ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

# Metric adaptation for supervised attribute rating

M. Strickert[a*], F.-M. Schleif[b], and T. Villmann[b]

[a] Leibniz Institute of Plant Genetics and Crop Plant Research Gatersleben,
[b] Research group Computational Intelligence, University of Leipzig
[*] Corresponding author: stricker@ipk-gatersleben.de

**Abstract**. A new approach for faithful relevance rating of attributes is proposed, enabling class-specific discriminatory data space transformations. The method is based on the adaptation of the underlying data similarity measure by using class information linked to the data vectors. For adaptive Minkowski metrics and parametric Pearson similarity, the obtained attribute weights can be used for back-transforming data for further analysis with methods utilizing non-adapted measures as demonstrated for benchmark and mass spectrum data.

**Keywords**. Supervised feature characterization, adaptive measures.

## 1 Introduction

In biomedical labs high-throughput facilities allow quantification of thousands of attributes per single probe. For specific diagnostic tasks, such as the identification of characteristically expressed cDNA or metabolite masses, only a minor fraction of the data variables contains informative measurements. Task-specific attribute rating is useful for noise cancellation and for focussing on interesting parts of the data, thereby, helping to lower the curse of dimensionality.

Generally, unsupervised methods for attribute characterization are available, such as factor loading analysis in principal component analysis (PCA). For task-related attribute rating, supervised strategies like iterative Relief [6] or linear discriminant analysis (LDA) [3] make use of additional class label information. The detection of interesting attributes is known as the feature subset selection (FSS) problem related to detecting specific data attributes for class-specific assignments [2]. So-called attribute filter methods are usually fast, because they do not require the construction of attribute probing classifiers. Since their truth is strictly dependent on the chosen filter criterion, such as information gain, their outcome, though, may be incompatible with subsequent methods using other data measures. In contrast to that, exhaustive probing by wrapper programs that pass on selected attributes to subsequent classifiers, is known to be time consuming and is mostly not applicable in large scenarios. Many programs for FSS make use of heavy statistical density estimation or classification indices.

Here, an appealingly simple approach is taken that optimizes attribute relevances during training as a function of within-class and between-class variance optimization. Thus, no classifier is built; instead, all pairs of data vectors contribute to change the similarity measure according to maximum class separation.

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

## 2 Method

Simultaneous attribute characterization of real-valued vector data is realized by supervised adaptation of a chosen similarity measure: the parameters are optimized for minimizing dissimilarities within classes while maximizing distances between classes. The underlying idea is related to LDA, but the iterative optimization and practical implementation are much different from that.

**Notation.** Using $\mathrm{d}_{\boldsymbol{\lambda}}^{ij} = \mathrm{d}_{\boldsymbol{\lambda}}(\mathbf{x}^i, \mathbf{x}^j)$ for $\boldsymbol{\lambda}$-parametrized dissimilarities between $n$ data vectors $\mathbf{x}^u \in \mathbb{R}^q$ $(u = 1 \ldots n)$, two pairwise sums of distances – $\sigma_1^2$ for matching class memberships and $\sigma_0^2$ for mismatching classes – are defined by

$$\sigma_l^2(\boldsymbol{\lambda}) = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{d}_{\boldsymbol{\lambda}}^{ij} \cdot \left(1 - l - (-1)^l \cdot \delta^{ij}\right) \;, \quad l = 0, 1 \,. \tag{1}$$

The Kronecker symbol $\delta^{ij}$ indicates identity of class memberships, i.e. $\delta^{ij} = 1$ if the class of data vector $\mathbf{x}^i$ equals the class of vector $\mathbf{x}^j$, $\delta^{ij} = 0$ in the case of disagreement. The determination of the parameter vector $\boldsymbol{\lambda}$, connected to adaptive similarity measures, is of interest. Here, adaptive Minkowski metrics ${}_p\mathrm{d}_{\boldsymbol{\lambda}}^{ij}$ and Pearson correlation ${}_r\mathrm{d}_{\boldsymbol{\lambda}}^{ij}$ are considered:

$${}_p\mathrm{d}_{\boldsymbol{\lambda}}^{ij} = \left(\sum_{k=1}^{q} \lambda_k \cdot |x_k^i - x_k^j|^p\right)^{1/p} \;, \tag{2}$$

$${}_r\mathrm{d}_{\boldsymbol{\lambda}}^{ij} = \frac{\sum_{k=1}^{q} \lambda_k^2 \cdot (x_k^i - \mu_{\mathbf{x}^i}) \cdot (x_k^j - \mu_{\mathbf{x}^j})}{\sqrt{(\sum_{k=1}^{q} \lambda_k^2 \cdot (x_k^i - \mu_{\mathbf{x}^i})^2) \cdot (\sum_{k=1}^{q} \lambda_k^2 \cdot (x_k^j - \mu_{\mathbf{x}^j})^2)}} = \frac{\mathscr{H}}{\sqrt{\mathscr{W} \cdot \mathscr{U}}} \,. \tag{3}$$

For values of $\lambda_k = 1, k = 1 \ldots q$, i.e. $\boldsymbol{\lambda} = \mathbf{1}$, standard versions are obtained. For adapted values, columns of the data matrix $\mathbf{X}$ with row vectors $\mathbf{x}^i$ can be rescaled by relevance factors $\lambda_k^{1/p}$ for Minkowski metrics in Eqn. 2 and by plain factors $\lambda_k$ for Pearson correlation in Eqn. 3 where $\mu_{\mathbf{x}^u}$ is average of $\mathbf{x}^u$. Standard methods, for example, PCA projection of rescaled Euclidean data, can be used after this simple transformation. The right-hand side abbreviation of Eqn. 3, defined by intuitive one-to-one matching, is used for derivatives later.

In Eqn. 1 the symbol $\sigma^2$ has been chosen on purpose: for the nonparametric squared Euclidean distance the variance $\sigma_k^2$ (second central moment) of attribute $k$ can be expressed up to a scaling factor as double sum over all distances

$$\sigma_{0,k}^2 + \sigma_{1,k}^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} (x_k^i - x_k^j)^2 = 2\,n\,(n-1) \cdot \sigma_k^2 \,. \tag{4}$$

**Optimization options.** There are many possibilities to meet the policy: make more similar which – by class membership – belongs together, make more dissim-

ilar which should be separated. Using derivatives of the general stress function:

$$\lambda_k \leftarrow \lambda_k - \gamma \cdot \frac{\partial \mathrm{s}}{\partial \lambda_k} \quad \Rightarrow \quad \frac{\partial \mathrm{s}}{\partial \lambda_k} = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial \mathrm{s}}{\partial \mathrm{d}_{\boldsymbol{\lambda}}^{ij}} \cdot \frac{\partial \mathrm{d}_{\boldsymbol{\lambda}}^{ij}}{\partial \lambda_k} \to 0 \tag{5}$$

parameter $\lambda_k$ of the similarity measure can be found by a stochastic gradient descent with step size $\gamma$. Two specific stress functions for minimization with mixing factors $f_0 \in \mathbb{R}$ for inter- and and $f_1 \in \mathbb{R}$ for intra-class influence are

$$\mathrm{s}_A(\boldsymbol{\lambda}) = \frac{\sigma_1^2(\boldsymbol{\lambda})}{\sigma_0^2(\boldsymbol{\lambda})} \qquad \text{and} \qquad \mathrm{s}_B(\boldsymbol{\lambda}) = f_1 \cdot \sigma_1^2(\boldsymbol{\lambda}) - f_0 \cdot \sigma_0^2(\boldsymbol{\lambda}). \tag{6}$$

**The first approach** $\mathrm{s}_A(\boldsymbol{\lambda})$ has been discussed recently [5]. For Euclidean spaces there is a clear link to LDA where the inverse proportional fraction $\sigma_0^2/\sigma_1^2$ gets maximized along an optimized class-separating direction in data space. In $\mathrm{s}_A(\boldsymbol{\lambda})$, the data measure is rescaled to obtain class separation. One potential disadvantage of the term is the impossibility to control the specific tradeoff of intra- vs. inter-class variability. A remedy is proposed in the following.
**The second approach** $\mathrm{s}_B(\boldsymbol{\lambda})$ with its partial derivative

$$\frac{\partial \mathrm{s}_B}{\partial \mathrm{d}_{\boldsymbol{\lambda}}^{ij}} = f_1 \cdot \delta^{ij} - f_0 \cdot (1 - \delta^{ij}) \tag{7}$$

allows to formulate an astonishingly simple feature weighting procedure:

```
 1: Read input data X and class labels
 2: λ ← 1
 3: repeat
 4:    Δ ← 0
 5:    for all data pairs x^i, x^j do
 6:       if class(i) = class(j) then
 7:          Δ ← Δ + f_1 · ∂d^ij_λ/∂λ
 8:       else
 9:          Δ ← Δ − f_0 · ∂d^ij_λ/∂λ
10:       end if
11:    end for
12:    Δ ← Δ/ (max(Δ) − min(Δ))
13:    λ ← λ − γ · Δ
14:    λ ← max(0, λ)
15:    λ ← λ · q/ Σ_{k=1}^{q} λ_k
16: until convergence.
```

Lines 7 and 9 implement Eqn. 7 and combine the right factor of the chain rule given by Eqn. 5. Using vector component-wise min and max, lines 12-15 normalize the parameter update – a learning rate of $\gamma = 0.1$ being appropriate – creating non-negative values $\lambda_k$ summing up to the number of dimensions $q$. The tradeoff parameters $f_0$ and $f_1$ can be naturally expressed by a single variable $f \in [0; 1]$:

$$f_1 = \frac{f}{g}\,, \quad f_0 = \frac{1-f}{1-g} \quad \text{with} \quad g := \frac{\left|\{\delta^{ij} = 1 | i, j = 1 \ldots q\}\right|}{q^2}\,. \tag{8}$$

For $f = g$ unit influence of intra- and interclass adaptation stress, $f_0 = f_1 = 1$, is obtained, irrespective of the number of data vectors in identical and different

classes. For $f = 0.5$, a natural normalization of intraclass derivatives is realized by the portion $g$ of within-class data pairs, as well as a normalization of interclass derivatives by the number of data pairs belonging to different classes. A choice of $f = 1$ yields extreme bias to only intraclass derivatives, while $f = 0$ regards only interclass relationships. Interesting balanced phase transitions are within these extremes, and currently found empirically by interval search on $f$.

For the proposed similarity measures the partial parameter derivatives are

$$\frac{\partial({}_p\mathrm{d}^{ij}_{\boldsymbol{\lambda}})}{\partial\lambda_k} = \frac{1}{p}\cdot|x^i_k - x^j_k|^p \cdot \left(\sum_{l=1}^{q}\lambda_l\cdot|x^i_l - x^j_l|^p\right)^{-1+1/p}, \quad (9)$$

$$\frac{\partial({}_r\mathrm{d}^{ij}_{\boldsymbol{\lambda}})}{\partial\lambda_k} = -\lambda_k\cdot\frac{({}_\mu x^i_k)^2\cdot\mathscr{H}\cdot\mathscr{U} - 2\cdot{}_\mu x^i_k\cdot{}_\mu x^j_k\cdot\mathscr{U}\cdot\mathscr{W} + ({}_\mu x^j_k)^2\cdot\mathscr{H}\cdot\mathscr{W}}{(\mathscr{U}\cdot\mathscr{W})^{3/2}}. \quad (10)$$

Letters in Eqn. 10 refer to the ones defined in Eqn. 3, the shortcut notation ${}_\mu x^u_k = x^u_k - \mu_{\mathbf{x}^u}$ denotes component $k$ of mean-subtracted vector $\mathbf{x}^u$.

## 3 Applications

**Tecator benchmark data.** The well-known benchmark data set contains 215 samples of 100-dimensional infrared absorbance spectra recorded on a Tecator Infratec Food and Feed Analyzer working in the wavelength range 850–1050$nm$ by the Near Infrared Transmission (NIT) principle [1]. The original regression data is changed into crisp labels of low and high fat content in meat probes according to [4] for feature identification aiming at separating observations. The results for adapted Pearson correlation after 15 iterations are shown in Fig. 1 for a tradeoff parameter of $f = 0.75$. The top row shows 10 sample spectra for each class (left) and the PCA projection of the labeled complete data set (right). The bottom left panel shows relevances $\lambda_k$, highlighting channels around 40 as highly important and channels around 18 and 78 of certain interest. The data set, rescaled by $\lambda_k$ and plotted as result of multi-dimensional scaling (MDS) of the correlation relationships $(1-{}_r\mathrm{d}^{ij}_{\boldsymbol{\lambda}})$ of the transformed spectra, is shown in the lower right; note that MDS of Euclidean distances would yield PCA. In contrast to unscaled Euclidean space, much better class separation can be observed for the rescaled correlation space.

**Mass spectrum cancer study.** Mass spectroscopy data from a clinical cancer study are analyzed. Frozen tissue material was sliced by a microtom from which, subsequently, 1050 mass spectra where taken using a linear MALDI-TOF MS, Autoflex, in a range of 2000–10000$Da$ (devices and spectra preparation software by Bruker Daltonik GmbH, Bremen). The data preparation protocol of the measured spectra followed the default workflow in ClinProTools 2.1 software (Bruker) for baseline correction, alignment and peak picking. Peaks with signal to noise ratio $> 5$ were used for analysis, and only maxima of the extracted peaks were considered. The final data set contains 32 mass values of 1050 high-quality spectra connected to $\mathrm{Her}^+$ or $\mathrm{Her}^-$ cancer or to normal samples, i.e. to one of three classes.

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

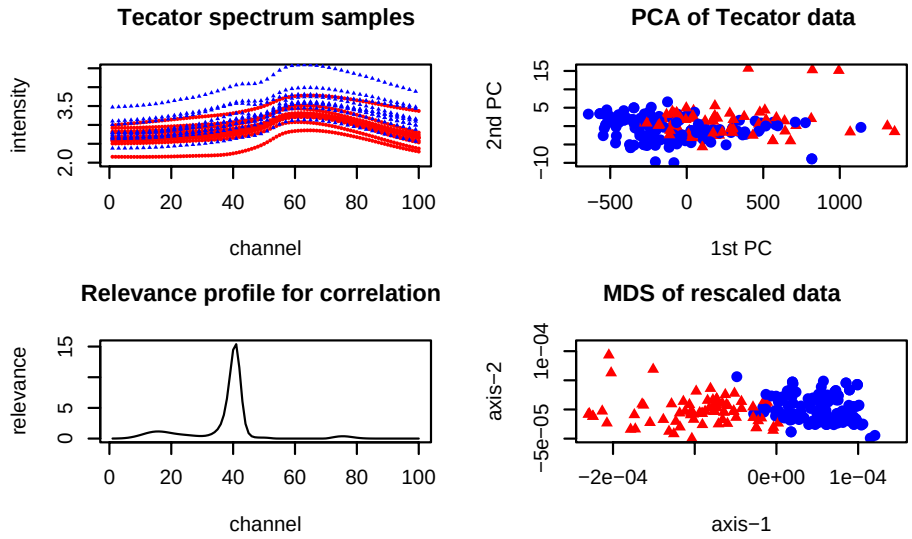


Fig. 1: Feature weighting for Tecator spectral data, high vs. low fat.

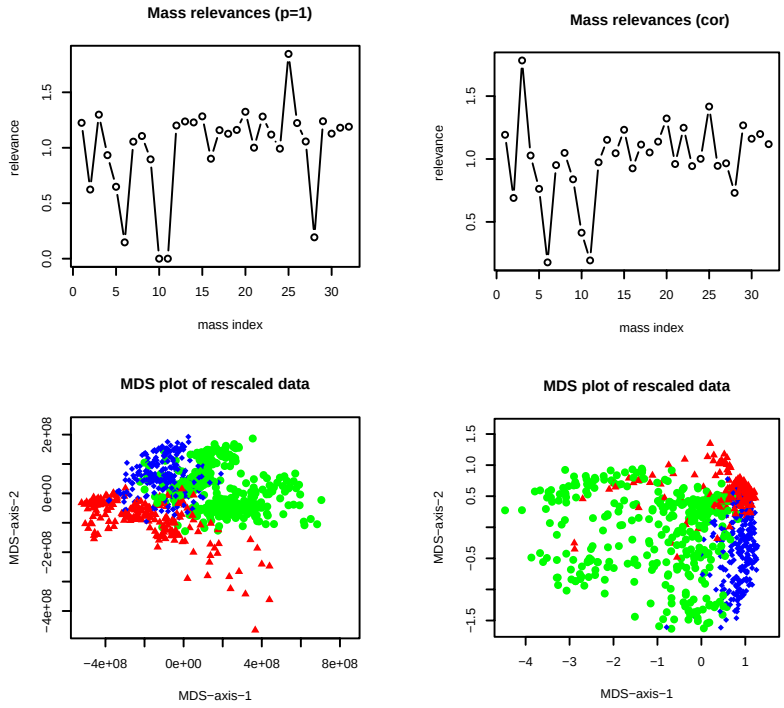


Fig. 2: Mass spectrum data. Left: Minkowski metric. Right Pearson correlation.

As illustrated by Fig. 2, the most interesting masses for improving the separation of three tissue classes depend on the underlying similarity measure, Minkowski metric with exponent $p = 1$ and correlation. After 100 iterations, with $f = 0.61$ for the Minkowski metric and $f = 0.69$ for correlation, chosen according to best visual separation, the MDS-based plots are different, yet, the feature ratings are quite similar. Particularly, indices 3 and 25 point out interesting masses of $2771Da$ and $11693Da$, respectively. Both measures achieve good separation of the initial configurations (not shown). For Minkowski metric, the k-NN classification error ($k = 7$) in Minkowski space drops from 1.9% to 1.5%, in the space of Pearson correlation from 1.5% to 1.1% on 25% test sets. These values along with the previous MDS plots only serve as illustration of the benefits of space rescaling for results of subsequently applied standard methods.

## 4 Conclusions

A new formulation of supervised attribute relevance detection using cross comparisons (SARDUX) has been presented. Generic adaptive similarity measures can be plugged into a structurally very simply algorithm realizing gradient descent for assessing the contribution of data attributes to intra-class compactification and inter-class separation. The tradeoff between compactification and separation can be controlled by a parameter $f$ of which can be selected, for example, by the outcome can be monitored by projection displays. As main result, data sets rescaled by the metric parameters show more class-specific global data arrangements which are helpful in discriminative visualizations, and even the performance of the locally operating k-NN classifier could be improved by rescaling. In future work, a less subjective choice of the tradeoff parameter $f$ will be addressed, and implicit self-normalization of the parameter vector will be implemented for use with generalized Mahalanobis distance.

Thanks for high quality measurements to A. Walch, GSF-Institute for Pathology, Neuherberg and S.-O. Deininger, Bruker Daltonik GmbH, Bremen, Germany. The work is supported by grant XP3624HP/0606T, Ministry of Culture Saxony-Anhalt.

## References

[1] C. Blake and C. Merz. UCI repository of machine learning databases, 1998.

[2] I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh. *Feature Extraction: Foundations and Applications*. Springer, Berlin, 2006.

[3] G. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley-Interscience, 2004.

[4] F. Rossi and N. Villa. Support vector machine for functional data classification. *Neurocomputing*, 69(7–9):730–742, March 2006.

[5] M. Strickert, K. Witzel, H.-P. Mock, F.-M. Schleif, and T. Villmann. Supervised attribute relevance determination for protein identification in stress experiments. In *Proceedings of Machine Learning in Systems Biology (MLSB 2007)*, pages 81–86, 2007.

[6] Y. Sun. Iterative relief for feature weighting: Algorithms, theories, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1035–1051, 2007.