ESANN'2005 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 27-29 April 2005, d-side publi., ISBN 2-930307-05-6.

Generalized Relevance LVQ with Correlation Measures for Biological Data

Marc Strickert[1], Nese Sreenivasulu[2], Winfriede Weschke[2], Udo Seiffert[1],
Thomas Villmann[3]

1 - Pattern Recognition Group, 2 - Gene Expression Group
Institute of Plant Genetics and Crop Plant Research Gatersleben, Germany
{stricker,srinivas,weschke,seiffert}@ipk-gatersleben.de

3 - University Leipzig, Clinic for Psychotherapy, Germany
villmann@informatik.uni-leipzig.de

Abstract. Generalized Relevance Learning Vector Quantization (GRLVQ) is combined with correlation-based similarity measures. These are derived from the Pearson correlation coefficient in order to replace the adaptive squared Euclidean distance which is typically used for GRLVQ. Patterns can thus be used without further preprocessing and compared in a manner invariant to data shifting and scaling transforms. High accuracies are demonstrated for a reference experiment of handwritten character recognition and good discrimination ability is shown for the detection of systematic differences between gene expression experiments.

Keywords. Prototype-based learning, adaptive metrics, correlation measure, Learning Vector Quantization, GRLVQ.

1 Introduction

Pattern classification is the key technology for solving tasks in diagnostics, automation, information fusion, and forecasting. Backbones of pattern classification are the underlying similarity measures: they define how data items are compared, and they control the grouping of data. Thus, depending on the notion of similarity, a data set can be viewed and processed from different perspectives. In learning vector quantization (LVQ) a data vector can be compared with a prototype vector for example according to the Euclidean distance or the Manhattan block distance, the former measuring diagonally across the vector space, the latter summing up distances along each dimension axis. Thereby, the block distance better maintains the independence of the considered attributes' physical meanings, while the Euclidean metric allows shortcuts the attribute space. In any case, the specific structure of the data space can and should be accounted for by selecting the appropriate metric. Alternatively, metrics evolve their specificity automatically during training within a certain range, as proposed by Kaski [5] for extensions of the self-organizing map (SOM) or by Hammer and Villmann [4] for LVQ-based learning. In biological sciences also the functional aspect of collected data plays an important role: the organization of spatio-temporal patterns for gene expression levels might be more revealing by comparing shapes of the expression profiles rather than finding spatially close expression vectors. A commonly used measure to meet this purpose is given by the Pearson correlation which describes the degree of linear dependence between two data sets. However attractive for pattern processing, attention must be paid to combining this measure with prototype-based learning methods, such as the unsupervised clustering with SOM [6] or neural gas (NG) [7] or the supervised classification with LVQ [6]. Ad-hoc solutions just replace the Euclidean distance, stated in the original formulations of the algorithms, by a correlation measure without paying attention to the prototype update. Thus, winner selection is changed, but update is still realized for minimizing Euclidean distances by $\boldsymbol{w}_{new} \propto \mathbf{x} - \boldsymbol{w}_{old}$ [2], not for maximizing correlations. This realization is also found in commercial bioinformatics tools, such as ArrayMiner, GeneSpring, or J-Express Pro for a SOM-based gene profile clustering and visualization. The common goal of these programs is gene expression analysis, i.e. the identification of key regulators and coexpressed genes that determine metabolic functions in developing organisms. Since expression profiles are usually assigned to underlying biological objects, auxiliary information for supervised classification is available, such as the developmental stage of the probed tissues, or the stress factors applied to the growing organisms. Here, the supervised generalized relevance learning vector quantization (GRLVQ, [4]) is taken as basis for extensions, because its large-margin generalization properties and its metric adaptivity are founded on strict mathematical derivations of the parametrized squared Euclidean metric [3]. The key issue of GRLVQ is the minimization of a classification cost function; this central idea is transferred to correlation-based similarity. Then we present an application of the new GRLVQ-variant to detect bias in gene expression studies.

2 Generalized Relevance LVQ (GRLVQ) and extensions

Given a set of training data $\mathbf{X} = \{(\mathbf{x}^i, y^i) \in \mathbb{R}^d \times \{1, \ldots, c\} \mid i = 1, \ldots, n\}$ to be classified with d-dimensional elements $\mathbf{x}^k = (x_1^k, \ldots, x_d^k)$ and c classes. A set $\mathbf{W} = \{\boldsymbol{w}^1, \ldots, \boldsymbol{w}^K\}$ of prototypes is used for the data representation, $\boldsymbol{w}^i = (w_1^i, \ldots, w_n^i, y^i) \in \mathbb{R}^d \times \{1, \ldots, c\}$, with class labels y^i attached to locations in the data space.

The classification cost function to be minimized is given in the generic form [4]:

$$\mathrm{E}_{\mathsf{GRLVQ}} := \sum_{i=1}^{n} \mathrm{g}\left(\mathrm{q}_{\boldsymbol{\lambda}}(\mathbf{x}^i)\right) \qquad \text{where} \qquad \mathrm{q}_{\boldsymbol{\lambda}}(\mathbf{x}^i) = \frac{\mathrm{d}^+_{\boldsymbol{\lambda}}(\mathbf{x}^i) - \mathrm{d}^-_{\boldsymbol{\lambda}}(\mathbf{x}^i)}{\mathrm{d}^+_{\boldsymbol{\lambda}}(\mathbf{x}^i) + \mathrm{d}^-_{\boldsymbol{\lambda}}(\mathbf{x}^i)} \,.$$

By summing up the classification costs of all patterns, $\mathrm{E}_{\mathsf{GRLVQ}}$ serves as a quality measure of the classification depending on the similarity, or likewise dissimilarity, of the presented pattern $\mathbf{x}^i$ and the two best-matching prototypes, $\boldsymbol{w}^{i+}$ representing the same label as $\mathbf{x}^i$ and $\boldsymbol{w}^{i-}$ a different label. Usually a sigmoid transfer function $\mathrm{g}(x) = \mathrm{sgd}(x) = 1/(1 + \exp(-x)) \in (0; 1)$ is applied [9]. The implicit degrees of freedom for the cost minimization are the locations of the prototypes in the weight space and, additionally, a set of free parameters $\boldsymbol{\lambda}$ connected to the function $\mathrm{d}_{\boldsymbol{\lambda}}(\mathbf{x}) = \mathrm{d}_{\boldsymbol{\lambda}}(\mathbf{x}, \boldsymbol{w})$ comparing pattern and prototype. In prior work, $\mathrm{d}_{\boldsymbol{\lambda}}(\mathbf{x})$ was supposed to be a metric in mathematical sense, i.e. taking only non-negative values, conforming to the triangle inequality, and giving a distance of $\mathrm{d} = 0$ only for $\boldsymbol{w} = \mathbf{x}$. These conditions make an intuitive interpretation of prototypes possible. However, if just a well-performing classifier invariant to certain features is wanted, distance conditions might be relaxed and instead a similarity measure be plugged into the algorithm. Overall similarity maximization can be expressed in the GRLVQ framework by flipping the sign of the measure and sticking to the minimization of $\mathrm{E}_{\mathsf{GRLVQ}}$. Since the iterative GRLVQ update implements a gradient descent on E, d must be differentiable almost everywhere, no matter if as distance or as similarity measure.

Partial derivatives of $\mathrm{E}_{\mathsf{GRLVQ}}$ yield the generic update formulas for the closest correct and the closest wrong prototype and the metric weights:

$$\begin{aligned}
\triangle \boldsymbol{w}^{i+} &= -\gamma^{+} \cdot \frac{\partial \mathrm{E}_{\mathsf{GRLVQ}}}{\partial \boldsymbol{w}^{i+}} = -\gamma^{+} \cdot \mathrm{g}'\left(\mathrm{q}_{\boldsymbol{\lambda}}(\mathbf{x}^i)\right) \cdot \frac{2 \cdot \mathrm{d}_{\boldsymbol{\lambda}}^{-}(\mathbf{x}^i)}{\left(\mathrm{d}_{\boldsymbol{\lambda}}^{+}(\mathbf{x}^i) + \mathrm{d}_{\boldsymbol{\lambda}}^{-}(\mathbf{x}^i)\right)^2} \cdot \frac{\partial \mathrm{d}_{\boldsymbol{\lambda}}^{+}(\mathbf{x}^i)}{\partial \boldsymbol{w}^{i+}} \\
\triangle \boldsymbol{w}^{i-} &= \gamma^{-} \cdot \frac{\partial \mathrm{E}_{\mathsf{GRLVQ}}}{\partial \boldsymbol{w}^{i-}} = \gamma^{-} \cdot \mathrm{g}'\left(\mathrm{q}_{\boldsymbol{\lambda}}(\mathbf{x}^i)\right) \cdot \frac{2 \cdot \mathrm{d}_{\boldsymbol{\lambda}}^{+}(\mathbf{x}^i)}{\left(\mathrm{d}_{\boldsymbol{\lambda}}^{+}(\mathbf{x}^i) + \mathrm{d}_{\boldsymbol{\lambda}}^{-}(\mathbf{x}^i)\right)^2} \cdot \frac{\partial \mathrm{d}_{\boldsymbol{\lambda}}^{-}(\mathbf{x}^i)}{\partial \boldsymbol{w}^{i-}} \\
\triangle \boldsymbol{\lambda} &= -\gamma^{\boldsymbol{\lambda}} \cdot \frac{\partial \mathrm{E}_{\mathsf{GRLVQ}}}{\partial \boldsymbol{\lambda}} = -\gamma^{\boldsymbol{\lambda}} \cdot \mathrm{g}'\left(\mathrm{q}_{\boldsymbol{\lambda}}(\mathbf{x}^i)\right) \cdot \frac{2 \cdot \partial \mathrm{d}_{\boldsymbol{\lambda}}^{+}(\mathbf{x}^i)/\partial \boldsymbol{\lambda} \cdot \mathrm{d}_{\boldsymbol{\lambda}}^{-}(\mathbf{x}^i) - 2 \cdot \mathrm{d}_{\boldsymbol{\lambda}}^{+}(\mathbf{x}^i) \cdot \partial \mathrm{d}_{\boldsymbol{\lambda}}^{-}(\mathbf{x}^i)/\partial \boldsymbol{\lambda}}{\left(\mathrm{d}_{\boldsymbol{\lambda}}^{+}(\mathbf{x}^i) + \mathrm{d}_{\boldsymbol{\lambda}}^{-}(\mathbf{x}^i)\right)^2}
\end{aligned}$$

Learning rates are $\gamma^{\boldsymbol{\lambda}}$ for the metric parameters λ_j, all initialized equally by $\lambda_j = 1/d, j = 1 \ldots d$; γ^{+} and γ^{-} describe the update amount. Their choice depends on the used measure – generally, they should be chosen according to the relation $0 \leq \gamma^{\boldsymbol{\lambda}} \ll \gamma^{-} \leq \gamma^{+} \leq 1$ and decreased within these constrains during training. Metric adaptation should be realized slowly, as a reaction to the quasi-stationary solutions for the prototype positions. Moreover, the normalization of $\sum_{i=1}^{d} \lambda_i = 1$ is necessary in order to prevent divergence of the parameters $\boldsymbol{\lambda}$. The above set of equations is a convenient starting point to test different concepts of similarity by just inserting the denoted partial derivatives of $\mathrm{d}_{\boldsymbol{\lambda}}(\mathbf{x})$.

3 Metrics and similarity measures

The missing ingredient for carrying out comparisons is either a distance metric or a more general (dis-)similarity measure $\mathrm{d}_{\boldsymbol{\lambda}}(\mathbf{x}, \boldsymbol{w})$. For reference, formulas for the weighted Euclidean distance will be given. Then, by relaxing the conditions of metrics, two measures are derived from the Pearson correlation, which inherit the invariance to shifting and amplitude scaling. The feature of prototype invariance implemented by the presented update dynamic is desirable in situations when mainly frequency information and simple graph-matching is accounted for. More details on graph-matching properties or general functional data processing with the prototype-based unsupervised SOM algorithm are given by Rossi et al. [8].

3.1 Weighted Euclidean metric

The weighted Euclidean metric yields the following set of equations [10]:

$$\begin{aligned}
\mathrm{d}^{\mathsf{EUC}}_{\boldsymbol{\lambda}}(\mathbf{x}, \boldsymbol{w}^i) &= \sum_{j=1}^{d} \lambda_j^{b_{\boldsymbol{\lambda}}} \cdot (x_j - w^i_j)^{b_{\boldsymbol{w}}}\ , \ \text{integers}\, b_{\boldsymbol{\lambda}}, b_{\boldsymbol{w}} \geq 0\ ,\ b_{\boldsymbol{w}}\ \text{even} \\
\Rightarrow \quad \frac{\partial \mathrm{d}^{\mathsf{EUC}}_{\boldsymbol{\lambda}}(\mathbf{x}, \boldsymbol{w}^i)}{\partial w^i_j} &= -b_{\boldsymbol{w}} \cdot \lambda_j^{b_{\boldsymbol{\lambda}}} \cdot (x_j - w^i_j)^{b_{\boldsymbol{w}}-1}, \\
\frac{\partial \mathrm{d}^{\mathsf{EUC}}_{\boldsymbol{\lambda}}(\mathbf{x}, \boldsymbol{w}^i)}{\partial \lambda_j} &= b_{\boldsymbol{\lambda}} \cdot \lambda_j^{b_{\boldsymbol{\lambda}}-1} \cdot (x_j - w^i_j)^{b_{\boldsymbol{w}}}.
\end{aligned}$$

For simplicity, roots have been omitted. In the squared case with $b_{\boldsymbol{w}} = 2$, the derivative for the prototype update $2 \cdot (x_j - w^i_j)$ is recognizable as Hebbian learning term. In other cases, large $b_{\boldsymbol{w}}$ tend to focus on dimensions with large differences, and small $b_{\boldsymbol{w}}$ focus on dimensions with small differences. Approved values for the exponents of the relevance factors are $b_{\boldsymbol{\lambda}} \in \{1, 2\}$.

3.2 Correlation measures

In the following, the common definition of the Pearson correlation

$$\mathsf{r} = \mathrm{d}^{\mathsf{r}}(\mathbf{x}, \boldsymbol{w}^i) = \frac{\sum_{j=1}^{d} (w^i_j - \mu_{\boldsymbol{w}^i}) \cdot (x_j - \mu_{\mathbf{x}})}{\sqrt{\sum_{j=1}^{d} (w^i_j - \mu_{\boldsymbol{w}^i})^2} \cdot \sqrt{\sum_{j=1}^{d} (x_j - \mu_{\mathbf{x}})^2}} \quad \in [-1; 1] \tag{1}$$

is not suitable in practice, because only a small range of values is taken and, furthermore, for well-matching vectors the calculated values are maximum instead of minimum. Therefore, inverse fractions of appropriately reshaped functions will be taken in the following. Since metric adaptivity has turned out to be beneficial, free parameters are added here to the covariance expression for weighting individual data dimensions. Then, the numerator of Eqn. 1 becomes

$$\mathscr{H} := \sum_{j=1}^{d} \lambda_j \cdot (w^i_j - \mu_{\boldsymbol{w}^i}) \cdot (x_j - \mu_{\mathbf{x}})\ .$$

The variable relevance factors are formally assigned to the adaptive *weight* derivations from the mean in order to scale the prototypes' influence, but not the static data. This deliberate asymmetry yields the following two separate variance terms for the denominator:

$$\mathscr{W} := \sum_{j=1}^{d} \lambda_j^2 \cdot (w^i_j - \mu_{\boldsymbol{w}^i})^2 \qquad \text{and} \qquad \mathscr{X} := \sum_{j=1}^{d} (x_j - \mu_{\mathbf{x}})^2\ .$$

Subsequently, these shortcuts for $\mathsf{r}_{\boldsymbol{\lambda}} = \mathscr{H}/\sqrt{\mathscr{W} \cdot \mathscr{X}}$ will become very handy in order to derive two application-specific heuristic correlation measures, the squared inverse correlation $\mathsf{r} \rightarrow \mathsf{r}^{-2}$ and the shifted inverse correlation $\mathsf{r} \rightarrow (1 + \mathsf{r})^{-\mathsf{k}}$. While the former r^{-2} treats the cases of correlation and anti-correlation as similar, the latter $(1 + \mathsf{r})^{-\mathsf{k}}$ distinguishes these cases. Both measures must be derived independently, because the domain $[-1; 1]$ of r must be transformed into appropriate new ones that are suitable for fast GRLVQ cost function minimization.

Squared inverse correlation

Since square roots in Eqn. 1 complicate calculations, the expression is taken to the inverse power of two. This negative power transforms the output from $[-1;1]$ to $[1;\infty)$ - a simpler formulation such as $1-\mathsf{r}^2$ did not exhibit satisfactory convergence in practice; maybe there exist many solutions close to zero that induce a plateau in the cost function. In contrast to that, inverse power cost functions yield large correction terms for badly correlated prototypes. The inverse correlation measure and its derivatives are expressed by:

$$\begin{aligned}
\mathrm{d}_{\boldsymbol{\lambda}}^{\mathsf{r}^{-2}}(\mathsf{x},\boldsymbol{w}^i) &= \frac{\left(\sum_{j=1}^{d}\lambda_j^2\cdot(w_j^i-\mu_{\boldsymbol{w}^i})^2\right)\cdot\left(\sum_{j=1}^{d}(x_j-\mu_{\mathsf{x}})^2\right)}{\left(\sum_{j=1}^{d}\lambda_j\cdot(w_j^i-\mu_{\boldsymbol{w}^i})\cdot(x_j-\mu_{\mathsf{x}})\right)^2}=\frac{\mathscr{W}\cdot\mathscr{X}}{\mathscr{H}^2}\\
\Rightarrow\ \frac{\partial\mathrm{d}_{\boldsymbol{\lambda}}^{\mathsf{r}^{-2}}(\mathsf{x},\boldsymbol{w}^i)}{\partial w_j^i} &= 2\cdot\mathscr{X}\cdot\frac{\lambda_j\cdot(w_j^i-\mu_{\boldsymbol{w}^i})\cdot\mathscr{H}\ -\ (x_j-\mu_{\mathsf{x}})\cdot\mathscr{W}}{\mathscr{H}^3}\cdot\lambda_j\\
\frac{\partial\mathrm{d}_{\boldsymbol{\lambda}}^{\mathsf{r}^{-2}}(\mathsf{x},\boldsymbol{w}^i)}{\partial\lambda_j} &= 2\cdot\mathscr{X}\cdot\frac{\lambda_j\cdot(w_j^i-\mu_{\boldsymbol{w}^i})\cdot\mathscr{H}\ -\ (x_j-\mu_{\mathsf{x}})\cdot\mathscr{W}}{\mathscr{H}^3}\cdot(w_j^i-\mu_{\boldsymbol{w}^i})\,.
\end{aligned}$$

But attention must be paid: minimum values are returned from $\mathrm{d}_{\boldsymbol{\lambda}}^{\mathsf{r}^{-2}}$ for both maximum correlation and maximum anti-correlation; hence, both data characteristics will become represented by the same prototype. It depends on the specific application if this is property is desirable or not: while gene coexpression analysis requires a clear distinction of the correlated and the anti-correlated profiles, multi-class problems with highly asymmetric feature vector profiles are likely to profit from the squared measure. A reformulation which maintains and emphasizes positive correlation is the shifted inverse measure discussed in the next section.

Shifted inverse correlation

In order to direct the learning process towards only positive correlations, a unit shift of r from its minimum of -1 to 0 is taken as denominator argument of a power fraction. This yields ∞ in the rare case of perfect anti-correlation and values close to zero for perfect correlation. The according expressions for the measure $(1+\mathsf{r})^{-\mathsf{k}}$ and its derivatives are:

$$\begin{aligned}
\mathrm{d}_{\boldsymbol{\lambda}}^{(1+\mathsf{r})^{-\mathsf{k}}}(\mathsf{x},\boldsymbol{w}^i) &= \frac{1}{(1+\mathrm{d}^{\mathsf{r}}(\mathsf{x},\boldsymbol{w}^i))^k}=\Big(1+\frac{\mathscr{H}}{\sqrt{\mathscr{W}}\cdot\sqrt{\mathscr{X}}}\Big)^{-k}=:\mathscr{R}^{-k}\\
\Rightarrow\ \frac{\partial\mathrm{d}_{\boldsymbol{\lambda}}^{(1+\mathsf{r})^{-\mathsf{k}}}(\mathsf{x},\boldsymbol{w}^i)}{\partial w_j^i} &= k\cdot\mathscr{R}^{-k-1}\cdot\frac{\lambda_j\cdot(w_j^i-\mu_{\boldsymbol{w}^i})\cdot\mathscr{H}-(x_j-\mu_{\mathsf{x}})\cdot\mathscr{W}}{\sqrt{\mathscr{W}^3}\cdot\sqrt{\mathscr{X}}}\cdot\lambda_j\\
\frac{\partial\mathrm{d}_{\boldsymbol{\lambda}}^{(1+\mathsf{r})^{-\mathsf{k}}}(\mathsf{x},\boldsymbol{w}^i)}{\partial\lambda_j} &= k\cdot\mathscr{R}^{-k-1}\cdot\frac{\lambda_j\cdot(w_j^i-\mu_{\boldsymbol{w}^i})\cdot\mathscr{H}-(x_j-\mu_{\mathsf{x}})\cdot\mathscr{W}}{\sqrt{\mathscr{W}^3}\cdot\sqrt{\mathscr{X}}}\cdot(w_j^i-\mu_{\boldsymbol{w}^i})\,.
\end{aligned}$$

The integer parameter $k>0$ takes influence on the convergence: too low values require large learning rates and induce many training cycles, whereas too large values inhibit the generalization capabilities and lead to numeric instabilities; in experiments, values in the range of $8\le k\le 20$ have turned out to be suitable – the experiments given below use a value of $k=16$.

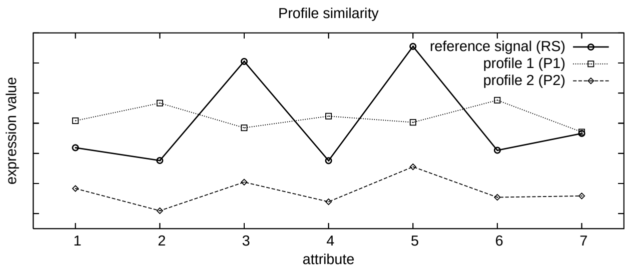

Fig. 1: Data profiles compared with different similarity functions. Relation signs for the squared Euclidean metric and the squared inverse correlation differ: $\mathrm{d}^{\mathsf{EUC}}(\mathsf{RS},\mathsf{P1})=0.82<\mathrm{d}^{\mathsf{EUC}}(\mathsf{RS},\mathsf{P2})=1.81$ but $\mathrm{d}^{\mathsf{r}^{-2}}(\mathsf{RS},\mathsf{P1})=3.55>\mathrm{d}^{\mathsf{r}^{-2}}(\mathsf{RS},\mathsf{P2})=1.25$.

As illustrated in Fig. 1, correlation measures can have fundamentally different properties than the Euclidean distance: the two profiles compared with a reference profile yield opposite relations, depending on the applied similarity function. Although the z-score transform —mean subtracted data scaled by standard deviation— can be roughly found in the Pearson term of Eqn. 1, data preprocessing cannot transform the classification problem into an equivalent one solvable with the Euclidean metric; the update formulas exhibit structural differences. As a rule of thumb, if a prototype is similar to input points in Euclidean sense, then it is very likely that it is also highly correlated to them. The other direction is untrue: if high correlation exists, there might a large Euclidean distance. Thus, potentially fewer prototypes are necessary for representations based on correlation similarities. This way sparser data models can be realized.

4 Experiments

4.1 Handwritten digit recognition

The first experiment is a multi-class problem of recognizing handwritten characters available from the FL3 handwritten symbol database [1]. 3471 binary coded images (32×32) of the digits $0\ldots9$ are given in the form of extracted 218-dimensional feature vectors as described in Villmann et al. [10]. For reference, the same data sets as in [10] are used, i.e. feature vectors of the original 32×32-images and for those from affine transforms to 64×64-images. Thus, two training sets containing 2280 patterns and two test sets with 1191 patterns are available. For comparison with recent extensions of LVQ given in [10], training utilizes 10 prototypes per class and applies 500 epochs. Results are summed up in table 1. For the original images, the shifted inverse correlation measure yields the best results for training and testing. This high accuracy is particularly remarkable in comparison to the supervised relevance neural gas algorithm (SRNG) which adapts not only the closest matching and the closest mismatching prototype but which additionally accounts for the neighborhood. Also squared inverse correlation measure performs well with GRLVQ. For reference, generalized LVQ (GLVQ) [9] with cost function and Kohonen's original LVQ3 [6], both Euclidean classifiers like SRNG, yield clearly lower accuracies. The transformed 64×64-images are still more difficult to learn: decreased generalization performance

Set/Method	$\mathsf{GRLVQ}_{(1+r)^{-16}}$	$\mathsf{GRLVQ}_{r^{-2}}$	SRNG	GLVQ	LVQ3
32 × 32-train	**98.3%**	97.7%	92.4%	91.5%	83.2%
32 × 32-test	**93.6%**	93.0%	89.4%	88.0%	80.1%
64 × 64-train	95.4%	**96.1%**	86.4%	86.1%	71.5%
64 × 64-test	78.6%	82.7%	**84.3%**	75.4%	68.7%

Table 1: Digit classification. Results for SRNG, GLVQ, LVQ3 are taken from [10].

of all models indicates a multi-modal data distribution or a non-representative training set. Although $\mathsf{GRLVQ}_{r^{-2}}$ exhibits the best accuracy on the training set, the generalization capability of SRNG is better. However, if the reference training conditions for $\mathsf{GRLVQ}_{r^{-2}}$ are relaxed to 2500 epochs (it is assumed that the reference results have optimally converged), accuracies of 96.4% and 85.3% are obtained for training and testing, respectively, with only 5 prototypes per class. These good results for difficult data indicate a general suitability of the correlation measures for other classification tasks.

4.2 Bias detection in gene expression experiments

The second study is connected to macroarray data. Expression profiles of 1421 genes were collected from filial tissues of barley seed during 7 developmental stages. For control purposes, each experiment has been repeated from 2 sets of independently grown plant material. The question of interest is, if a systematic difference can be found in the gene expression profiles resulting from the two experimental series. Thus, 1421 data vectors in 7 dimensions, are considered for each of the two classes. Since only rough tendencies are of interest, a single prototype is used for each class. 7500 epochs of 25 separate runs on random half splits of the available data have been run for the weighted squared Euclidean and both correlation measures with the best manually found parameters. The training accuracy of the Euclidean-based classifier is $51.30 \pm 1.34\%$ and the testing accuracy is $50.02 \pm 1.23\%$, i.e. this model does not perform better than guessing, which is expected for two identically conducted experiments. Anyway, the shifted inverse correlation yields a generalization accuracy of approximately 54%. Even better, the squared inverse correlation increases the test set accuracy to $64.57 \pm 1.60\%$ at a training accuracy of $68.34 \pm 1.88\%$. These results point out the a significant difference between the expression profiles from either experiments. A look at the average metric parameters $\mu(\lambda_j)$ with $\sigma(\lambda_j) < 0.0065 \, \forall \, j$

$$\mu(\boldsymbol{\lambda}) = (0.137_{(1)}, 0.139_{(2)}, 0.150_{(3)}, 0.149_{(4)}, 0.145_{(5)}, 0.140_{(6)}, 0.139_{(7)})$$

reveals emphasis on components $j = \{3, 4, 5\}$, $\lambda_j > 0.143$ which are greater than the average of $0.143 \approx 1/7$. Further biological investigations indicated a very slight shift in assigning developmental stages between the two sets of independent experiments. In the conducted gene expression experiments a robust transcriptional reprogramming occurred during intermediate stage related to components 4 and 5 of filial tissue development. Although overall expression data between the two sets of experiments are hardly distinguishable in practice, the slight systematic influence depending on a precise assigning of the developmental stages affects gene expression during the intermediate phase. These slight differences in the mutual correlations were detected and could be exploited by the $\mathsf{GRLVQ}_{r^{-2}}$-classifier, a useful property for processing biological observations.

5 Conclusions and future work

Adaptive correlation-based similarity measures have been successfully integrated into the existing mathematical framework of GRLVQ learning. The experiments show that there is much potential in using non-Euclidean similarity measures. High sensitivity to specific differences in the data is realized, and very good classification results can be obtained with a small number of prototypes. A potential drawback of the obtained prototypes is the difficulty to interpret them, especially in case of the squared inverse correlation, for which both correlated and anti-correlated data are matched by the same prototype. Further studies must reveal in which way the adapted metric $\boldsymbol{\lambda}$-parameters emphasize certain data dimensions; preliminary results show differences and similarities to the results obtained for the adaptive Euclidean measure, but specific characterization will be necessary. The next step will be the integration of the correlation measures into the supervised relevance neural gas (SRNG) method in order to further improve convergence and accuracy. Future applications of the proposed correlation-based classifiers will be implemented for the analysis of high-throughput gene expression data in order to identify key regulators in clusters of coexpressed genes.

Acknowledgments

Thanks to Dr. Volodymyr Radchuk for helping probe preparation for macroarray experiments and to Frank Schleif for the preprocessed FL3 symbol data sets. The present work is supported by BMBF grant FKZ 0313115, GABI-SEED-II.

References

[1] FL3 Handwritten Symbol Database - Subset of the NIST Special Database 1, NIST. ftp://sequoyah.ncsl.nist.gov/pub/databases/data.

[2] I. Fischer. Similarity-based neural networks for applications in computational molecular biology. *Advances in Intelligent Data Analysis V*, 5(3):208–218, 2003.

[3] 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 Softcomputing*, volume 3070 of *Springer Lecture Notes in Artificial Intelligence*, pages 592–597. Springer, 2004.

[4] B. Hammer and T. Villmann. Generalized Relevance Learning Vector Quantization. *Neural Networks*, 15:1059–1068, 2002.

[5] S. Kaski. Bankruptcy analysis with self-organizing maps in learning metrics. *IEEE Transactions on Neural Networks*, 12:936–947, 2001.

[6] T. Kohonen. *Self-Organizing Maps.* Springer-Verlag, Berlin, 3rd edition, 2001.

[7] T. Martinetz, S. Berkovich, and K. Schulten. "Neural-gas" network for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks*, 4(4):558–569, 1993.

[8] F. Rossi, B. Conan-Guez, and A. E. Golli. Clustering functional data with the SOM algorithm. In *Proceedings of ESANN 2004*, pages 305–312, Bruges, Belgium, April 2004.

[9] A. Sato and K. Yamada. Generalized Learning Vector Quantization. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems 7 (NIPS)*, volume 7, pages 423–429. MIT Press, 1995.

[10] T. Villmann, F. Schleif, and B. Hammer. Supervised Neural Gas and Relevance Learning in Learning Vector Quantization. In T. Yamakawa, editor, *Proc. of the Workshop on Self-Organizing Networks (WSOM)*, pages 47–52, Kyushu Institute of Technology, 2003.