination with feature relevance for prototype-based classification learning to obtain reliable and interpretable classification decisions.

1 Introduction

Learning vector quantization (LVQ) is a popular classification learning approach originally introduced in [1]. it constitutes a prototype-based model realizing a shallow but powerful scheme for learning and inference. The underlying nearest prototype approach implies inherent model interpretability [2, 3]. A sophisticated approximation of the classification error as provided in [4] allows gradient descent learning for the prototype adjustment to be known as *generalized* LVQ (GLVQ). It has been mathematically proven that GLVQ is a robust classifier maximizing implicitly the hypothesis margin during learning [5, 6].

A substantial performance progress can be achieved if the (matrix) *relevance learning* strategies for the determination of data feature importance and correlations are incorporated into the basic learning scheme – GMLVQ [7, 8]. Relevance learning also improves the GLVQ-interpretability by evaluation of the adjusted feature relevances and correlations contributing to the classification decisions. This has been demonstrated to be beneficial in particular in the context of biomedical research [9, 10, 11, 12]. Yet, interpretation of the relevance learning result can be impractical if the number of data features is large as, for instance, in gene expression data analysis, where hundreds or thousands of gene values are investigated. To overcome related challenges, recently developed variants try to regularize relevance learning by incorporation of external expert knowledge [13]. However, this attempt in general does not lead to a substantial reduction of feature influence for the classification decisions. This is considered disadvantageous if the relevant features have to be further investigated, e.g. by bio-chemical analyzes or medical treatment research [14]. For those problems a sparse feature relevance provides an option to reduce costs and efforts significantly.

*J.V. and M.K. were partially supported by ILEAS-project *Stiftung Innovation in der Hochschullehre* – (1001-2196) and EFRE-project *KIMed* – (100734114)).

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Feature sparseness in machine learning frequently is realized by a respective l_0-regularization of feature weighting being well-considered in (deep) multilayer perceptron learning [15, 16]. The l_0-regularization adjustment using an approximated gradient approach is known from [17, 18]. A respective approach in the context of relevance learning for GMLVQ is not considered so far. Thus, we propose in this contribution how to combine relevance learning with feature sparse by l_0-regularization for GMLVQ and provide a respective gradient descent learning scheme. We denote this variant as *sparse* GMLVQ (sGMLVQ) and demonstrate the model abilities for bio-chemical classification problems.

2 Generalized Matrix LVQ (GMLVQ)

Suppose classification training data $\mathcal{T} = \{(\mathbf{x}, c(\mathbf{x}))\} \subset \mathbb{R}^n \times \mathcal{C}$ in the feature data space $\mathbb{R}^n$ where $\mathcal{C} = \{1, \ldots, C\}$ is the set of the available classes such that $c(\mathbf{x})$ is the class label for a sample $\mathbf{x}$. Further, GMLVQ requires a set $\mathcal{P} = \{(\mathbf{p}, c(\mathbf{p}))\} \subset \mathbb{R}^m \times \mathcal{C}$ of M prototypes $\mathbf{p}$ such that each class is represented by at least one prototype. GMLVQ inference for an unknown sample $\hat{\mathbf{x}} \in \mathbf{R}^n$ is based on the nearest prototype classification principle (NPCP) $\hat{\mathbf{x}} \mapsto c\left(\mathbf{p}_{s(\hat{\mathbf{x}})}\right)$ where $s(\hat{\mathbf{x}}) = \underset{i=1,\ldots,M}{\operatorname{argmin}} d_{\mathbf{\Omega}}(\hat{\mathbf{x}}, \mathbf{p}_i)$ is the index of the closest prototype to $\hat{\mathbf{x}}$ with respect to the parametrized dissimilarity measure

$$d_{\mathbf{\Omega}}(\mathbf{x}, \mathbf{p}) = (\mathbf{\Omega}\mathbf{x} - \mathbf{p})^\top (\mathbf{\Omega}\mathbf{x} - \mathbf{p})$$

representing the squared Euclidean in the data embedding space $\mathbb{R}^m$ and the embedding map $\mathbf{\Omega} \in \mathbb{R}^{m \times n}$. In GMLVQ the cost function

$$E_{\mathrm{GMLVQ}}(\mathcal{T}, \mathcal{P}, \mathbf{\Omega}) = \sum_{(\mathbf{x}, c(\mathbf{x})) \in \mathcal{T}} f\left(\frac{d_{\mathbf{\Omega}}(\mathbf{x}, \mathbf{p}^+) - d_{\mathbf{\Omega}}(\mathbf{x}, \mathbf{p}^-)}{d_{\mathbf{\Omega}}(\mathbf{x}, \mathbf{p}^+) + d_{\mathbf{\Omega}}(\mathbf{x}, \mathbf{p}^-)}\right)$$

is minimized, with $f(\cdot)$ being a monotonically increasing function frequently chosen as sigmoidal. The parameters $\mathbf{p}^+$, $\mathbf{p}^-$ are the best matching correct and incorrect prototypes for the training pair $(\mathbf{x}, c(\mathbf{x}))$, respectively. This cost function approximates the classification error and allows stochastic gradient descent learning (SGDL) for the prototypes [4]. Further, SGDL for the mapping matrix $\mathbf{\Omega}$ realizes a relevance learning scheme adjusting the embedding in this way that best class separation by means of the prototype set $\mathcal{P}$ is achieved [8]. The resulting matrix $\Lambda = \mathbf{\Omega}^\top \cdot \mathbf{\Omega}$ is denoted as relevance matrix or classification correlation matrix, because the entries Λ_{ij} can be interpreted as feature correlations contributing to class separation. Original relevance learning as proposed in [7] by immediate feature weighting is obtained for the choice $m = n$ and diagonal $\mathbf{\Omega}$. Yet, neither GMLVQ nor original relevance learning tackle the problem of feature sparseness.

3 Feature Sparseness in GMLVQ

As already pointed out in the introduction, feature sparseness becomes important if the number of data features required for sufficient classification performance has to be reduced according to external constraints like minimizing feature generation costs or better model interpretability. To incorporate feature sparseness in GMLVQ we parametrize the mapping matrix $\mathbf{\Omega} \in \mathbb{R}^{m \times n}$ by

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

$\mathbf{\Omega_s} = (s_1 \cdot \mathbf{\Omega}_1 | \ldots | s_n \cdot \mathbf{\Omega}_n)$ where $\mathbf{\Omega}_i = (\Omega_{1i}, \ldots, \Omega_{mi})^\top \in \mathbb{R}^m$ are the column vectors of $\mathbf{\Omega}$ and $\mathbf{s} = (s_1, \ldots, s_n)$ is the sparseness vector. Further we assume $s_k = r_k^2$ to ensure positive values for s_k and observe that

$$\mathbf{\Omega_s} \cdot \mathbf{x} = \mathbf{\Omega}(\mathbf{s} \circ \mathbf{x}) \tag{1}$$

is valid. Thus, the sparseness vector forces feature sparsity for $\|\mathbf{s}\|_0 \searrow \min$, which is known as l_0-regularization [16].[1]

The quantity $\|\mathbf{s}\|_0 = \#\{j \in \{1, \ldots, n\} | s_j \neq 0\}$ is known as l_0-(pseudo)-norm and counts the non-zero elements of $\mathbf{s}$.It shares three out of the four axioms of a norm: non-negativity, positivity and sub-additivity ($\triangle$-inequality) [20, 21, 19]. Yet, the 1-homogeneity does not hold true, i.e. $\|\alpha \cdot \mathbf{r}\|_0 \neq |\alpha| \|\mathbf{r}\|_0$ is valid, which implies non-convexity. In contrast, it is 0-homogeneous: $\|\alpha \cdot \mathbf{r}\|_0 = \|\mathbf{r}\|_0$.

Using this regularization, we consider the cost function

$$E_{\text{sGMLVQ}}(\mathcal{T}, \mathcal{P}, \mathbf{\Omega}) = (1 - \gamma) \cdot E_{\text{GMLVQ}}(\mathcal{T}, \mathcal{P}, \mathbf{\Omega}) + \gamma \cdot \|\mathbf{s}\|_0 \tag{2}$$

for feature-sparse GMLVQ (sGMLVQ) where the l_0-regularization determines a penalty term. SGDL for this cost function with respect to both, the prototypes $\mathcal{P}$ and the mapping matrix $\mathbf{\Omega}$, is only marginally influenced by the sparseness vector interpreting $\mathbf{s} \circ x$ as a modified data sample. To avoid standard LASSO optimization to determine sparse $\mathbf{s}$-vectors, we prefer SGDL for the sparseness penalty $\|\mathbf{s}\|_0$, which, however, requires smooth approximations. Following [17], we take the approximation $l_0(\mathbf{s}) = \sum_{k=1}^{n} g(w_k)$ where $g(w_k)$ are smooth functions. In [17, 18] the functions

$$g_\epsilon(r_k) = 1 - \exp\left(-\frac{r_k^2}{2\sigma^2}\right) \quad \text{and} \quad g_\sigma(r_k) = \frac{r_k^2}{r_k^2 + \epsilon} \tag{3}$$

are considered where $\epsilon, \sigma > 0$ are a hyperparameters, respectively, and $s_k = r_k^2$ as defined before. Other approaches can be found in [22] Further, additional knowledge regarding supposed feature importance can be included using appropriate initializations $\mathbf{s}_0$.

4 Numerical Experiments and Results

4.1 Datasets and Standardized Learning Procedure

The experimental evaluation employs three standard benchmark datasets from the UCI-database [23] that span medical diagnostics and chemical analysis: (i) the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, comprising 569 tumor samples (benign or malignant) described by $n = 30$ real-valued morphological features; (ii) the Pima Indians Diabetes (PIMA) dataset, containing $n = 8$ clinical values for 768 female patients of Pima heritage with a binary outcome indicating the presence or absence of diabetes; and (iii) the Wine dataset (WINE), which includes 178 instances of three cultivars of the Italian wine each described by $n = 13$ physicochemical attributes. A fourth dataset of cancer patients (P-NET) was chosen which already has a vastly reduced feature set $n = 60$ in order

[1] Further, the level sets $\mathcal{L}_0^k = \{\mathbf{s} \in \mathbb{R}^n | \|\mathbf{s}\|_0 \leq k\}$ are established and for the subset $\mathcal{K} \subseteq \{1, \ldots n\}$ the set $\boldsymbol{S}_\mathcal{K} = \{\mathbf{s} \in \mathbb{R}^n | s_k = 0 \iff k \notin \mathcal{K}\}$ determines the subspace of $\mathbb{R}^n$ containing those vectors whose elements s_k are vanishing if k does not belong to $\mathcal{K}$ [19].

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

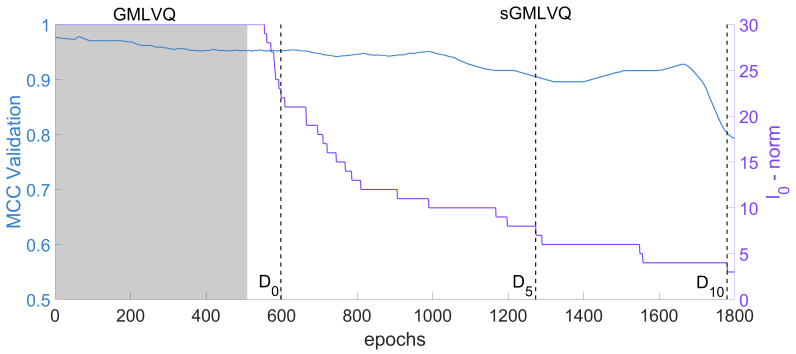

Fig. 1: WDBC: validation performance (blue) and $||s||_0$ (purple) for GMLVQ followed by sGMLVQ. Three checkpoints are marked: D_0 with 23 features, D_5 with 8 and D_{10} with only 4 remaining features (more explanation in text).

to see whether it can be reduced further. The data stems from [24] and has been examined in [13]. The training process consists of the following steps:

- GMLVQ with $\Omega \in \mathbb{R}^{m \times n}$ with $m \leq n$ is trained for fixed $\mathbf{s}$ ($\forall j : \; s_j = 1$) until convergence; classification validation measure (Matthews correlation coefficient – MCC) is determined and referred as original MCC ($oMCC$)
- Ω is then fixed and $\mathbf{s}$ is enabled to receive updates from the optimizer but $\gamma = 0$ is constant until current MCC is converged
- Continue training with slow linearly increasing penalty factor γ from (2) enforcing feature sparseness using the approximation g_σ from (3) for gradient decent learning of l_0. influence/importance of the penalty term is gradually and linearly increased slightly each epoch.
- Training is stopped if the current MCC drops down significantly compared to the original MCC

4.2 Analysis and Results

As the sparsity constraint gains influence over the loss to limit the number of used features, performance inevitably drops. The amount of decrease depends on the specific dataset and its domain. To deal with this data specificity but keeping comparability we identify this tipping point D_0 by an one-sided t-test [25] between current MCC (sliding average) and $oMCC$ for each experiment separately and determine the corresponding number η_0 of remaining features. Further tipping points D_i were determined at which the performance loss significantly is greater than 5% and 10% compared to $oMCC$ yielding η_5 and η_{10} as the number of remaining features, respectively. An illustration of this behavior for WDBC is depicted in Fig. 1.

The results for all four datasets according to this setting are summarized in Tab. 1. We observe that frequently a still high performance can be kept although the number of remaining features is significantly reduced. As expected, if the information loss by feature elimination becomes too large, model performance collapses.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

	GMLVQ	sGMLVQ					
		D_0		D_5		D_{10}	
dataset	MCC	MCC	η_0	MCC	η_5	MCC	η_{10}
WDBC	0.962	0.936	15.0	0.879	5.2	0.815	3.5
n=30	± 0.02	± 0.04	± 7.8	± 0.04	± 1.5	± 0.07	± 1.2
WINE	0.976	0.965	8.3	0.913	4.7	0.860	3.2
n=13	± 0.04	± 0.04	± 2.9	± 0.06	± 1.8	± 0.08	± 0.9
PIMA	0.533	0.478	4.4	0.439	3.0	0.399	2.6
n=8	± 0.08	± 0.08	± 1.7	± 0.07	± 0.9	± 0.07	± 0.5
P-NET	0.680	0.633	51.8	0.606	28.7	0.564	11.8
n=60	± 0.05	± 0.05	± 17.0	± 0.05	± 24.1	± 0.06	± 17.5

Table 1: Validation MCC values together with the l_0-norm value η for the standard GMLVQ and the sGMLVQ for different D-levels (D_i means i% difference to the original MCC value). All constellations run 10× with 5-fold cross validation.

5 Conclusions

In this contribution we propose a feature elimination scheme by l_0-norm penalty, which can be used in matrix relevance learning vector quantization. In this way, the idea of relevance learning to determine important feature correlations for classification decision can be preserved and is overlayed by the enforcement of feature sparseness. Further investigations should include the interpretation and evaluation of the corresponding re-adjustement of relevance matrix, which could contribute to causal inference patterns.

References

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

[2] M. Biehl, B. Hammer, and T. Villmann. Prototype-based models in machine learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(2):92–111, 2016.

[3] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16:1–85, 2022.

[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. Proceedings of the 1995 Conference*, pages 423–9. MIT Press, Cambridge, MA, USA, 1996.

[5] K. Crammer, R. Gilad-Bachrach, A. Navot, and N.Tishby. Margin analysis of the LVQ algorithm. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing (Proc. NIPS 2002)*, volume 15, pages 462–469, Cambridge, MA, 2003. MIT Press.

[6] S. Saralajew, L. Holdijk, M. Rees, and T. Villmann. Robustness of generalized learning vector quantization models against adversarial attacks. In A. Vellido, K. Gibert, C. Angulo, and J.D.M. Guerrero, editors, *Advances in Self-Organizing Maps, Learning Vector Quantization, Clustering and Data Visualization – Proceedings of the 13th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization, WSOM+2019, Barcelona*, volume 976 of *Advances in Intelligent Systems and Computing*, pages 189–199. Springer Berlin-Heidelberg, 2019.

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

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

[9] W. Arlt, M. Biehl, A. E. Taylor, S. Hahner, R. Libe, B. A. Hughes, P. Schneider, D. J. Smith, H. Stiekema, N. Krone, E. Porfiri, G. Opocher, J. Bertherat, F. Mantero, B. Allolio, M. Terzolo, P. Nightingale, C. H. L. Shackleton, X. Bertagna, M. Fassnacht, and P. M. Stewart. Urine steroid metabolomics as a biomarker tool for detecting malignancy in adrenal tumors. *J Clinical Endocrinology and Metabolism*, 96:3775–3784, 2011.

[10] A. Rana, A. Shaker, S. Saralajew, T. Suzuki, K. Yasuda, S. Kato, T. Wada, T. Fujikawa, and T. Kikutsuji. Prototype-based learning for healthcare: A demonstration of interpretable AI. In *Workshop Proceedings of the IEEE International Conference on Data Mining (ICDM), Washington DC, USA*, page to appear, 2025.

[11] T. Villmann, M. Kaden, M. Lange, P. Stürmer, and W. Hermann. Precision-recall-optimization in learning vector quantization classifiers for improved medical classification systems. In *Proceedings IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2014), Orlando*, pages 71–77, Los Alamitos, 2014. IEEE Press.

[12] J. Voigt, M. Kaden, L. Reuss, and T. Villmann. Reliable classification learning for medical data analysis using prototype-based models. In L. Rutkowski, R. Scherer, M. Korytkowski, W. Pedrycz, R. Tadeusiewicz, and J.M. Zurada, editors, *Proceedings of the 24th International Conference on Artificial Intelligence and Soft Computing - ICAISC, Zakopane*, volume Part I of *LNCS/LNAI*, page accepted, Cham, 2025. Springer International Publishing, Switzerland.

[13] J. Voigt, S. Saralajew, M. Kaden, K. Bohnsack, L. Reuss, and T. Villmann. Biologically-informed shallow classification learning integrating pathway knowledge. In *Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC'2024) – Bioinformatics Workshop*, volume 1, pages 357–367. SCITEPRESS – Science and Technology Publications, Lda., 2024.

[14] J. Oldenburg, J. Wagner, S. Troschke-Meurer, J. Plietz, L. Kaderali, H. Völzke, M. Nauck, G. Homuth, U. Völker, and S. Simm. XModNN: Explainable modular neural network to identify clinical parameters and disease biomarkers in transcriptomic datasets. *Biomolecules*, 14(12, 1501):1–17, 2024.

[15] L. Breimann. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384, 1995.

[16] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, 58(1):267–288, 1996.

[17] G.H. Mohimani, M. Babaie-Zadeh, and C. Jutten. Fast sparse representation based on smoothed l_0 norm. In M.E. Davies, C.J. James, S.A. Abdallah, and M.D Plumbley, editors, *Proceedings of the International Conference on Independent Component Analysis and Signal Separation (ICA)*, volume 4666 of *Lecture Notes in Computer Science*, pages 389–396, 2007.

[18] M.R. Osborne, B. Presnell, and B.A. Turlach. On the LASSO and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337, 2000.

[19] O. Hanner. On the uniform convexity of L^p and l^p. *Arkiv för Matematik*, 3(19):239–244, 1956.

[20] R. Tibshirani. The LASSO method for variable selection in the Cox model. *Statistics in Medicine*, 16(4):385–395, 1997.

[21] M.M. Day. The spaces L^p with $0 < p < 1$. *Bulletin of the American Mathematical Society*, 46:816–823, 1940.

[22] J. Li, H. Che, and X. Liu. Circuit design and analysis of smoothed l_0 norm approximation for sparse signal reconstruction. *Circuits, Systems, and Signal Processing*, 42:2321–2345, 2022.

[23] 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.

[24] H.A. Elmarakeby, J. Hwang, R. Arafeh, J. Crowdis, S. Gang, D. Liu, S.H. AlDubayan, K. Salari, S. Kregel, C. Richter, T.E. Arnoff, J. Park, W.C. Hahn, and E.M. Van Allen. Biologically informed deep neural network for prostate cancer discovery. *Nature*, 598:348–352, 2021.

[25] B. L. WELCH. The generalization of 'student's' problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35, 01 1947.