Biomedical Applications of Prototype Based Classifiers and Relevance Learning

Michael Biehl

University of Groningen, Johann Bernoulli Institute for Mathematics and Computer Science, P.O. Box 407, 9700 AK Groningen, The Netherlands

Abstract. In this contribution, prototype-based systems and relevance learning are presented and discussed in the context of biomedical data analysis. Learning Vector Quantization and Matrix Relevance Learning serve as the main examples. After introducing basic concepts and related approaches, example applications of Generalized Matrix Relevance Learning are reviewed, including the classification of adrenal tumors based on steroid metabolomics data, the analysis of cytokine expression in the context of Rheumatoid Arthritis, and the prediction of recurrence risk in renal tumors based on gene expression.

Keywords: Prototype-based classification, Learning Vector Quantization, Relevance Learning, biomedical data analysis

1 Introduction

The development of novel technologies for biomedical research and clinical practice have led to an impressive increase of the amount and complexity of electronically available data. Large amounts of potentially high-dimensional data are available from different imaging platforms, genomics, proteomics and other *omics* techniques, or longitudinal studies of large patient cohorts. At the same time there is a clear trend towards personalized medicine in complex diseases such as cancer or heart disorders.

As a consequence, an ever-increasing need for powerful automated data analysis is observed. Machine Learning can provide efficient tools for tasks including problems of unsupervised learning, e.g. in the context of clustering, and supervised learning for classification and diagnosis, regression, risk assessment or outcome prediction.

In biomedical and more general life science applications, it is particularly important that algorithms provide *white box* solutions. For instance, the criteria which determine the outcome of a particular diagnosis system or recommendation scheme, should be transparent to the user. On the one hand, this increases the acceptance of automated systems among practitioners. In basic research, on the other hand, interpretable systems may provide novel insights into the nature of the problem at hand.

Prototype-based classifiers constitute a powerful family of tools for supervised data analysis. These systems are parameterized in terms of class-specific representatives in the original feature space and, therefore, facilitate direct interpretation

of the classifiers. In addition, prototype-based systems can be further enhanced by the data-driven optimization of adaptive distance measures. The framework of relevance learning increases the flexibility of the approaches significantly and can provide important insights into the role of the considered features.

In Sec. 2, the basic concepts of prototype based classification is introduced with emphasis on the framework of Learning Vector Quantization (LVQ). The use of standard and unconventional distances is briefly discussed before relevance learning is introduced in Sec. 2.5. Emphasis is on the so-called Generalized Relevance Matrix LVQ (GMLVQ). Section 3 presents the application of GMLVQ in several relevant biomedical problems, before a brief summary is given in Sec. 4.

2 Distance-based classification and prototypes

Here, a brief review of distance based systems is provided. First, the concepts of Nearest Prototype Classifiers and Learning Vector Quantization (LVQ) are presented in Sec. 2.1 and 2.2. The presentation focusses on their relation to the classical Nearest Neighbor classifier. In Sec. 2.3 examples of non-standard distance measures are briefly discussed. Eventually, adaptive dissimilarities in the framework of relevance learning are introduced in Section 2.4.

2.1 Nearest Prototype Classifiers

Similarity based schemes constitute an important and successful framework for the supervised training of classifiers in machine learning [1–4]. The basic idea of comparing observations with a set of reference data is at the core of the classical Nearest-Neighbor (NN) or, more generally, k-Nearest-Neighbor (kNN) scheme [1–3, 5]. This very popular approach is easy to implement and serves as an important baseline for the evaluation of alternative algorithms.

A given set of P feature vectors and associated class labels

$$\mathbb{D} = \{\mathbf{x}^\mu, y^\mu = y(\mathbf{x}^\mu)\}_{\mu=1}^{P} \quad \text{where } \mathbf{x}^\mu \in \mathbb{R}^N \text{ and } y^\mu \in \{1, 2, \dots C\} \qquad (1)$$

is stored as a reference set. An arbitrary feature vector or *query* $\mathbf{x} \in \mathbb{R}^N$ is then classified according to its similarity to the reference samples: The vector $\mathbf{x}$ is assigned to the class of its Nearest Neighbor in $\mathbb{D}$. Very frequently, the (squared) Euclidean distance with $d(\mathbf{x}, \mathbf{x}^\mu) = (\mathbf{x} - \mathbf{x}^\mu)^2$ is employed for the comparison. The more general kNN classifier determines the majority class membership among the k closest samples. Figure 1 (a) illustrates the concept in terms of the NN-classifier.

While kNN classification is very intuitive and does not require an explicit *training phase*, an essential drawback is obvious: For large data sets $\mathbb{D}$, storage needs are significant and, moreover, computing and sorting all distances $d(\mathbf{x}, \mathbf{x}^\mu)$ becomes costly, even if sophisticated bookkeeping and sorting strategies are employed. Most importantly, NN or kNN classifiers tend to realize very complex decision boundaries which may be subject to over-fitting effects, because all reference samples are taken into account explicitly, cf. Fig. 1(a).

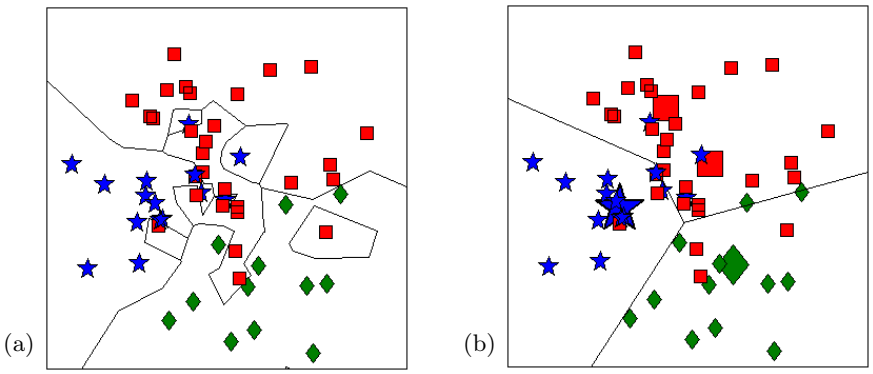


Fig. 1: Illustration of Nearest-Neighbor classification (panel a) and Nearest-Prototype classification in LVQ (panel b). The same two-dimensional data set with three different classes (marked by squares, diamonds and pentagrams) is shown in both panels. Piecewise linear decision boundaries, based on Euclidean distance are shown for the NN classifier in (a), while panel (b) corresponds to an NPC with prototypes marked by large symbols.

These particular difficulties of kNN schemes motivated the idea to replace the complete set of exemplars $\mathbb{D}$ by a few representatives already in [6]. Learning Vector Quantization (LVQ) as a principled approach to the identification of suitable prototypes $\mathbf{w}^k \in \mathbb{R}^N \;\; (k = 1, 2, \ldots K)$ was suggested by Kohonen [7, 8]. The prototypes carry fixed labels $y^k = y(\mathbf{w}^k)$ indicating which class they represent. Obviously, the LVQ system should comprise at least one prototype per class.

Originally, LVQ was motivated as an approximate realization of a Bayes classifier with the prototypes serving as a robust, simplified representation of class-conditional densities [7–9]. Ideally, prototypes constitute typical representatives of the classes, see [10] for a detailed discussion of this property. Recent reviews of prototype based systems in general and LVQ in particular can be found in [9, 11–13].

A Nearest Prototype Classifier (NPC) assigns any feature vector $\mathbf{x}$ to the class $y^* = y(\mathbf{w}^*)$ of the closest prototype $\mathbf{w}^*(\mathbf{x})$, or $\mathbf{w}^*$ for short, which satisfies

$$d(\mathbf{w}^*, \mathbf{x}) \leq d(\mathbf{w}^j, \mathbf{x}) \quad \text{for} \quad j = 1, 2, \ldots K. \tag{2}$$

Assuming that meaningful prototype positions have been determined from a given data set $\mathbb{D}$, an NPC scheme based on Euclidean distance also implements piece-wise linear class boundaries. However, since usually $K \ll P$, these are much smoother than in an NN or kNN scheme and the resulting classifier is less specific to the training data. Moreover, the NPC requires only the computation and ranking of K distances $d(\mathbf{w}^j, \mathbf{x})$. Figure 1 (b) illustrates the NPC scheme

as parameterized by a few prototypes and employing Euclidean distance for the same data set as shown in panel (a).

In binary problems with classes A and B, a bias can be introduced by modifying the NPC scheme: A data point $\mathbf{x}$ is assigned to class A if

$$d(\mathbf{w}^A, \mathbf{x}) \leq d(\mathbf{w}^B, \mathbf{x}) + \Theta \tag{3}$$

and to class B, else. Here, $\mathbf{w}^A$ and $\mathbf{w}^B$ denote the closest prototypes carrying label A or B, respectively. The threshold Θ can be varied from large negative to large positive values, yielding true positive rate (sensitivity) and false positive rate (1-specificity) as functions of Θ. Hence, the full Receiver Operator Characteristics (ROC) can be determined [14].

2.2 Learning Vector Quantizaton

A variety of schemes have been suggested for the iterative identification of LVQ prototypes from a given dataset. Kohonen's basic LVQ1 algorithm [7] already comprises the essential ingredients of most modifications which were suggested later. It is conceptually very similar to unsupervised competitive learning [1] but takes class membership information into account, explicitly.

Upon presentation of a single feature vector $\mathbf{x}^\mu$ with class label $y^\mu = y(\mathbf{x}^\mu)$, the currently closest prototype, i.e. the so-called *winner* $\mathbf{w}^* = \mathbf{w}^*(\mathbf{x}^\mu)$ is identified according to condition (2). The Winner-Takes-All (WTA) update of LVQ1 leaves all other prototypes unchanged:

$$\mathbf{w}^* \leftarrow \mathbf{w}^* + \eta_w \, \Psi(y^*, y^\mu) \, (\mathbf{x}^\mu - \mathbf{w}^*) \quad \text{with} \quad \Psi(y, \tilde{y}) = \begin{cases} +1 \text{ if } y = \tilde{y} \\ -1 \text{ else.} \end{cases} \tag{4}$$

Hence, the winning prototype is moved even closer to $\mathbf{x}^\mu$ if both carry the same class label: $y^* = y^\mu \Rightarrow \Psi = +1$. If the prototype is meant to represent a different class, it is moved further away ($\Psi = -1$) from the feature vector. The learning rate η_w controls the step size of the protoytpe updates.

All examples in $\mathbb{D}$ are presented repeatedly, for instance in random sequential order. A possible initialization is to set prototypes identical to randomly selected feature vectors from their class or close to the class-conditional means.

Several modifications of the basic scheme have been considered in the literature, aiming at better generalization ability or convergence properties, see [11, 15, 16] for examples and further references.

LVQ1 and many other modifications cannot be formulated as the optimization of a suitable objective function in a straightforward way [17]. However, several cost function based LVQ schemes have been proposed in the literature [9, 17, 18]. A popular example is the so–called Generalized Learning Vector Quantization (GLVQ) as introduced by Sato and Yamada [17]. The suggested cost function is given as a sum over all examples in $\mathbb{D}$:

$$E = \sum_{\mu=1}^{P} \Phi(e^\mu) \quad \text{with} \quad e^\mu = \frac{d(\mathbf{w}^J, \mathbf{x}^\mu) - d(\mathbf{w}^K, \mathbf{x}^\mu)}{d(\mathbf{w}^J, \mathbf{x}^\mu) + d(\mathbf{w}^K, \mathbf{x}^\mu)}. \tag{5}$$

For a given $\mathbf{x}^\mu$, $\mathbf{w}^J$ represents the *closest correct* prototype carrying the correct label $y(\mathbf{w}^J) = y^\mu$ and $\mathbf{w}^K$ is the *closest incorrect* prototype with $y(\mathbf{w}^K) \neq y^\mu$, respectively. A monotonically increasing function $\Phi(e^\mu)$ specifies the contribution of a given example in dependence of the respective distances $d(\mathbf{w}^J, \mathbf{x}^\mu)$ and $d(\mathbf{w}^K, \mathbf{x}^\mu)$. Frequent choices are the identity $\Phi(e^\mu) = e^\mu$ and the sigmoidal $\Phi(e^\mu) = 1/[1 + exp(-\gamma\, e^\mu)]$ where $\gamma > 0$ controls the *steepness* [17]. Note that e^μ in Eq. (5) satisfies $-1 \leq e^\mu \leq 1$. The misclassification of a particular sample is indicated by $e^\mu > 0$, while negative e^μ correspond to correctly classified training data. As a consequence, the cost function can be interpreted as to approximate the number of misclassified samples for large γ, i.e. for *steep* Φ.

Since E is differentiable with respect to the prototype components, gradient based methods can be used to minimize the objective function for a given data set in the training phase. The popular *stochastic gradient descent* (SGD) is based on the repeated, random sequential presentation of single examples [1, 3, 19, 20].

The SGD updates of the correct and incorrect winner for a given example $\{\mathbf{x}, y(\mathbf{x})\}$ read

$$\begin{aligned}
\mathbf{w}^J \leftarrow \mathbf{w}^J - \eta_w \frac{\partial}{\partial \mathbf{w}^J}\, \Phi(e) &= \mathbf{w}^J - \eta_w\, \Phi'(e)\, \frac{2 d_K}{(d_J + d_K)^2}\, \frac{\partial d_J}{\partial \mathbf{w}^J}, \\
\mathbf{w}^K \leftarrow \mathbf{w}^K - \eta_w \frac{\partial}{\partial \mathbf{w}^K}\, \Phi(e) &= \mathbf{w}^K + \eta_w\, \Phi'(e)\, \frac{2 d_J}{(d_J + d_K)^2}\, \frac{\partial d_K}{\partial \mathbf{w}^K},
\end{aligned} \tag{6}$$

where the abbreviation $d_L = d(\mathbf{w}^L, \mathbf{x})$ is used. For the squared Euclidean distance we have $\partial d_L / \partial \mathbf{w}^L = -2(\mathbf{x} - \mathbf{w}^L)$. Hence, the displacement of the correct winner is along $+(\mathbf{x} - \mathbf{w}^J)$ and the update of the incorrect winner is along $-(\mathbf{x} - \mathbf{w}^K)$, very similar to the attraction and repulsion in LVQ1. However, in GLVQ, both winners are updated simultaneously.

Theoretical studies of stochastic gradient descent suggest the use of time-dependent learning rates η_w following suitable schedules in order to achieve convergent behavior of the training process, see [19, 20]. for mathematical conditions and example schedules. Alternatively, automated procedures can be employed which adapt the learning rate in the course of training, see for instance [21, 22]. Methods for adaptive step size control have also been devised for batch gradient versions of GLVQ, employing the full gradient in each step, see e.g. [23, 24].

Alternative cost functions have been considered for the training of LVQ systems, see, for instance, [18, 25] for a likelihood based approach. Other objective functions focus on the generative aspect of LVQ [10], or aim at the optimization of the classifier's ROC [26].

2.3 Alternative distances

Although very popular, the use of the standard Euclidean distance is frequently not further justified. It can even lead to inferior performance compared with

problem specific dissimilarity measures which might, for instance, take domain knowledge into account.

A large variety of meaningful measures can be considered to quantify the dissimilarity of N-dim. vectors. Here, we mention only briefly a few important alternatives to Euclidean metrics. A more detailed discussion and further examples can be found in [12, 27, 28], see also references therein.

The family of Minkowski distances of the form

$$d_p(\mathbf{x},\mathbf{y}) = \left(\sum_{j=1}^{N} |x_j - y_j|^p\right)^{1/p} \quad \text{for} \quad \mathbf{x},\mathbf{y} \in I\!R^N \tag{7}$$

provides an important set of alternatives [29]. They fulfill metric properties (for $p \geq 1$) and Euclidean distance is recovered with $p = 2$. Employing Minkowski distances with $p \neq 2$ has proven advantageous in several practical applications, see for instance [30–32].

A different class of more general measures is based on the observation that the Euclidean distance can be written as

$$d_2(\mathbf{x},\mathbf{y}) = \left[(\mathbf{x}\cdot\mathbf{x}) - 2\mathbf{x}\cdot\mathbf{y} + (\mathbf{y}\cdot\mathbf{y})\right]^{1/2}. \tag{8}$$

Replacing inner products of the form $\mathbf{a}\cdot\mathbf{b} = \sum_j a_j b_j$ by a suitable kernel function $\kappa(\mathbf{a},\mathbf{b})$, one obtains so-called kernelized distances [33, 34]. In analogy to the *kernel-trick* used in the Support Vector Machine [33], kernelized distances can be used to implicitly transform non-separable complex data to simpler problems in a higher-dimensional space, see [35] for a discussion in the context of GLVQ.

A very popular dissimilarity measure that takes statistical properties of the data into account explicitly, was suggested very early by Mahalanobis [36]. The *point-wise* version

$$d_M(\mathbf{x},\mathbf{y}) \;=\; \left[(\mathbf{x}-\mathbf{y})^\top C^{-1}\,(\mathbf{x}-\mathbf{y})\right]^{1/2} \tag{9}$$

employs the (empirical) covariance matrix C of the data set for the comparison of two particular feature vectors. The Mahalonobis distance is widely used in the context of the unsupervised and supervised analysis of given data sets, see [2] for a more detailed discussion.

As a last example we mention statistical divergences which can be used when observations are described in terms of densities or histograms. For instance, text can be characterized by *word counts* while color histograms are often used to summarize properties of images. In such cases, the comparison of sample data amounts to evaluating the dissimilarity of histograms. A variety of statistical divergences is suitable for this task [37]. The non-symmetric Kullback-Leibler divergence [2] constitutes a well-known measure for the comparison of densities. An example of a symmetric dissimilarity is the so-called Cauchy-Schwarz divergence [37]:

$$d_{CS}(\mathbf{x},\mathbf{y}) = 1/2\,\log\left[(\mathbf{x}\cdot\mathbf{x})(\mathbf{y}\cdot\mathbf{y})\right] - \log\left[\mathbf{x}\cdot\mathbf{y}\right]. \tag{10}$$

It can be interpreted as a special case of more general γ-divergences, see [37, 38].

In LVQ, meaningful dissimilarities do not have to satisfy metric properties, necessarily. Unlike the kNN approach, LVQ classification does not rely on the pair-wise comparison of data points. A non-symmetric measure $d(\mathbf{w}, \mathbf{x}) \neq d(\mathbf{x}, \mathbf{w})$ can be employed for the comparison of prototypes and data points as long as one version is used consistently in the winner identification, update steps, and the actual classification after training [38].

In cost function based GLVQ, cf. Eq. (5), it is straightforward to replace the squared Euclidean by more general, suitable differentiable measures $d(\mathbf{w}, \mathbf{x})$. Similarly, LVQ1-like updates can be devised by replacing the term $(\mathbf{w} - \mathbf{x})$ in Eq. (4) by $1/2\, \partial d(\mathbf{w}, \mathbf{x})/\partial \mathbf{w}$. Obviously, the winner identification has to make use of the same distance measure in order to be consistent with the update.

It is also possible to extend gradient-based LVQ to non-differentiable distance measures like the *Manhattan distance* with $p = 1$ in Eq. (7), if differentiable approximations are available [29]. Furthermore, the concepts of LVQ can be transferred to more general settings, where data sets do not comprise real-valued feature vectors in an N-dimensional Euclidean space [13]. Methods for classification problems where only pair-wise dissimilarity information is available, can be found in [39, 40], for instance.

2.4 Adaptive distances and relevance learning

The choice of a suitable distance measures constitutes a key step in the design of a prototype-based classifier. It usually requires domain knowledge and insight into the problem at hand. In this context, Relevance Learning constitutes a very elegant and powerful conceptual extension of distance based classification. The idea is to fix only the basic form of the dissimilarity a priori and optimize its parameters in the training phase.

2.5 Generalized Matrix Relevance Learning

As an important example of this strategy we consider here the replacement of standard Euclidean distance by the more general quadratic form

$$d_{\Lambda}(\mathbf{x}, \mathbf{w}) \;=\; (\mathbf{x} - \mathbf{w})^{\top}\; \Lambda\; (\mathbf{x} - \mathbf{w}) \;=\; \sum_{i,j=1}^{N} (x_i - w_i)\, \Lambda_{ij}\, (x_j - w_j). \qquad (11)$$

While the measure is formally reminiscent of the Mahalonobis distance defined in Eq. (9), it is important to note that Λ cannot be directly computed from the data. On the contrary, its elements are considered adaptive parameters in the training process as outlined below.

Note that Euclidean distance is recovered by setting Λ proportional to the N-dim. identity matrix. A restriction to diagonal matrices Λ corresponds to the original formulation of relevance LVQ, which was introduced as RLVQ or GRLVQ in [41] and [42] respectively. There, each feature is weighted by a single adaptive factor in the distance measure.

Measures of the form (11) have been employed in various classification schemes [43–46]. Here we focus on the so-called Generalized Matrix Relevance LVQ (GMLVQ), which was introduced and extended in [47–49]. Applications from the biomedical and other domains are discussed in Sec. 3.

As a minimal requirement, $d_\Lambda(\mathbf{x},\mathbf{w}) \geq 0$ should hold true for all $\mathbf{x},\mathbf{w} \in I\!R^N$. This can be guaranteed by assuming a re-parameterization of the form

$$\Lambda = \Omega^\top \Omega, \quad \text{i.e.} \quad d_\Lambda(\mathbf{x},\mathbf{w}) = \left[\Omega\,(\mathbf{x}-\mathbf{w})\right]^2 \tag{12}$$

with the auxiliary matrix $\Omega \in I\!R^{M\times N}$. It also implies the symmetries $\Lambda_{ij} = \Lambda_{ji}$ and $d_\Lambda(\mathbf{x},\mathbf{w}) = d_\Lambda(\mathbf{w},\mathbf{x})$. Frequently, a normalization $\sum_{ii}\Lambda_{ii} = \sum_{ij}\Omega_{ij}^2 = 1$ is imposed in order to avoid numerical problems.

According to Eq. (12), d_Λ corresponds to conventional Euclidean distance after a linear transformation of all data and prototypes. The transformation matrix can be $(M \times N)$-dimensional, in general, where $M < N$ corresponds to a low-dimensional intrinsic representation of the original feature vectors. Note that, even for $M = N$, the matrix Λ can become singular and d_Λ is only a *pseudo-metric* in $I\!R^N$: for instance, $d_\Lambda(\mathbf{x},\mathbf{y}) = 0$ is possible for $\mathbf{x} \neq \mathbf{y}$.

In the training process, all elements of the matrix Ω are considered adaptive quantities. From Eqs. (12) we obtain the derivatives

$$\frac{\partial d_\Lambda(\mathbf{w},\mathbf{x})}{\partial \mathbf{w}} = \Omega^\top\Omega\,(\mathbf{w}-\mathbf{x}), \qquad \frac{\partial d_\Lambda(\mathbf{w},\mathbf{x})}{\partial \Omega} = \Omega\,(\mathbf{w}-\mathbf{x})(\mathbf{w}-\mathbf{x})^\top \tag{13}$$

which can be used to construct heuristic updates along the lines of LVQ1 [12, 13, 50]. From the GLVQ cost function, cf. Eq. (5), one obtains the matrix update

$$\Omega \leftarrow \Omega - \eta_\Omega\,\Phi'(e)\left(\frac{2d_\Lambda^K}{(d_\Lambda^J+d_\Lambda^K)^2}\frac{\partial\, d_\Lambda(\mathbf{w}^J,\mathbf{x})}{\partial\Omega} - \frac{2d_\Lambda^J}{(d_J+d_\Lambda^K)^2}\frac{\partial\, d_\Lambda(\mathbf{w}^K,\mathbf{x})}{\partial\Omega}\right) \tag{14}$$

which can be followed by a normalization step achieving $\sum_{ij}\Omega^2 = 1$. Prototypes are updated as given in Eq. (6) with the gradient terms replaced according to Eq. (13). The matrix learning rate is frequently chosen smaller than that of the prototype updates: $\eta_\Omega < \eta_w$, details can be found in [47, 49]. The matrix $\Omega \in I\!R^{M\times N}$ can be initialized by, for instance, drawing independent random elements or by setting it proportional to the N-dim. identity matrix for $M = N$.

In the measure (11), the diagonal elements of Λ quantify the weight of single features in the distance. The inspection of the relevance matrix can provide valuable insights into the structure of the data set after training, examples are discussed in Sec. 3. Off-diagonal elements correspond to the contribution of pairs of features to d_Λ and their adapation enables the system to cope with correlations and dependencies between the features. Note that this heuristic interpretation of Λ is only justified if all features are of the same order of magnitude, strictly speaking. In any given data set, this can be achieved by applying a z-score transformation, yielding zero mean and unit variance features. Alternatively, potentially different magnitudes of the features could be taken into account after training by rescaling the elements of Λ accordingly.

2.6 Related schemes and variants of GMLVQ

Adaptive distance measures of the form (11) have been considered in several realizations of distance based classifiers. For example, Weinberger et *al.* optimize a quadratic form in the context of nearest neighbor classification [43, 44]. An explicit construction of a relevance matrix from a given data set is suggested and discussed in [45], while the gradient based optimization of an alternative cost function is presented in [46].

Localized versions of the distance (11) have been considered in [44, 47, 49]. In GMLVQ, it is possible to assign an individual relevance matrix Λ^j to each $\mathbf{w}^j$ or to devise class-wise matrices. Details and the corresponding modified update rules can be found in [47, 49]. While this can enhance the classification performance significantly in complex problems, we restrict the discussion to the simplest case of one global measure of the form (11).

The GMLVQ algorithm displays an intrinsic tendency to yield singular relevance matrices which are dominated by a few eigenvectors corresponding to the leading eigenvalues. This effect has been observed empirically in real world applications and benchmark data sets, see [13, 47] for examples. Moreover, a mathematical investigation of stationarity conditions explains this typical property of GMLVQ systems [50]. Very often, the effect allows for an interpretable visualization of the labeled data set in terms of projections onto two or three leading eigenvectors [13, 47, 51].

An explicit *rank control* can be achieved by using a rectangular $(M \times N)$ matrix Ω in the re-parameterization (12), together with the incorporation of a penalty term for $\mathrm{rank}(\Lambda) < M$ in the cost function [49, 48]. For $M = 2$ or 3, the approach can also be used for the discriminative visualization of labelled data sets [51].

An important alternative to the intrinsic dimension reduction provided by GMLVQ is the identification of a suitable linear projection in a pre-processing step. This can be advantageous, in particular for nominally very high-dimensional data as encountered in e.g. bioinformatics, or in situations where the number of training samples P is smaller than the dimension of the feature space. Assuming that a given projection of the form

$$\mathbf{y} = \Psi\,\mathbf{x}, \quad \mathbf{v} = \Psi\,\mathbf{w} \;\; \text{with} \;\; \Psi \in I\!R^{M\times N} \tag{15}$$

maps N-dim. feature vectors and prototypes to their M-dim. representations we can re-write the distance measure of the form (11) as

$$(\mathbf{x}-\mathbf{w})^\top\; \Lambda\,(\mathbf{x}-\mathbf{w}) \;=\; (\mathbf{x}-\mathbf{w})^\top\; \Psi^\top\, \widetilde{\Lambda}\, \Psi\,(\mathbf{x}-\mathbf{w}) \;=\; (\mathbf{y}-\mathbf{v})^\top\;\; \widetilde{\Lambda}\;(\mathbf{y}-\mathbf{v}). \tag{16}$$

Hence, training and classification can be formulated in the M-dimensional space, employing prototypes $\mathbf{v}^j \in I\!R^M$ and an $M \times M$ relevance matrix $\widetilde{\Lambda}$. Moreover, the relation $\Lambda = \Psi^\top \widetilde{\Lambda}\, \Psi$ facilitates its interpretation in the original feature space.

This versatile framework allows to combine GMLVQ with, for instance, Principal Component Analysis (PCA) [2] or other linear projection techniques. Furthermore, it can be applied to the classification of funtional data, where the

components of the feature vectors represent an ordered sequence of values rather than a collection of more or less independent quantities. This is the case in, for instance, time series data or spectra obtained from organic samples, see [52] for examples and further references. The coefficients of a, for instance, polynomial approximation of observed data are typically obtained by a linear transformation of the form (15), where the rows of Ψ represent the basis functions. Hence, training can be performed in the lower-dimensional coefficient space, while the resulting classifier is still interpretable in terms of the original features [52].

3 Biomedical applications of GMLVQ

In the following, selected bio-medical applications of the GMLVQ approach are highlighted. The example problems illustrate the flexibility of the approach and range from the direct analysis of relatively low-dim. data in steroid metabolomics (Sec. 3.1), the combination of relevance learning with dimension reduction for cytokine data (Sec. 3.2), and the application of GMLVQ to selected gene expression data in the context of tumor recurrence prediction (Sec. 3.3). A brief discussion with emphasis on the interpretability of the relevance matrix. Evenutally, further applications of GMLVQ for biomedical and life science data are briefly mentioned in Sec. 3.4.

3.1 Steroid Metabolomics in Endocrinology

A variety of disorders can affect the human endocrine system. For instance, tumors of the adrenal glands are relatively frequent and often found incidentally [53, 54]. The adrenals produce a number of steroid hormones which regulate important body functions. The differential diagnosis of malignant Adrenocortical Carcinoma (ACC) vs. benign Adenoma (ACA) based on non-invasive methods constitutes a highly relevant diagnostic challenge [53]. In [54], Arlt et *al.* explore the possibility to detect malignancy on the basis of the patient's steroid excretion pattern obtained from 24h urine samples by means of gas chromatography/mass spectrometry (GC/MS).

The analysis of data comprising the excretion of 32 steroids and steroid metabolites was presented in [54] and [55]: A data set representing a study population of 102 ACA and 45 ACC samples was analysed by means of training a GMLVQ system with one prototype per class and a single, global relevance matrix $\Lambda \in I\!R^{32\times 32}$. In a pre-processing step, excretion values were log-transformed and in every individual training process a z-score transformation was applied.

In order to estimate the classification performance with respect to novel data representing patients with unknown diagnosis, random splits of the data set were considered: about 90% of the samples were used for training, while 10% served as a validation set. Results were obtained on average over 1000 randomized splits, yielding, for instance the threshold-averaged ROC [14], see Eq. (3).

A comparison of three scenarios provides evidence for the benefitial effect of relevance learning: When applying Euclidean GLVQ, the classifier achieves

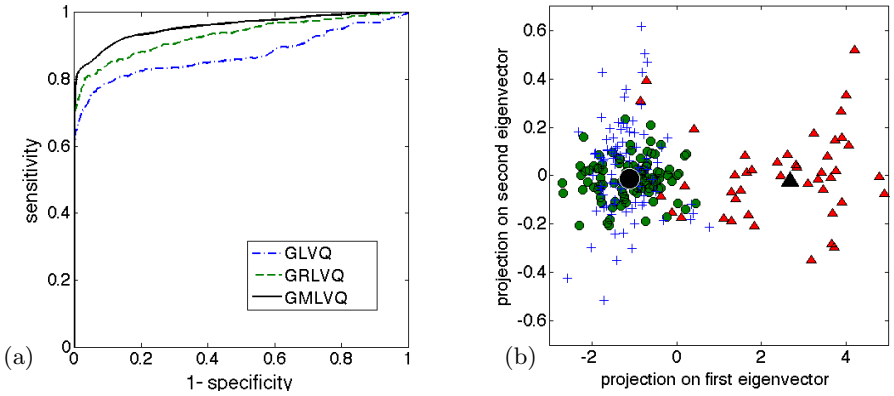


Fig. 2: Detection of malignancy in adrenocortical tumors, see Sec. 3.1.
Panel (a): Test set ROC as obtained in the randomized validation procedure by applying GLVQ with Euclidean distance (dash-dotted line), GRLVQ with diagonal Λ (dashed) and GMLVQ with a full relevance matrix (solid).
Panel (b): Visualization of the data set based on the GMLVQ analysis in terms of the projection of steroid profiles on the leading eigenvectors of Λ. Circles correspond to patients with benign ACA while triangles mark malignant ACC. Prototypes are marked by larger symbols. In addition, healthy controls (not used in the analysis) are displayed as crosses.

an ROC with an *area under the curve* of $AUC \approx 0.87$, see Fig. 2 (a). The consideration of an adaptive diagonal relevance matrix, corresponding to GRLVQ [42], yields an improved performance with $AUC \approx 0.93$. The GMLVQ approach, cf. Sec. 2.5, with a fully adaptive relevance matrix achieves an AUC of about 0.97. In the latter case, a working point with equal sensitivity and specificity of 0.90 can be selected by proper choice of the threshold Θ in Eq. (3). As reported in [54], the GMLVQ system outperformed alternative classifiers of comparable complexity.

The resulting relevance matrix Λ turned out to be dominated by the leading eigenvector corresponding to its largest eigenvalue; subsequent eigenvalues are found to be significantly smaller. As discussed above, this property can be exploited for the discriminative visualization of the data set and prototypes, see Fig. 2 (b). The figure displays, in addition, a set of feature vectors representing healthy controls, which were not explicitly considered in the training process. Reassuringly, control samples cluster close to the ACA prototype and appear clearly separated from the malignant ACC.

By inspecting the relevance matrix of the trained system, further insight into the problem and data set can be achieved. Figure 3 (a) displays the diagonal elements of Λ on average over 1000 randomized training runs. Subsets of markers can be identified, which are consistently rated as particularly important for

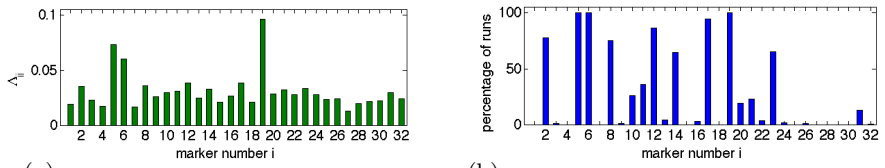


Fig. 3: Relevance of steroid markers in adrenal tumor classification, see Sec. 3.1 for details. **Panel (a):** Diagonal elements Λ_{ii} of the GMLVQ relevance matrix on average over the 1000 randomized training runs. **Panel (b):** Percentage of training runs in which a particular steroid appeared among the 9 most relevant markers.

the classification. For instance, markers $5, 6$ and 19 appear significantly more relevant than all others, see [54] for a detailed discussion from the endocrinological perspective. There, the authors suggest a panel of nine leading steroids, which could serve as a reduced marker set in a practical realization of the diagnosis tool. Figure 3 (b) displays the fraction of training runs in which a single marker is rated among the nine most relevant ones, providing further support for the selection of the subset [54]. Repeating the GMLVQ training for selected subsets of leading markers yielded slightly inferior performance compared to the full panel of 32 markers, with $AUC \approx 0.96$ for nine steroids, and $AUC \approx 0.94$ with 3 leading markers only, see [54] for details of the analysis.

The analysis of steroid metabolomics data by means of GMLVQ and related techniques is currently explored in the context of various disorders, see [56–58] for recent examples. In the context of adrenocortical tumors, the validation of the diagnostic approach in prospective studies and the development of efficient methods for the detection of post-operative recurrence are in the center of interest [59].

3.2 Cytokine Markers in Inflammatory Diseases

Rheumatoid Arthritis (RA) constitutes an important example of chronic inflammatory disease. It is the most common form of autoimmune arthritis with symptoms ranging from stiffness and swelling of joints to, in the long term, bone erosion and joint deformity.

Cytokines play an important role in the regulation of inflammatory processes. Yeo et *al.* [60] investigated the role of 117 cytokines in early stages of RA. Their mRNA expression was determined by means of PCR techniques for four different patient groups: Uninflamed healthy controls (group A, 9 samples), patients with joint inflammations that resolved within 18 months after symptom onset (group B, 9 samples), *early RA* patients developing Rheumatoid Arthritis in this period of time (group C, 17 samples), and patients with an established diagnosis of RA (group D, 12 samples).

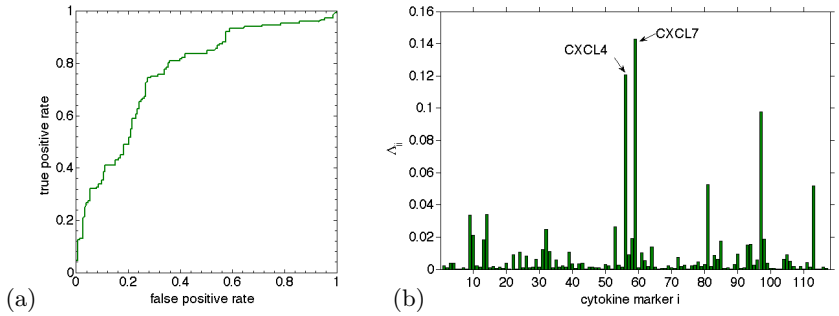


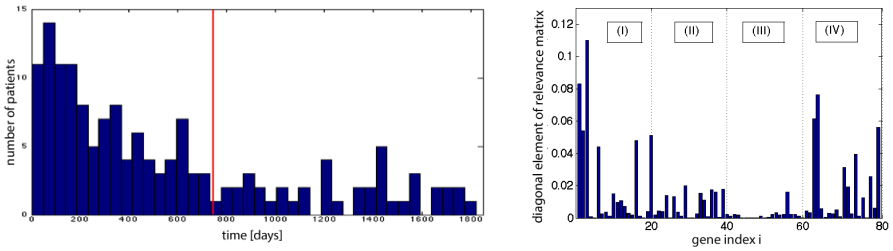
Fig. 4: GMLVQ analysis of Rheumatoid Arthritis data, see Sec. 3.2 for details. Discrimination of patients with early RA (class B) vs. resolving cases (class C). **Panel (a)** shows the ROC ($AUC \approx 0.763$) as obtained in the Leave-One-Out (from each class) validation. **Panel (b)** displays the diagonal elements of the back-transformed relevance matrix $\Lambda \in I\!R^{117\times 117}$ on average over the validation runs.

Note that the total number of samples is small compared to the dimension $N = 117$ of the feature vectors $\mathbf{x}$ comprising log-transformed RNA expression values. Hence, standard PCA was applied to identify a suitable low-dimensional representation of the data. The analysis revealed that 95% of the variation in the data set was explained by the 21 leading principal components already. Attributing the remaining 5% mainly to *noise*, all cytokine expressions data were represented in terms of $M = 21$-dim. feature vectors corresponding to the $\mathbf{y} \in I\!R^M$ in Eq. (15).

GMLVQ was applied to two classification subproblems: The first addressed the discrimination of healthy controls (class A) and established RA patients (class D). While this problem does not constitute a diagnostic challenge at all, it served as a consistency check and revealed first insights into the role of cytokine markers. In the second setting, the much more difficult problem of discriminating early stage RA (class C) from resolving cases (class B) was considered.

The performances of the respective classifier systems were evaluated in a validation procedure by leaving out one sample from each class for testing and training on the remaining data. Results were reported on average over all possible test set configurations. Reassuringly, the validation set ROC obtained for the classification of A vs. D displayed almost error free performance with $AUC \approx 0.996$. The expected greater difficulty of discriminating patient groups C and D was reflected in a lower AUC of approximately 0.763, see Fig. 4 (a).

It is important to note that it was not the main aim of the investigation to propose a practical diagnosis tool for the early detection of Rheumatoid Arthritis. As much as an early diagnosis would be desirable, the limited size of the study population would not provide enough supporting evidence for such a suggestion. However, the GMLVQ analysis revealed important and surprising insights into



(a) (b)

Fig. 5: Recurrence risk prediction in ccRCC, see Sec. 3.3 for details.
Panel (a): Number of recurrences registered in the 469 patients vs. time in days. The vertical line marks a threshold of 24 months, before which 109 patients developed a recurrence. **Panel (b):** Diagonal entries of the relevance matrix with respect to the discrimination of low risk vs. high risk patients from the expression of the 80 selected genes.

the role of cytokines. Computing the back-transformed relevance matrix $\Lambda \in \mathbb{R}^{117\times117}$ with respect to the original cytokine expression features along the lines of Eq. (16), makes possible an evaluation of their significance in the respective classification problem. Figure 4 (b) displays the cytokine relevances as obtained in the discrimination of classes B and C. Two cytokines, CXCL4 and CXCL7, were identified as clearly dominating in terms of their discriminative power. A discussion of further relevant cytokines also with respect to the differences between the two classification problems can be found in [60].

The main result of the machine learning analysis triggered additional investigations by means of a direct inspection of synovial tissue samples. Careful studies employing staining techniqes confirmed that CXCL4 and CXCL7 play an important role in the early stages of RA [60]. Significantly increased expression of CXCL4 and CXCL7 was confirmed in early RA patients compared with those with resolving arthritis or with clearly established disease. The study showed that the two cytokines co-localize, in particular, with extravascular macrophages in early stage Rheumatoid Arthritis. Implications for future research into the onset and progression of RA are also discussed in [60].

3.3 Recurrence Risk Prediction in Clear Cell Renal Cell Carcinoma

Mukherjee et al. [62] investigated the use of mRNA-Seq expression data to evaluate recurrence risk in clear cell Renal Cell Carcinoma (ccRCC). The corresponding data set is publicly available from *The Cancer Genome Atlas* (TCGA) repository [63] and is also hosted at the Broad Institute (http://gdac.broadinstitute.org). It comprises mRNA-Seq data (raw and RPKM normalized) for 20532 genes, accompanied by clinical data for survival and recurrences for 469 tumor samples.

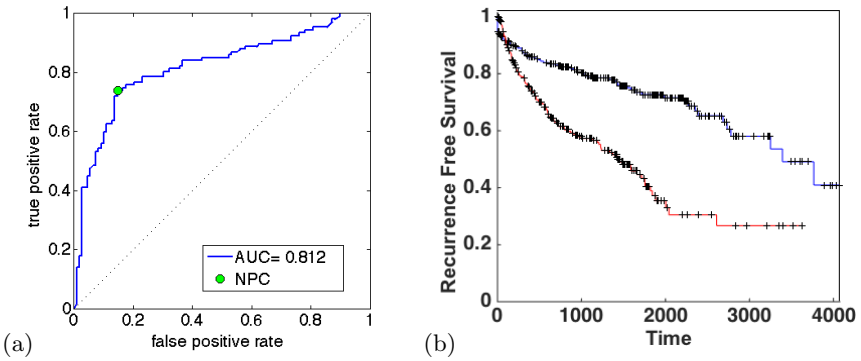


Fig. 6: Recurrence risk prediction in ccRCC, see Sec. 3.3 for details.
Panel (a): ROC for the classification of low-risk (no or late recurrence) vs. high risk (early recurrence) as obtained in the Leave-One-Out validation of the GMLVQ classifier trained on the subset of 216 patients, cf. Sec. 3.3. The circle marks the performance of the Nearest Prototype Classifier. **Panel (b):** Kaplan-Meier plot [61] showing recurrence free survival rates in the high-risk (lower curve) and low-risk (upper curve) group as classified by the GMLVQ system applied to all 469 samples. Time is given in days.

Preprocessing steps, including normalization, log-transformation, and median centering, are described in [62].

By means of an outlier analysis [64], a drastically reduced panel of 80 genes was identified for further use, see also [62] for a description of the method in this particular example. The panel consists of four different groups, each comprising 20 selected genes: In group (I), high expression can be correlated with low risk, i.e. late or no recurrence. In group (II), however, low expression is associated with low risk. Group (III) contains genes where high expression is correlated with a high risk for early recurrence, while in group (IV) low expression of the genes is an indication of high risk.

In [62], a risk index is presented, which is based on a voting scheme with respect to the 80 selected genes. Here, the focus is on the further analysis of the corresponding expression values using GMLVQ, also discussed in [62].

In order to define a meaningful classification problem, two extreme groups of patients were considered: group A with poor prognosis / high risk, comprises 109 patients with recurrence within the first 24 months after the initial diagnosis. Group B corresponds to 107 patients with favorable prognosis / low risk, who did not develop tumor recurrence within 60 months after diagnosis. The frequency of recurrence times observed over five years in the complete set of 469 patients is shown in Figure 5 (a), the vertical line marks the threshold of two years after diagnosis.

A GMLVQ system with one 80-dim. prototype per class (A, B) and a global relevance matrix $\Lambda \in \mathbb{R}^{80\times 80}$ was trained on the subset of the 216 clear-cut cases

in groups A and B. Leave-One-Out validation yielded the averaged ROC shown in Fig. 6 (a) with $AUC \approx 0.812$.

The diagonal elements of the averaged relevance matrix are displayed in Fig. 5 (b). The results show that genes in the groups (I) and (IV) seem to be particularly discriminative and suggest that a further reduction of the gene panel should be well possible [62].

In order to further evaluate the GMLVQ classifier, it was employed to assign all 469 samples in the data set to the groups of high risk or low risk patients, respectively. In case of the 216 cases with early recurrence (≤ 24 months) or no recurrence within 60 months, the Leave-One-Out prediction was used. For the remaining 253 patients, the GMLVQ classifier obtained from the 216 reference samples was used.

In Fig. 6 the resulting Kaplan-Meier plot [61] is shown. It displays the recurrence free survival rate of the low risk (upper) and high risk (lower) groups according to GMLVQ classification, corresponding to a pronounced discrimination of the groups with log-rank p-value 1.2×10^{-8}.

In summary, the work presented in [62] shows that gene expression data makes possible an efficient risk assessment with respect to tumor recurrence. Further analysis, taking into account healthy cell samples as well, shows that the panel of genes is not only prognostic but also diagnostic [62].

3.4 Further bio-medical and life science applications

Apart from the studies discussed in the previous sections, variants of LVQ have been employed successfully in a variety of biomedical and life science applications. In the following, a few more examples are briefly mentioned and references are provided for the interested reader.

An LVQ1-like classifier was employed for the identification of exonic vs. intronic regions in the genome of C. Elegans based on features derived from sequence data [30]. In this application, the use of the Manhattan distance in combination with heuristic relevance learning proved advantageous.

Simple LVQ1 with Euclidean distance measure was employed successfully in the inter-species prediction of protein phosphorylation in the sbv IMPROVER challenge [65]. There, the goal was to predict the effect of chemical stimuli on human lung cells, given information about the reaction of rodent cells under the same conditions.

The detection and discrimination of viral crop plant diseases, based on color and shape features derived from photographic images was studied in [38]. The authors applied divergence-based LVQ, cf. Sec. 2.3, for the comparison of feature histograms derived from Cassava plant leaf images. A comparison with alternative approaches, including GMLVQ is presented in [66].

The analysis of flow-cytometry data was considered in [67] in the context of the DREAM6/FlowCAP2 challenge [68]. For each subject, 31 markers were provided, including measures of cell size and intracellular granularity as well as 29 expression values of surface proteins for thousands of individual cells. Hand-crafted features were determined in terms of statistical moments over the entire

cell population, yielding a 186-dim. representation for each patient. GMLVQ applied in this feature space yielded error-free prediction of AML in the test set [67, 68].

The detection and discrimination of different Parkinsonian syndromes was addressed in [69, 70]. Three-dimensional brain images obtained by fluorodeoxyglucose positron emission tomography (FDG-PET) comprise several hundreds of thousands voxels per subject, providing information about the local glucose metabolism. An appropriate dimension reduction by *Scaled Subprofile Model with Principal Component Analysis* (SSM/PCA), yields a data set dependent, low-dimensional representation in terms of subject scores, see [69, 70] for further references. In comparison with Decision Trees and Support Vector Machines, the GMLVQ classifier displayed competetive or superior performance [70].

4 Concluding remarks

This contribution merely serves as a starting point for studies into the application of prototype and distance based classification in the biomedical domain. It provides by no means a complete overview and focusses on the example framework of Generalized Matrix Relevance Learning Vector Quantization, which has been applied to a variety of life science datasets. The specific application examples were selected in order to demonstrate the flexibility of the approach and illustrate its interpretability.

A number of open questions and challenges deserve attention in future research – to name only a few examples: A better understanding of feature relevances should be obtained, for instance, by exploiting the approaches presented in [71]. Combined distance measures can be designed for the treatment of different sources of information in an integrative manner [72]. The analysis of functional data plays a role of increasing importance in the biomedical domain, see e.g. [52]. In general, the development of efficient methods for the analysis of biomedical data, which are at the same time powerful and transparent, constitutes a major challenge of great importance. Prototype based classifiers will continue to play a central role in this context.

Acknowledgments

The author would like to thank the collaboration partners and co-authors of the publications which are reviewed in this contribution or could be mentioned only briefly.

References

1. C.M. Bishop. *Pattern Recognition and Machine Learning.* Cambridge University Press, Cambridge, UK, 2007.
2. D.G. Stork R.O. Duda, P.E. Hart. *Pattern Classification.* John Wiley & Sons, Hoboken, NJ, 2001.

3. T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, second edition, 2009.
4. M. Biehl, B. Hammer, M. Verleysen, and T. Villmann, editors. *Similarity based clustering - recent developments and biomedical applications*, volume 5400 of *Lecture Notes in Artificial Intelligence*. Springer, 2009. 201 pages.
5. P.E. Hart T.M. Cover. Nearest neighbor pattern classification. *IEEE Trans. Information Theory*, 13:21–27, 1967.
6. P.E. Hart. The condensed nearest neighbor rule. *Information Theory, IEEE Transactions on*, 14:515–516, 1968.
7. T. Kohonen. Learning vector quantization for pattern recognition. Technical Report TKK-F-A601, Helsinki University of Technology, Espoo, Finland, 1986.
8. T. Kohonen. *Self-Organizing Maps*. Springer, Heidelberg, 1997.
9. Bohnsack A. Kaden M. Villmann, T. Can Learning Vector Quantization be an alternative to SVM and Deep Learning? - Recent trends and advanced variants of Learning Vector Quantization for classification learning. *J. of Artificial Intelligence and Soft Computing Research*, 7:65–81, 2017.
10. M. Riedel T. Villmann B. Hammer, D. Nebel. Generative versus discriminative prototype based classification. In M. Kaden M. Lange T. Villmann, F.-M. Schleif, editor, *Advances in Self-Organizing Maps and Learning Vector Quantization: Proc. of the 10th Intl. Workshop WSOM 2014*, pages 123–132, Cham, 2014. Springer.
11. D. Nova and P.A. Estévez. A review of learning vector quantization classifiers. *Neural Computing and Applications*, 25(3-4):511–524, 2014.
12. M. Biehl, B. Hammer, and T. Villmann. Distance measures for prototype based classfication. In L. Grandinetti, N. Petkov, and T. Lippert, editors, *BrainComp 2013, Proc. International Workshop on Brain-Inspired Computing, Cetraro/Italy, 2013*, volume 8603 of *Lecture Notes in Computer Science*, pages 100–116. Springer, 2014.
13. T. Villmann M. Biehl, B. Hammer. Prototype-based models in machine learning. *Wileys Interdisicplinary Reviews (Wires) Cognitive Science*, 7.
14. T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861–874, 2006.
15. T. Kohonen. Improved versions of Learning Vector Quantization. In *International Joint Conference on Neural Networks*, volume 1, pages 545–550, 1990.
16. M. Biehl, A. Ghosh, and B. Hammer. Dynamics and generalization ability of LVQ algorithms. *Journal of Machine Learning Research*, 8:323–360, 2007.
17. A. S. Sato and K. Yamada. Generalized Learning Vector Quantization. In M. C. Mozer D. S. Touretzky and M. E. Hasselmo, editors, *Proc. Neural Information Processing Systems (NIPS)*, volume 8, pages 423–429, Cambridge, MA, USA, 1996. MIT Press.
18. K. Obermayer S. Seo. Soft nearest prototype classification. *IEEE Trans. on Neural Networks*, 14:390–398, 2003.
19. L. Bottou. Online algorithms and stochastic approximations. In D. Saad, editor, *Online Learning and Neural Networks*, pages 9–42. Cambridge University Press, Cambridge, UK, 1998.
20. H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22:405, 1951.
21. Y. LeCun T. Schaul, S. Zhang. No More Pesky Learning Rates. *JMLR: W&CP*, 28:342–351, 2013.
22. Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2014.

23. G. Papari, K. Bunte, and M. Biehl. Waypoint averaging and step size control in learning by gradient descent (technical report). In F.-M. Schleif and T. Villmann, editors, *MIWOCI 2011, Mittweida Workshop on Computational Intelligence*, volume MLR-2011-06 of *Machine Learning Reports*, pages 16–26. Univ. of Bielefeld, 2011.
24. M. Biehl. GMLVQ demo code, website: http://www.cs.rug.nl/˜biehl (last visited: 16 march 2017), 2015.
25. K. Obermayer S. Seo. Soft Learning Vector Quantization. *Neural Computation*, 15:1589–1604, 2003.
26. T Villmann, M Kaden, W Hermann, and M Biehl. Learning Vector Quantization classifiers for ROC-optimization. *Computational Statistics*, pages 1–22, 2016. published online.
27. B. Hammer and T. Villmann. Classification using non-standard metrics. In M. Verleysen, editor, *European Symposium on Artificial Neural Networks, ESANN 2005*, pages 303–316. d-side publishing, 2005.
28. 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, *Advances in Neural Information Paradigms*, volume 247 of *Springer Studies in Computational Intelligence*, pages 183–199. Springer, 2010.
29. M. Lange and T. Villmann. Derivatives of Lp-norms and their approximations. *Machine Learning Reports*, MLR-03-2013, 2013.
30. M. Biehl, R. Breitling, and Y. Li. Analysis of Tiling Microarray Data by Learning Vector Quantization and Relevance Learning. In H. Yin, P. Tino, E. Corchado, W. Byrne, and X. Yao, editors, *Proc. Intelligent Data Engineering and Automated Learning, IDEAL 2007*, volume 4881 of *Lecture Notes in Computer Science*, pages 880–889. Springer, 2007.
31. O. Golubitsky and S.M. Watt. Distance-based classification of handwritten symbols. *International Journal on Document Analysis and Recognition (IJDAR)*, 13(2):133–146, 2010.
32. T. Villmann, M. Kästner, D. Nebel, and M. Riedel. ICMLA face recognition challenge – results of the team 'Computational Intelligence Mittweida'. In *Proc. of the International Conference on Machine Learning Applications (ICMLA'12)*, pages 7–10. IEEE Computer Society Press, 2012.
33. J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004. 474 pages.
34. B. Schölkopf. The kernel trick for distances. *Advances in Neural Information Processing Systems*, 13:301–307, 2001.
35. F.-M. Schleif, T. Villmann, B. Hammer, P. Schneider, and M. Biehl. Generalized Derivative Based Kernelized Learning Vector Quantization. In *IDEAL*, pages 21–28, 2010.
36. P.C. Mahalanobis. On the generalised distance in statistics. *Proc. of the National Inst. of Sciences of India*, 2(1):49–55, 1936.
37. A. Cichocki, R. Zdunek, A. Phan, and S.-I. Amari. *Nonnegative Matrix and Tensor Factorizations*. Wiley, 2009.
38. E. Mwebaze, P. Schneider, F.-M. Schleif, J.R. Aduwo, J.A. Quinn, S. Haase, T. Villmann, and M. Biehl. Divergence based classification in Learning Vector Quantization. *Neural Computation*, 74(9):1429–1435, April 2011.
39. B. Hammer, F.-M. Schleif, and X. Zhu. Relational extensions of Learning Vector Quantization. In B.-L. Lu, L. Zhang, and J. Kwok, editors, *Neural Information Processing*, volume 7063 of *Lecture Notes in Computer Science*, pages 481–489. Springer Berlin Heidelberg, 2011.

40. D. Nebel, B. Hammer, and T. Villmann. A median variant of generalized learning vector quantization. In *International Conference on Neural Information Processing*, pages 19–26. Springer Berlin Heidelberg, 2013.
41. T. Bojer, B. Hammer, D. Schunk, and K. Tluk von Toschanowitz. Relevance determination in Learning Vector Quantization. In M. Verleysen, editor, *European Symposium on Artificial Neural Networks*, pages 271–276, 2001.
42. B. Hammer and T. Villmann. Generalized Relevance Learning Vector Quantization. *Neural Networks*, 15(8-9):1059–1068, 2002.
43. K.Q. Weinberger, J. Blitzer, and L. Saul. Distance Metric Learning for Large Margin Nearest Neighbor Classification. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1473–1480. MIT Press, Cambridge, MA, 2006.
44. K.Q. Weinberger and L.K. Saul. Distance metric learning for Large Margin Nearest Neighbor classification. *The Journal of Machine Learning Research*, 10:207–244, 2009.
45. M. Boareto, J. Cesar, V.B.P. Leite, and N. Caticha. Supervised Variational Relevance Learning, an analytic geometric feature selection with applications to omic data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(99):705–711, 2015.
46. J. Hocke and T. Martinetz. Global metric learning by gradient descent. In *Artificial Neural Networks and Machine Learning–ICANN 2014*, pages 129–135. Springer, 2014.
47. P. Schneider, M. Biehl, and B. Hammer. Adaptive relevance matrices in learning vector quantization. *Neural Computation*, 21:3532–3561, 2009.
48. P. Schneider, K. Bunte, H. Stiekema, B. Hammer, T. Villmann, and M. Biehl. Regularization in Matrix Relevance Learning. *IEEE Transactions on Neural Networks*, 21:831–840, 2010.
49. K. Bunte, P. Schneider, B. Hammer, F.-M. Schleif, T. Villmann, and M. Biehl. Limited Rank Matrix Learning, discriminative dimension reduction, and visualization. *Neural Networks*, 26:159–173, 2012.
50. M. Biehl, B. Hammer, F. M. Schleif, P. Schneider, and T. Villmann. Stationarity of Matrix Relevance LVQ. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, July 2015.
51. M. Biehl, K. Bunte, F. M. Schleif, P. Schneider, and T. Villmann. Large margin linear discriminative visualization by matrix relevance learning. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, June 2012.
52. F. Melchert, U. Seiffert, and M. Biehl. Functional Representation of Prototypes in LVQ and Relevance Learning. In E. Merényi, M.J. Mendenhall, and P. O'Driscoll, editors, *Advances in Self-Organizing Maps and Learning Vector Quantization: Proceedings of the 11th International Workshop WSOM 2016*, pages 317–327, Cham, 2016. Springer International Publishing.
53. European Network for the Study of Adrenal Tumours. ENS@T website: http://www.ensat.org (last visited: 16 march 2017), 2002.
54. 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.

55. M. Biehl, P. Schneider, D. Smith, H. Stiekema, A. Taylor, B. Hughes, C. Shackleton, P. Stewart, and W. Arlt. Matrix relevance LVQ in steroid metabolomics based classification of adrenal tumors. In M. Verleysen, editor, *20th European Symposium on Artificial Neural Networks (ESANN 2012)*, pages 423–428. d-side publishing, 2012.
56. S. Ghosh, E. Baranowski, R. van Veen, G. de Vries, M. Biehl, W. Arlt, P. Tino, and K. Bunte. Comparison of strategies to learn from imbalanced classes for computer aided diagnosis of inborn steroidogenic disorders. In M. Verleysen, editor, *25th European Symposium on Artificial Neural Networks (ESANN 2017)*. d-side publishing, 2017. in press.
57. A. Moolla, A. Amin, B. Hughes, W. Arlt, Z. Hassan-Smith, M. Armstrong, P. Newsome, T. Shah, L. Van Gaal, A. Verrijken, S. Francque, M. Biehl, and J. Tomlinson. The urinary steroid metabolome as a non-invasive tool to stage non-alcoholic fatty liver disease. *Endocrine Abstracts*, 44:OC1.4, 2016.
58. K. Lang, F. Beuschlein, M. Biehl, A. Dietz, A. Riester, B.A. Hughes, D.M. O'Neil, S. Hahner, M. Quinkler, J.W. Lenders, C. Shackleton, M. Reincke, and W. Arlt. Urine steroid metabolomics as a diagnostic tool in primary aldosteronism. Presented at BES 2015, Edinburgh, UK. *Endocrine Abstracts*, 38:OC1–6, 2015.
59. V. Chortis, I. Bancos, A.J. Sitch, A.E. Taylor, D. O'Neil, K. Lang, M. Quinkler, M. Terzolo, M. Manelli, D. Vassiliadi, U. Ambroziak, M. Conall Dennedy, M. Sherlock, J. Bertherat, F. Beuschlein, M. Fassnacht, J. Deeks, M. Biehl, and W. Arlt. Urine steroid metabolomics is a highly sensitive tool for post-operative recurrence detection in adrenocortical carcinoma. *Endocrine Abstracts*, 41:OC1.4, 2016.
60. L. Yeo, N. Adlard, M. Biehl, M. Juarez, T. Smallie, M. Snow, C.D. Buckley, K. Raza, A. Filer, and D. Scheel-Toellner. Expression of chemokines CXCL4 and CXCL7 by synovial macrophages defines an early stage of rheumatoid athritis. *Annals of the Rheumatic Diseases*, 75:763–771, 2015.
61. E.L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *J. Amer. Stat. Assn.*, 53:457–481, 1958.
62. G. Mukherjee, G. Bhanot, K. Raines, S. Sastry, S. Doniach, and M. Biehl. Predicting recurrence in clear cell Renal Cell Carcinoma: Analysis of TCGA data using outlier analysis and generalized matrix LVQ. In *2016 IEEE Congress on Evolutionary Computation (CEC)*, pages 656–661, July 2016.
63. National Cancer Institute and National Human Genome Research Institute. The Cancer Genome Atlas (TCGA) Portal: http://cancergenome.nih.gov (last visited: 16 March 2017),.
64. C.C. Aggarwal. *Outlier Analysis*. Springer, New York, 2013.
65. M. Biehl, P. Sadowski, G. Bhanot, E. Bilal, A. Dayarian, P. Meyer, R. Norel, K. Rhrissorrakrai, M.D. Zeller, and S. Hormoz. Inter-species prediction of protein phosphorylation in the sbv IMPROVER species translation challenge. *Bioinformatics*, 31(4):453–461, 2015.
66. E. Mwebaze and M. Biehl. Prototype-Based Classification for Image Analysis and Its Application to Crop Disease Diagnosis. In E. Merényi, M.J. Mendenhall, and P. O'Driscoll, editors, *Advances in Self-Organizing Maps and Learning Vector Quantization: Proceedings of the 11th International Workshop WSOM 2016*.
67. M. Biehl, K. Bunte, and P. Schneider. Analysis of Flow Cytometry Data by Matrix Relevance Learning Vector Quantization. *PLoS ONE*, 8(3):e59401, 2013.
68. N. Aghaeepour, G. Finak, The FlowCAP Consortium, The DREAM Consortium*, H. Hoos, T.R. Mosmann, R. Brinkman, R. Gottardo, and R.H. Scheuermann. Critical assessment of automated flow cytometry data analysis techniques. *Nature Methods*, 10(3):228–238, 2013.

69. D. Mudali, M. Biehl, K.L. Leenders, and J.B.T.M. Roerdink. LVQ and SVM classification of FDG-PET brain data. In Erzsébet Merényi, J. Michael Mendenhall, and Patrick O'Driscoll, editors, *Advances in Self-Organizing Maps and Learning Vector Quantization: Proc. of the 11th Intl. Workshop WSOM 2016, Houston, Texas, USA, January 6-8, 2016*, pages 205–215, Cham, 2016. Springer.
70. D. Mudali, M. Biehl, S.K. Meles, R.J. Renken, D. Garcia-Garcia, P. Clavero, J. Arbizu, J.A. Obeso, M.C. Rodriguez-Oroz, K. Leenders, and J.B.T.M. Roerdink. Differentiating Early and Late Stage Parkinson's Disease Patients from Healthy Controls. *Journal of Biomedical Engineering and Medical Imaging*, 3:33–43, 2016.
71. B. Frenay, D. Hofmann, A. Schulz, M. Biehl, and B. Hammer. Valid interpretation of feature relevance for linear data mappings. In *Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on*, pages 149–156. IEEE, 2014.
72. E. Mwebaze, G. Bearda, M. Biehl, and D. Zühlke. Combining dissimilarity measures for prototype-based classification. In M. Verleysen, editor, *23rd European Symposium on Artificial Neural Networks (ESANN 2015)*, pages 31–36. d-side publishing, 2015.