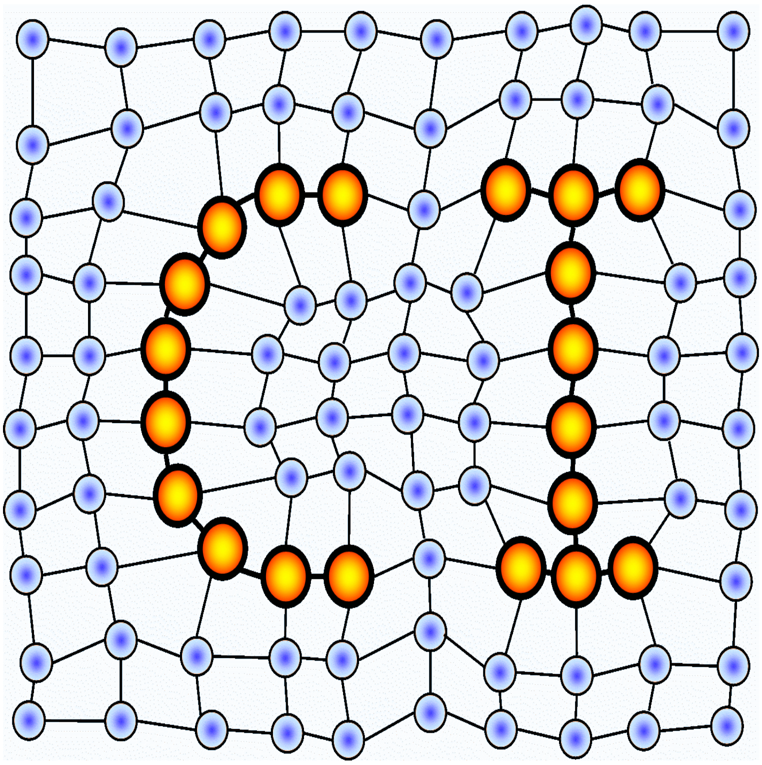
MIWOCI 2010, Mittweida Workshop on Computational Intelligence

F.-M. Schleif, T.Villmann (Eds.)

Machine Learning Reports **MLR-2010-05**



MIWOCI 2010, MITTWEIDA WORKSHOP ON COMPUTATIONAL INTELLIGENCE

Impressum

Publisher: University of Applied Sciences Mittweida
Technikumplatz 17,
09648 Mittweida, Germany
Editor: Prof. Dr. Thomas Villmann
Dr. Frank-Michael Schleif

Techical-Editor: Dr. Frank-Michael Schleif
Contact: schleif@informatik.uni-leipzig.de
URL: http://www.uni-leipzig.de/compint/mlr/mlr.html
ISSN: 1865-3960

Contents

2nd Mittweida Workshop on Computational Intelligence

F.-M. Schleif[1]

1 The 2nd Mittweida Workshop on Computational Intelligence (MiWoCi'2010)

From June 29th to July 03rd, 2010, 12 scientists from the University of Bielefeld, University of Siegen, University of Groningen (NL), the University of Applied Sciences Mittweida, the Fraunhofer Insitute FIT and the Fraunhofer Institute IFF met in Mittweida, Germany to continue the tradition of the Mittweida Workshops on Computational Intelligence. The aim was to present their current research, discuss scientific questions, and exchange their ideas. The seminar centered around topics in machine learning, signal processing and data analysis, covering fundamental theoretical aspects as well as recent applications, partially in the frame of innovative industrial cooperations. This volume contains a collection of extended abstracts which accompany these talks to give some insight into the research presented in Mittweida.

Apart from the scientific merrits, this year's seminar came up with a few highlights which demonstrate the excellent possibilities offered by the surroundings of Mittweida. Feared by the last year journey in the caves and mines of muria the this year adventures were explored under sunlight. The participants climbed to the high forests of Mittweida (Kletterwald) and enjoyed the exciting and fearing adventures provided on the top of the trees. During a *wild water* journey (Paddeltour) the outstanding fitness of the researcher was demonstrated and some of them also demonstrated their braveness by swimming in the rapids.

Our particular thanks for a perfect local organization of the workshop go to Thomas Villmann as spiritus movens of the seminar.

Bielefeld, November, 2010
Frank-M. Schleif

[1] E-mail: `fschleif@techfak.uni-bielefeld.de`

[2] University of Bielefeld, CITEC, Chair of Theoretical Computer Science, Leipzig, Germany

Linear Data Association Mapping by Supervised Metric Adaptation

Marc Strickert[1]

Keywords: Matrix metric learning, supervised linear mappings, generalized regression, association mapping, distance matrix correlation, dimension reduction.

1 Introduction

Linear mappings are ubiquitous in all areas of science. They are often used for the reduction of data dimensionality with visualization being a special case, and for inverse problems where coefficients for the superposition of known vectors to an observed mixture are sought. The results are well interpretable as a linear mixture of data attributes. Since, in many cases, algebraic solutions can be obtained, linear methods are supposed to be fast compared to nonlinear or iterative methods such as artificial neural networks [11]. The 'zoo' of linear methods includes popular approaches like

PCA Principal component analysis [9] projects vector data to axes of maximum variance, a concept connected to Euclidean space.

PCOA Principal coordinates analysis [15] turns a distance matrix into a cloud of points in Euclidean space where the original distances are approximated.

ICA Independent component analysis [10] seeks for unmixing linearly superimposed data using the assumption of independent non-Gaussian sources.

PP Projection pursuit [14] iteratively finds projections according to a custom PP criterion permitting concepts like information-optimized data spreading and class label-driven separation.

LDA Linear discriminant analysis [8] projects data on the axis best separating the class categories connected to the data vectors.

[1] University of Siegen, Institute for Vision and Graphics,
E-mail: strickert@informatik.uni-siegen.de

GLS Generalized least squares regression [13] allows to associate the input vectors with real-valued dependent variables, thereby accounting for heteroscedasticity of the data.

CCA Canonical correlation analysis [7] transforms both vectors from input space and a their dependent vectors to a common subspace with maximum correlation of the projections.

The ordering of these methods roughly reflects the amount of prior knowledge about the data that can be integrated into the model, ranging from implicit assumptions like the level of variance (PCA) via discrete labels (LDA) to associated vector spaces (CCA). This corresponds to label-free and label-based, i.e. unsupervised and supervised, data processing.

Generally, input data vectors and their dependent categories, variables, or vectors can be considered as association context. Another level of abstraction is the connection of an input data relationship matrix to a target relationship matrix, with relationships being distances, divergence measures or (dis)similarity relations. Here, the input vector relationships are quantified by an adaptive data metric, while the output relationships are calculated as Euclidean distance for intuitive interpretation or a symmetric dissimilarity matrix.

2 Linear mappings from matrix metric adaptation

Recently the idea of adaptive subspace mapping was proposed [1] which seeks for a mapping transformation of N M-dimensional data vectors $\mathsf{x}^j \in \mathsf{X} \subset \mathbb{R}^M, \mathsf{x}^j = (x^j_k)_{k=1\dots M}, 1 \leq j \leq N$ in such a way that their pairwise distances $\mathbf{D}^{\boldsymbol{\lambda}}_{\mathsf{X}}$ are in maximum correlation with those distances $\mathbf{D}_{\mathsf{L}}$ defined on the label space L with q-dimensional labels $\boldsymbol{l}^j \in \mathsf{L} \subset \mathbb{R}^q, \boldsymbol{l}^j = (l^j_k)_{k=1\dots q}$, i.e.

$$\mathsf{r}(\mathbf{D}_{\mathsf{L}}, \mathbf{D}^{\boldsymbol{\lambda}}_{\mathsf{X}}) = \max\,. \tag{1}$$

Therein, $\mathbf{D}^{\boldsymbol{\lambda}}_{\mathsf{X}}$ is the matrix of adaptive input vector distances which depend on the parameter matrix $\boldsymbol{\lambda} \in \mathbb{R}^{M \times u}$ to maximize the Pearson correlation (r) between the distance matrices of input and label space. For two input vectors $\boldsymbol{x}^i$ and $\boldsymbol{x}^j \in \mathsf{X}$ being column vectors the data-driven adaptive distance is defined as

$$(\mathbf{D}^{\boldsymbol{\lambda}}_{\mathsf{X}})_{i,j} = \sqrt{(\boldsymbol{x}^i - \boldsymbol{x}^j)^{\top} \cdot \boldsymbol{\lambda} \cdot \boldsymbol{\lambda}^{\top} \cdot (\boldsymbol{x}^i - \boldsymbol{x}^j)}. \tag{2}$$

Despite of its similarity to the Mahalanobis distance the rank of $\boldsymbol{\Lambda} = \boldsymbol{\lambda} \cdot \boldsymbol{\lambda}^{\top}$ is in practice not of full rank but $u \ll M$. Instead of learning $\boldsymbol{\Lambda}$ the matrix $\boldsymbol{\lambda}$ is adapted to ensure that $\boldsymbol{\Lambda}$ gets positive semi-definite and thus Eqn. 2 is a metric.

Since Eqn. 2 can be decomposed into transposed and non-transposed linear mappings $\boldsymbol{\lambda}^{\top} \cdot \boldsymbol{x}^k$, choices of $u \leq 3$ could be used for visualizations of the data space mapped according to the requirements imposed by the label space. Locally optimum solutions for $\boldsymbol{\lambda}$ are by obtained by maximizing Eqn. 1, for example, by using the memory limited quasi Newton Broyden-Fletcher-Goldfarb-Shanno method.

The *run time* depends a lot on the data. For example, many equidistant points lead to an ill-structured distance matrix, preventing efficient correlation-based optimization. Of course, the size of the data set is another important factor. While the label distance matrix is only calculated once, the distance matrix of the transformed data is calculated repeatedly. This yields a run time and memory complexity of $\mathcal{O}(u \cdot M \cdot N^2)$. Despite of this potential run time and memory bottleneck, the proposed method is very appealing in many domains of application. Specifically, computations may need to be carried out only once, or existing mappings may just be re-adjusted with a few iterations for new data. Once computed, the mappings are very fast in the applications.

3 Matrix Initialization

The simplified expression $\boldsymbol{x}^{\mathsf{T}} \cdot \boldsymbol{\Lambda} \cdot \boldsymbol{x}$ taken from Eqn. 2 describes the mixing of the k-th and m-th attribute of $\boldsymbol{x}$ by the matrix components $(\boldsymbol{\Lambda})_{km}$. The identity matrix $\boldsymbol{\Lambda} = \boldsymbol{E}$ leads to the squared Euclidean norm of $\boldsymbol{x}$, accounting only for attributes paired with themselves, while $\boldsymbol{\Lambda} = \frac{1}{2} \cdot (\boldsymbol{E} + \mathbf{1})$ leads to an equal contribution of all attribute pairs.

Without prior knowledge both choices are desirable options for starting the matrix adaptation in an unbiased way. Yet, $\boldsymbol{\lambda}$ gets adapted, not $\boldsymbol{\Lambda}$, which leads to low ranks of $\boldsymbol{\Lambda} = \boldsymbol{\lambda} \cdot \boldsymbol{\lambda}^{\mathsf{T}}$ being incompatible with full rank of the identity matrix or its average with the unity matrix. Despite this clash the discrepancy between the possible and the desired initialization matrix should be minimum. A gradient descent can be used for an initialization approach targeting all attribute pairs, starting with a random matrix $\boldsymbol{V}$ on the cost function

$$S = \left\| \boldsymbol{V} \cdot \boldsymbol{V}^{\mathsf{T}} - \frac{1}{2} \cdot (\boldsymbol{E} + \mathbf{1}) \right\|_{\mathsf{F}}^{2} . \qquad (3)$$

Therein, the squared Frobenius norm $\| \cdot \|_{\mathsf{F}}^2$ is used to express a minimum least square approach for optimizing V_{km}. The final matrix $\boldsymbol{V}$ is used to initialize $\boldsymbol{\lambda}$ for the desired metric adaptation.

Certainly, rank discrepancies will always lead to sub-optimum solutions with $S > 0$, but the optimization will distribute the mismatches evenly over the matrix, such that for each column and row the same accumulated mismatch will occur. In other words: mismatches are distributed equally over all data attributes and, consequently, the sum of all mixing coefficients per attribute, i.e. the initial influence of each attribute, is the same, which is a desirable property.

Note that for a large number of data attributes such as $M > 10\,000$ the involved matrices get quite big, but the structurally very simple cost function in Eqn. 3 allows a C or CUDA solution with a memory footprint of $\mathcal{O}(1)$. Furthermore, in case of $\boldsymbol{\lambda}$ mapping M-dimensional data to a one-dimensional target subspace, i.e. for regression or binary classification problems, an optimum analytical initialization is the vector $\lambda_i = \sqrt{((M-1)/2+1)/M}, i = 1 \ldots M$.

4 Matrix Interpretation

For the interpretation of the finally obtained parameter matrix $\boldsymbol{\lambda}$ it is natural to look at the mixing matrix $\boldsymbol{\Lambda} = \boldsymbol{\lambda} \cdot \boldsymbol{\lambda}^{\mathsf{T}}$ connected to the learned metric expressed by Eqn. 2. Basically, large absolute values $|\Lambda_{ij}|$ denote important contributions of attribute pairs i, j to the given association task.

Yet, the covariance level of the attributes affect the magnitude of the mixing factors, i.e. two merely noisy attribute pairs would carry the same semantic of being to the same degree irrelevant, but their corresponding mixing magnitudes are inversely related to their different variances in order to suppress these two noise sources. Thus, a scaled mixing matrix that erases the *covariance structure* the same ways as if the data was whitened before matrix adaptation is obtained by inserting $\boldsymbol{K}$ into the central expression of Eqn. 2 to $\boldsymbol{x}^{\mathsf{T}} \cdot [\boldsymbol{K} \cdot (\boldsymbol{\lambda} \cdot \boldsymbol{\lambda}^{\mathsf{T}}) \cdot \boldsymbol{K}^{\mathsf{T}}] \cdot \boldsymbol{x}$, where $\boldsymbol{K} \cdot \boldsymbol{K}^{\mathsf{T}} = \mathsf{cov}(\mathbf{X})$ is the $M \times M$ covariance matrix of the data attributes. Thus, $\boldsymbol{K} = \mathsf{cov}(\mathbf{X})^{1/2}$ can be multiplied to the parameter matrix $\boldsymbol{\lambda}$ prior to calculating the transformed mixing matrix $\boldsymbol{\Lambda}$.

Likewise the influence of attribute variances, disregarding further covariance structure, can be removed by multiplying the k-th row of $\boldsymbol{\lambda}$ by the standard deviation of the k-th data attribute prior to calculating $\boldsymbol{\Lambda}$.

5 Application domains

The proposed method provides a very flexible approach to tasks including classification, regression, and faithful data visualization, as will be demonstrated in the examples below. Currently, the method is able to deal with data sets of medium size such as 1000 samples by 1000 dimensions, 100 samples by $10\,000$ dimensions, or $10\,000$ samples by 100 dimensions. This allows to deal with important practical applications in bioinformatics, where some thousands of metabolite abundances or gene expression intensities are available, but only some tens or hundreds of samples. Also spectrum-based medical analyses or technical diagnosis tasks can be approached.

As an important feature, the label space, i.e. the costly and usually reliable part of the measurements will be unchanged. Complementarily, many screening technologies deliver massive array data or wide-band spectra with partially unnecessary or redundant information to be transformed for better matching the label space.

As additional benefit, missing values can be quite easily dealt with. In the label data, only few missing values are to be expected anyway, and the distance matrix can still be approximated despite of some missing attributes. In the set of data measurements missing data attributes can be also easily handled, because the matrix product in Eqn. 2 can be decomposed into many scalar products for which missing values set to zero do not contribute to the final value. This way, missing and undesired values can be conveniently masked out.

To give some notes on the selection of the dimensionality u of the subspace. From the global perspective, in a t-class classification problem a no more than $(t-1)$-dimensional Euclidean subspace should suffice to host the corresponding category simplex. In other words, the number of classes, i.e. the number of dimensions of the label space, is a convenient upper bound for the rank of the metric's matrix. Alternatively, for an auto-encoding problem with N samples and M dimensions, the maximum required dimension of a subspace for a perfect linear mapping is $\min(M, N) - 1$. In other words, you can perfectly map a 10 000-dimensional array data set containing 100 samples into a 99 dimensional subspace without loosing information about the relationships between the samples. This offers a great compression potential if only relational classification and clustering methods are subsequently used.

It is important to note that the proposed method does not provide a classification system; it just yields reasonable transformations for being subsequently used by existing classifiers. Recently, integrated methods have been proposed to combine matrix learning and classification, such as limited rank learning vector quantization (LiRaM-LVQ) [3, 2] and the large margin nearest neighbor classifier [6]. In the two-class case, the method presented here allows to generate receiver operating characteristic (ROC) curves, because a threshold can be easily varied between those results obtained for the solved regression task.

Matrix inversion might be considered as another application by just setting the label space to the identity matrix. Unfortunately, this does not work, because the off-diagonal elements of the distance matrix of the identity matrix are all one, and due to an induced zero denominator in the correlation term proper optimization becomes impossible.

6 Examples for spectral data

A benchmark spectral data set taken from the StatLib repository of machine learning [12] is taken for illustrating several features of adaptive subspace mapping. The data set contains 215 samples of 100-dimensional infrared absorption spectra recorded on a Tecator Infratec Food and Feed Analyzer working at a wavelength range of 850–1050*nm* in Near Infrared Transmission (NIT) mode. An overview of the spectra is given in Fig. 1.

The regression problem consists of predicting fat content of meat from these publicly available spectra. The benchmark data comes split into training set (samples 1–172) and test set (samples 173–215). The following three subsections address illustrative tasks related to this data set, only varying in the use of prior knowledge: (a) only related to the real-valued variable of fat content, (b) related to the fat content discretized into three disjoint classes of about the same size, and (c) related to the data set itself.

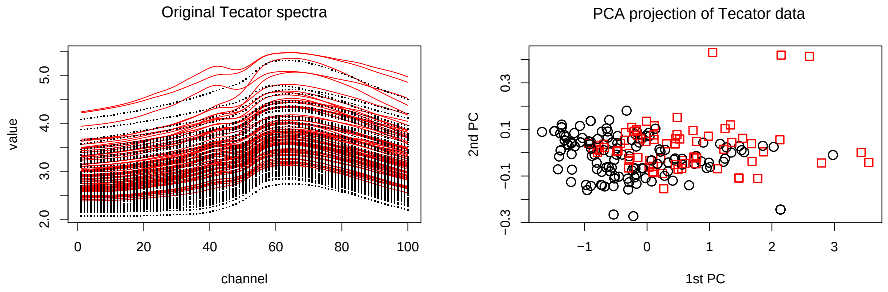


Figure 1: Tecator spectra, raw (left) and a PCA projection (right). For illustration, fat content below a value of 20% is displayed by black dashed lines, high fat content above 20% is indicated by red solid lines.

6.1 Multiple regression

The dependent real-valued variable of fat content is provided along with the data set. Prior to optimization the parameter vector $\boldsymbol{\lambda}$ is initialized with identical components. The results are compared with the left division operator ’\’, $\boldsymbol{\lambda} = \mathbf{X}^{\mathsf{T}} \backslash \mathbf{L}$, and the more stable Moore-Penrose pseudoinverse ’pinv’ $\boldsymbol{\lambda} = \mathsf{pinv}(\mathbf{X}^{\mathsf{T}}) \cdot \mathbf{L}$, both available in MATLAB and GNU Octave. A summary of the correlation results of the models $\mathsf{r}(\mathbf{L}, \boldsymbol{\lambda}^{\mathsf{T}} \cdot \mathbf{X})$ is given in Tab. 1.

r	ASM	Octave V3: ’pinv’	Octave V3: ’\’
train	0.9875	**0.9980**	0.9964
test	**0.9879**	0.9595	0.9022

Table 1: Regression results as Pearson correlation of the new method compared to two different approaches based on matrix pseudoinverse calculations. Best values for training and test set are in highlighted in bold.

On the training data set, the proposed model, based on distance matrices rather than the data and label vectors, performs relatively poorly compared to the other two methods. For example, the Moore-Penrose approach ’pinv’ provides close to optimum ($\mathsf{r} = 1$) solutions. Yet, the new model provides by far the best generalization, expressed by both the high correlation values and by the small discrepancy between training and test performance. Similar observations not reported here were made for other data sets too. In contrast to plain data matrix inversion the squared number of pairwise relationships contained in distance matrices may lead to a more representative model, compensating for outliers or too specific traits. Thus, the proposed model is empirically well-suited for dealing with a few number of samples of high-dimensional data.

The model allows to interpret $\boldsymbol{\lambda} \cdot \boldsymbol{\lambda}^{\mathsf{T}}$ as mixing matrix. This does also apply to the matrix inversion models, because $\mathbf{X}^{\mathsf{T}} \cdot \boldsymbol{\lambda} \approx \mathbf{L} \Rightarrow \mathbf{X}^{\mathsf{T}} \cdot \boldsymbol{\lambda} \cdot \boldsymbol{\lambda}^{\mathsf{T}} \cdot \mathbf{X} \approx \mathbf{L} \cdot \mathbf{L}^{\mathsf{T}}$ being transfered to $(\boldsymbol{x}^i - \boldsymbol{x}^j)^{\mathsf{T}} \cdot \boldsymbol{\lambda} \cdot \boldsymbol{\lambda}^{\mathsf{T}} \cdot (\boldsymbol{x}^i - \boldsymbol{x}^j) \approx (l^i - l^j)^2$ yields the corresponding squared distance matrices for $i, j = 1 \ldots N$. Four mixing matrices with absolute entries are shown in Fig. 2. The columns in Tab. 1 correspond to the top left, bottom right, and bottom left sub-panels, respectively. These matrices share the properties of more intensity of pairwise channels around 40, 50, and 80, and some suppression in the right upper area corresponding to pairwise intervals between 55 and 75. The overall structures are somewhat different though. Interestingly, a much more pronounced and smoother structure is displayed in the top right matrix panel of Fig. 2 for the parameter vector transformed according to whitened data.

Based on the mean square error measure, the model performance is further compared with a multivariate regression model for functional data (FMR) [4]. Two comparison variants are studied here: one with original data, and one with z-score transformed spectra as used in the reference publication. Note that this spectrum z-score cannot be realized like the attribute z-score by a linear transformation, because each spectrum mean and variance require completely independent calculations.

Results are summarized in Table 2. Basically, FMR provides the best test set result, although the noticeable drop of error compared to the training set is not explained by the authors. Using non-transformed data the proposed methods performs well in terms of training, and it is average for the test data. These results are completely different from those for the z-score transformed spectra which result in a very low training error and an exceedingly high test set error.

For a better understanding of this strong discrepancy the original data was checked. It turned out that one sample (sample number 13 in the test set, being spectrum number 185 in the whole data set) exhibited both maximum mean and variance values. Consequently these values create maximum impact during the z-score transformation. A PCA plot of the non-transformed test set (not shown) confirms a very outlier status of that sample. Despite of being extremal terms of the mean, the variance, and the location in the PCA scatter plot, its associated meat content value of 34.8, representing the overall 85% percentile threshold, is not extremal. If these rather handwaving arguments are used to eliminate only that critical data point, the test set MSE drops down from 9.42 to an excellent value of 2.26. In a fair comparison this removal should not be done, but we will come across that data point again in the next section. Finally, the authors' indication of channels 60–80 being relevant to fat content association [4] is in contrast to the identified channels around 40 being of interest using the method proposed here. This trivially reminds us that attribute importance is strongly connected to the method of choice and not an intrinsic property of the data.

MSE	ASM	ASM (z-score)	MR	FMR
train	2.43	**1.32**	2.94	4.27
test	3.73	9.42 / 2.26*	4.00	**3.52**

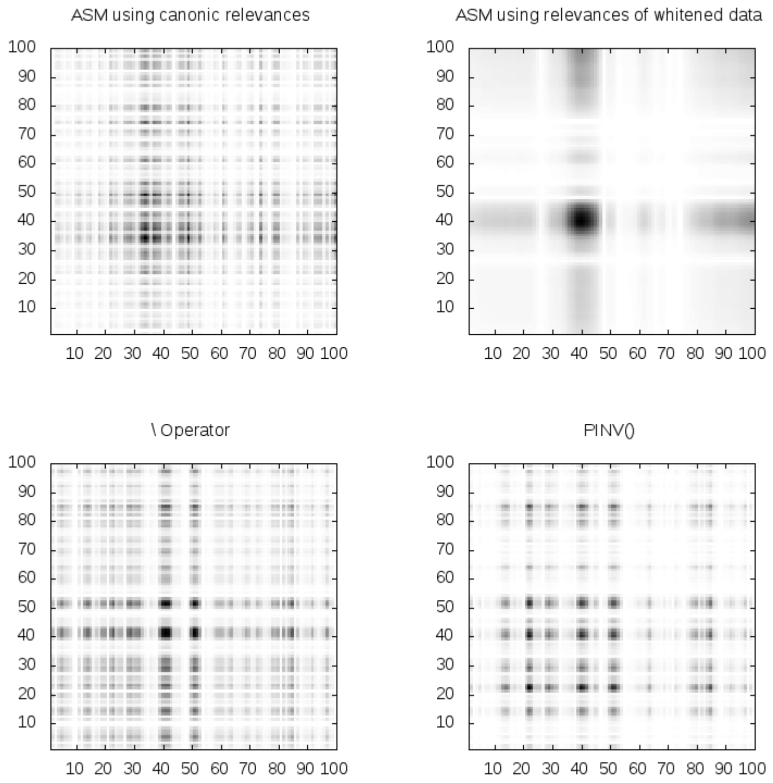
Table 2: Mean square errors, from left to right for the proposed method using original data and z-score transformed spectra, using standard multivariate regression (MR) and functional MR (FMR) [4]. The result indicated by * refers to a test set with spectrum number 13 (185 in complete data set) eliminated.

6.2 Multiclass subspace regression

In the previous section a very good association of the spectra and their corresponding meat content values was shown. Here, an artificial problem is defined by creating three partition labels for low, middle, and high value percentiles, i.e. a split of the training data set into 57:58:57 samples. The two partition boundaries are used to also categorize the test labels. Instead of the ordinal categories 0, 1, and 2 an orthogonal encoding of $\boldsymbol{a} = (0, 0, 1)$, $\boldsymbol{b} = (0, 1, 0)$, and $\boldsymbol{c} = (1, 0, 0)$ is chosen in order to provide maximum independence of the labels. The task is to map the data into a subspace with equidistant data arrangement between the three categories, i.e. corners of an equilateral triangle. Obviously such target arrangement fits into a 2D space, thus, motivating to a choice of $u = 2$ columns for the parameter matrix $\boldsymbol{\lambda}$.

A representative result of the optimization is shown in Fig. 3. No equilateral triangle, yet, a faint triangular scattering can be observed in the left plot, leading to a relatively poor correlation of $\mathsf{r} = 0.68$ of the projection and the label space distances. According to the 'nature' of the data to associate with real-valued fat content, the categories are arranged in a rather cloudy structure along the original partition ordering. Still the clouds of the training categories and of the a posteriori mapped test set spectra do not overlap strongly. Getting back to ordinal categories of high, middle, and low fat content, a better separation of the high label (green) from the middle (red) and low (blue) content can be observed than for the red and the blue category. The green outlier point in the top left is again the critical test spectrum number 13 identified for z-score transformed data in the previous section.

Looking at the mixing matrix in the right panel of Fig. 3, we find an intensity distribution similar but more structured than for the regression task displayed in Fig. 2. This observation coincides with the fact of a rank-2 matrix allowing more structure than the just the rank-1 matrix needed for scalar regression tasks. To conclude, the proposed method cannot override the intrinsic characteristics of the data, but it provides considerable results to fulfill the desired, maybe ill-posed, mapping problem.



ASM using canonic relevances

ASM using relevances of whitened data

\ Operator

PINV()

Figure 2: Tecator attribute mixing matrices with absolute mixing components. Darker pairs have more influence on the regression task.

6.3 Multivariate regression for multi-dimensional scaling

In this section traditional label information is ignored. Instead an interesting task is to associate data vectors with themselves by setting $\mathsf{L} = \mathsf{X}$. This is a non-trivial problem, because it requires finding an u-dimensional subspace in which the high-dimensional data relationships are best approximated. This is very similar to the principal coordinate extraction problem (PCOA), but instead of a mere reconstruction of a given distance matrix, a linear mapping operation is sought such that adding one point leads to mapping that point rather than to compute new reconstruction of an $(N+1) \times (N+1)$ matrix.

Here, rather than aiming at the Euclidean distance with $\mathsf{L} = \mathsf{X} \rightarrow \mathsf{r}(\mathbf{D}_{\mathsf{L}}, \mathbf{D}_{\mathsf{X}}^{\boldsymbol{\lambda}}) = \max$ a dissimilarity matrix $\mathbf{D}_{\mathsf{L}}$ is defined based on the 1 minus Kendall correlation values

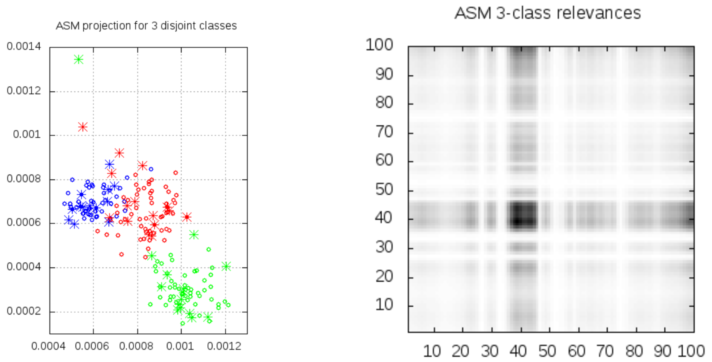


Figure 3: Three-class association for Tecator data set. Left: scatter plot of 2D projection space. Colored points refer to blue=low, red=medium, and green=high fat content. Small points denote the training set, large points the test set. Right: corresponding attribute mixing matrix, cf. Fig. 2 top right.

between all pairs of spectra. Thus, an adapted Mahalanobis-like distance is wanted of which the induced linear mapping approximates the given correlationships in $\mathbf{D}_{\mathsf{L}}$. The results of the optimization is shown in the right panel of Fig. 4, providing a correlation of $\mathsf{r}(\mathbf{D}_{\mathsf{L}}, \mathbf{D}^{\lambda}_{\mathsf{X}}) = 0.958$. For comparison, a PCOA reconstruction based on MDSLocalize [5] is shown in the left panel of Fig. 4. A strong visual correspondence between both scatter plots can be stated, although the PCOA approach yields a much higher correlation of $\mathsf{r}(\mathbf{D}_{\mathsf{L}}, \mathbf{D}^{\lambda}_{\mathsf{X}}) = 0.996$. This very good reconstruction is achieved at the cost of losing the data mapping functionality. In practice many algorithms for computing the Kendall correlation coefficient possess an $\mathcal{O}(M^2)$ time complexity, which gets excessive for some thousand-dimensional spectra. Thus, turning the reconstruction problem into a linear mapping operation, new data points can be integrated at very low costs. This feature makes the proposed optimization method attractive for processing large incoming data sets exhibiting complex relationships between the samples.

7 Summary and Conclusions

A generalized linear projection pursuit method has been discussed for illustrating the open potential for using linear mappings in tasks aiming at the association of data transformable vectors with their constant label space. In the current version, for a given training set the set size must remain constant during training because of the underlying batch optimization scheme. Optimization is rather stable, because no critical numerical operation like matrix inversion is needed. Training itself is limited to medium size data

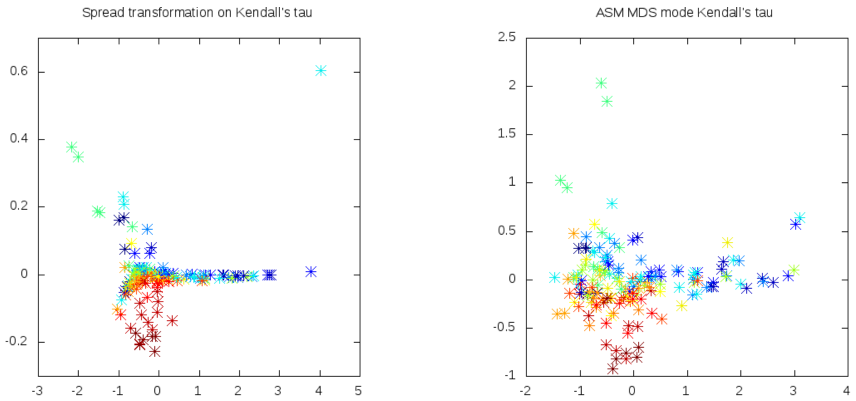


Figure 4: Low-dimensional representation of the 1 minus Kendall correlation dissimilarity matrix of Tecator spectra. Left: Reconstruction by an MDSLocalize-based spread transformation. Right: Mapping result of the proposed method. Colors representing fat content (red=high, green=medium, blue=low) are only for orientation purposes and don't affect computations.

sets, because of the squared complexity induced by the utilization of distance matrices. After optimization, though, mappings of new data vectors can be easily computed by applying the trained linear model. In addition, the distance matrix approach allows for labels to be expressed as pairwise relationships rather than as associated scalar values or vectors.

In practice, initialization of the mapping is not a critical issue, because the convergence is usually quite good. Still, a way of initialization was presented compensating for the lack of rank by distributing overall attribute mixing equally to all attributes.

The interpretation of the linear models is supposed to be easier by looking at the attribute mixing matrix $\boldsymbol{\Lambda} = \boldsymbol{\lambda} \cdot \boldsymbol{\lambda}^{\mathsf{T}}$ rather than at the obtained parameter matrix $\boldsymbol{\lambda}$. Absolute entries of $\boldsymbol{\Lambda}$ can be related to the strength of their attribute pairs to contribute to the association task. Additionally to the original data view, alternative views compensating for attribute variance and covariance have been indicated.

The well-known Tecator spectral benchmark data set has been used for illustrating the versatility of the method by doing classical regression, multi-class regression, and a self-association auto-encoder task. These examples are not yet very systematic, but they give an initial impression of the potential of the proposed optimization scheme for the supervised calculation of linear mappings.

Acknowledgments

This work was started during a fruitful cooperation project with Gustavo Vazquez and Axel Soto from the Universidad del Sur, Argentina, gratefully funded by BMBF/IB grant ARG 08/016. The continuation of this work was possible thanks to fundings in the DFG Graduiertenkolleg 1564 "Imaging New Modalities". Many thanks to Prof. Thomas Villmann for organizing and hosting the inspiring 2nd Mittweida Workshop on Computational Intelligence 2010.

References

[1] Marc Strickert, Axel J. Soto, and Gustavo E. Vazquez. Adaptive matrix distances aiming at optimum regression subspaces. In Michel Verleysen, editor, *European Symposium on Artificial Neural Networks (ESANN)*, pages 93–98. D-facto Publications, 2010.

[2] Petra Schneider, Michael Biehl, and Barbara Hammer. Adaptive relevance matrices in learning vector quantization. *Neural Computation*, 21(12):3532–3561, 2009. PMID: 19764875.

[3] Petra Schneider, Kerstin Bunte, Barbara Hammer, Thomas Villmann, and Michael Biehl. Regularization in matrix relevance learning. Tech. rep. mlr-02-2008, University of Leipzig, http://www.uni-leipzig.de/˜compint/mlr/mlr_02_2008.pdf, 2008.

[4] Hidetoshi Matsui, Yuko Araki, and Sadanori Konishi. Multivariate regression modeling for functional data. *Journal of Data Science*, 6(3):313–331, 2007.

[5] Marc Strickert, Nese Sreenivasulu, and Udo Seiffert. Sanger-driven MDSLocalize - A comparative study for genomic data. In Michel Verleysen, editor, *European Symposium on Artificial Neural Networks (ESANN)*, pages 265–270. D-facto Publications, 2006.

[6] Kilian Weinberger, John Blitzer, and Lawrence Saul. Distance metric learning for large margin nearest neighbor classification. In Yair Weiss, Bernhard Schölkopf, and John Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1473–1480. MIT Press, Cambridge, MA, 2006.

[7] David R. Hardoon, Sandor R. Szedmak, and John R. Shawe-taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[8] Geoffrey J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley-Interscience, 2004.

[9] Ian T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 2nd edition, 2002.

[10] Aapo Hyvärinen and Erkki Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4–5):411–430, 2000.

[11] Simon Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 2nd edition, 1999.

[12] Pantelis Vlachos. *StatLib Data Repository*. Carnegie Mellon University, http://lib.stat.cmu.edu/, 1998.

[13] Guido del Pino. The unifying role of iterative generalized least squares in statistical algorithms. *Statistical Science*, 4(4):394–403, 1989.

[14] Jerome H. Friedman and Werner Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76:817–823, 1981.

[15] John C. Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53:325–338, 1966.

Methods for validating the results of fuzzy cluster and classification algorithms

Tina Geweniger[1,2]

1 Introduction

In the course of our research we developed a variety of fuzzy cluster and classification algorithms like Median Fuzzy c-Means (M-FCM, [1]), Fuzzy Robust Soft Learning Vector Quantization (FRSLVQ, [2]), or Fuzzy Neural Gas (FNG, [7]). Some of them work with fuzzy data sets, others result in fuzzy cluster or classification solutions. While trying to compare the results, either for several trial runs or for different methods, we noticed a lack of methods to measure the quality of the solutions. Inspired by the discussion following my presentation at the workshop I discovered a couple of publications which are addressing exactly this topic. The purpose of this paper is to give an overview of already available evaluation measures for fuzzy data sets and comment on them if applicable.

2 Summary of available evaluation methods

To compare fuzzy cluster or classification results it is necessary to use special measures designed to handle fuzzy data. The measures or indexes listed in this chapter commonly are derived from their crisp versions and employ *t-norms* to handle the fuzzy aspects. In the following subsections a short description of the original measure and its adaption to fuzzy data is given. Detailed descriptions can be found in the referenced articles.

2.1 Kappa value

Cohen's κ_C [10] and Fleiss' κ_F [8] are two statistical measures of inter-rater agreement of two (Cohen) ore more (Fleiss) crisp classifiers C_1 and C_2 or C_1 to C_M respectively, taking into account the agreement occurring by chance. They are both given by

[1] E-mail: `tina@geweniger.org`

[2] University of Applied Sciences Mittweida, AG Computational Intelligence, Germany

$$\kappa = \frac{p_0 - p_c}{1 - p_c}$$

where p_0 is the relative agreement among the classifiers and p_c is the expected agreement by chance, which is the expected value of the joined event that the classifiers classify a data point to the same one of the C class. In [5] and [4] the fuzzy versions of both Kappa values have been derived. p_0 and p_c are then defined as

Fuzzy Cohen's Kappa

$$\begin{aligned} p_0 &= \frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{C} T(\mu_i^{C_1}(\mathbf{x}_k), \mu_i^{C_2}(\mathbf{x}_k)) \\ p_c &= \sum_{i=1}^{C}\int_{\mu_i^{C_1}=0}^{1}\int_{\mu_i^{C_2}=0}^{1} p(\mu_i^{C_1})\cdot p(\mu_i^{C_2}) T(\mu_i^{C_1}, \mu_i^{C_2}) d\mu_i^{C_2} d\mu_i^{C_1} \end{aligned}$$

Fuzzy Fleiss' Kappa

$$\begin{aligned} p_0 &= \frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{C} T(\mu_i^{C_1}(\mathbf{x}_k), \ldots, \mu_i^{C_M}(\mathbf{x}_k)) \\ p_c &= \sum_{i=1}^{C}\int_{\mu_i^{C_1}=0}^{1}\cdots\int_{\mu_i^{C_M}=0}^{1} \left(\prod_{j=1}^{M} p(\mu_i^{C_j})\right)\cdot\Big(T(\mu_i^{C_1}, \ldots, \mu_i^{C_M})\Big) d\mu_i^{C_1}\ldots d\mu_i^{C_M} \end{aligned}$$

where $\mu_i^C(\mathbf{x})$ is the fuzzy assignment of data point $\mathbf{x}$ to class i. N is the number of data points and M the number of classes or clusters. The function $T : [0,1]^2 \to [0,1]$ is a *t-norm* and the values $p_i^{C_1}$ to $p_i^{C_M}$ the densities $p_i^{C_j} = \frac{1}{N}\sum_{k=1}^{N}\mu_i^{C_j}(\mathbf{x}_k), j = 1\ldots M$.

For both kappa the relation $\kappa \in [-1, 1]$ is valid and the values are interpreted according to the scheme given in table 1.

Remark 1 Using different *t-norms* like Minimum, Sum or Łukasiewicz leads to different Kappa values. In [4] it is recommended to use the Minimum norm.

Remark 2 If this measure is to be used to compare two or more cluster solutions with each other, all the possible permutations of the clusters have to be considered.

2.2 Rand Index and related indexes

A further criterion for evaluating cluster or classification solutions is the Rand Index as proposed by Rand [9]. This measure compares pairs of objects, counts their agreements

κ-value	meaning
κ<0	poor agreement
$0 = \kappa \leq 0.2$	slight agreement
$0.2 < \kappa \leq 0.4$	fair agreement
$0.4 < \kappa \leq 0.6$	moderate agreement
$0.6 < \kappa \leq 0.8$	substantial agreement
$0.8 < \kappa \leq 1$	perfect agreement

Table 1: Interpretation of the κ-values

in terms of class memberships and calculates the index value. There are four different possibilities to define pairwise class memberships:

a - an object pair belongs to one class or cluster in the first solution and also belongs to one class or cluster in the second solution

b - an object pair belongs to one class or cluster in the first solution, but to different classes or clusters in the second solution

c - an object pair belongs to different classes or clusters in the first solution, but to one class or cluster in the second solution

d - an object pair belongs to different classes or clusters in the first solution, and also to different classes or clusters in the second solution

Based on the values for a to d the Rand index can be calculated by

Rand Index

$$RI = \frac{a+d}{a+b+c+d}$$

In [6] Campello derived a Fuzzy Rand Index using *t-norms* and *t-conorms* to obtain valid values for a to d based on fuzzy assignments. Detailed descriptions can be found in [6].

The resulting index is value a in $[0, 1]$, where 1 implies complete agreement.

Remark 1 Using different *t-norms* like Minimum, Sum or Łukasiewicz leads to different values for the Fuzzy Rand Index.

Remark 2 The Fuzzy Rand Index can only be used to compare a crisp with an fuzzy solution. It is not suitable for comparing two fuzzy solutions with each other, since in the case of perfect agreement, the Fuzzy Rand Index does not result in 1, which is caused by the use of the *t-norms*.

There are a couple of related indexes like

Adjusted Rand Index

$$ARI = \frac{a - \frac{(a+c)(a+b)}{d}}{\frac{(a+c)+(a+b)}{2} - \frac{(a+c)(a+b)}{d}}$$

Jaccard coefficient

$$J = \frac{a}{a+b+c}$$

Fowlkes-Mallows Index

$$FM = \frac{a}{\sqrt{(a+b)(a+c)}}$$

Minkowski Measure

$$M = \sqrt{\frac{b+c}{b+a}}$$

Γ **Statistics**

$$\Gamma = \frac{Ma - (a+b)(a+c)}{\sqrt{(a+b)(a+c)(M-(a+b))(M-(a+c))}}$$

$$\texttt{with} \quad M = N(N-1)/2$$

which are also suitable measures and are also using a to d. Each of them shows some special characteristics. For further details I again refer to the article by Campello [6] and also to the literature mentioned there.

2.3 Fuzzy variant of the Rand Index

In this measure, proposed by Hüllermeier in [3], the pairwise comparing and counting of class memberships as employed for the Rand Index is replaced by comparing the pairwise distances between two data points for two different cluster or classification solutions. Normalizing the sum over the differences between all the pairwise distances respectively gives the proposed measure

$$d(C_1, C_2) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} |E^{C_1}(\mathbf{x}_j, \mathbf{x}_j) - E^{C_2}(\mathbf{x}_j, \mathbf{x}_j)|}{n(n-1)/2}$$

where

$$E^{C_1}(\mathbf{x}_j, \mathbf{x}_j) = ||\mu^{C_1}(\mathbf{x}_i) - \mu^{C_1}(\mathbf{x}_i)||$$

with $\mu^{C_1}(\mathbf{x})$ as the fuzzy assignments of data point x to the clusters of the solution C_1. The same yields for E^{C_2} respectively.

The value of this measure again is in $[0, 1]$, where 1 implies complete agreement.

Remark Although called Fuzzy Rand Index by the author of [3], this claim holds only for the very special case that both cluster or classification solutions are crisp. Which again is equivalent with the normal Rand Index.

References

[1] T. Geweniger, D. Zühlke, B. Hammer, and T. Villmann. Median variant of fuzzy c-means. In M. Verleysen, editor, *Proc. Of European Symposium on Artificial Neural Networks (ESANN'2009)*, pages 523–528, Evere, Belgium, 2010. d-side publications.

[2] Tina Geweniger, Petra Schneider, Frank-Michael Schleif, Michael Biehl, and Thomas Villmann. Extending rslvq to handle data points with fuzzy class assignments. Machine Learning Report 02/2009, University of Leipzig, 2010.

[3] E. Hüllermeier and M. Rifqi. A fuzzy variant of the rand index for comparing clustering structures. *IFSA/EUSFLAT*, 2009.

[4] D. Zühlke, T. Geweniger, U. Heimann, and T. Villmann. Fuzzy Fleiss-Kappa for comparison of fuzzy classifiers. In M. Verleysen, editor, *Proc. Of European Symposium on Artificial Neural Networks (ESANN'2009)*, pages 269–274, Evere, Belgium, 2009. d-side publications.

[5] W. Dou, Y. Ren, Q. Wu, S. Ruan, Y. Chen, D. Bloyet, and J.-M. Constans. Fuzzy kappa for the agreement of fuzzy classifications. *NEUROCOMPUTING*, 70:726–734, 2007.

[6] R.J.G.B.Campello. A fuzzy extension of the rand index and other related indexes for clustering and classification assessment. *Pattern recognition Letters 28*, 2007.

[7] T. Villmann, B. Hammer, F.-M. Schleif, T. Geweniger, and W. Herrmann. Fuzzy classification by fuzzy labeled neural gas. *Neural Networks*, 19:772–779, 2006.

[8] J.L. Fleiss. *Statistical Methods for Rates and Proportions*. Wiley, New York, 2nd edition, 1981.

[9] William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[10] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960.

Application-adaptive dissimilarity measures for hyperspectral images

***Felix Bollenbeck*[2] *and Udo Seiffert*[1,2]**

Acknowledgements: The authors gratefully acknowledge scientific and technical support from Hans-Peter Mock and Andrea Matros, Leibniz-Institute of Plant Genetics and Crop Plant Research Gatersleben, Germany; Joachim Schiemann, Julius-Kühn Institute Quedlinburg, Germany; as well as technical support from Lars O. Lierstuen, Norsk Elektro Optikk A/S, Lørenskog, Norway.

1 Introduction

Hyperspectral imaging is a powerful method to transform the abundance of different molecules of a sample into a set of images. Each of these images contains a spatial distribution of the reflected light of the acquired scenery at a particular and typically narrow wavelength band. The reflected light intensity depends on the molecule structure of the sample near to its surface. All images together form a reflectance spectrum. Each pixel of the acquired image set can be considered as a vector that contains the so-called spectral fingerprint at this very local position. Fig. 5 shows an example of a spectral fingerprint.

This technique and its utilization to characterize and quantify biochemical compounds in plants has been known for decades [14, 13]. Due to a strong demand by plant breeders and farmers along with a recently higher availability of suitable hyperspectral cameras, applications in plant phenotyping and precision agriculture can increasingly be noticed [11, 4, 5, 1].

Besides specific and always required preprocessing and possibly some visualization of the acquired image sets (see Fig. 6), the analysis of the obtained hyperspectral signatures is a complex and delicate task. This generally ranges from

[1] E-mail: `Udo.Seiffert@iff.fraunhofer.de`

[2] Fraunhofer IFF Magdeburg, Biosystems Engineering, Magdeburg, Germany

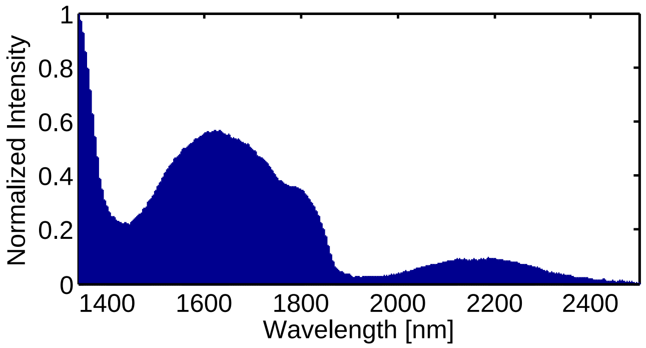


Figure 5: Sample spectral fingerprint (reflectance spectrum) at one pixel of an acquired scenery within the wavelength range of about 1,340 nm. . . 2,500 nm. Besides the molecule specific details of the sample the general physical effect of decreasing energy along with decreasing frequency (increasing wavelength) can be seen as well.

- unsupervised *clustering* to find similar spectral fingerprints to
- *classification* that links the data to some a-priori known categories, such as genotype or nutrition conditions, up to
- *regression* where typically the obtained spectra are mapped to externally obtained biochemical reference data.

Due to its complexity, both in terms of the number of variables (dimensions = spectral bands) and the number of sample vectors (pixels), and the fact that the analyses are data driven (analytical context is typically unknown), artificial neural networks and machine learning paradigms [12, 7] can beneficially be applied to tackle it [9, 3, 1].

2 Neural networks based spectral data analysis

Whereas regression (refer to list in previous chapter) requires a mapping of two different functions and one of them (output function) needs to be obtained by external wet lab analysis, clustering and classification are basically based on calculating similarities between per se available input samples. This leads to the discussion of suitable similarity or dissimilarity measures [10, 8]. In particular, prototype-based neural networks offer an elegant way to implement different metrics.

This is the starting point to develop task or application specific dissimilarity measures. One possible way is to consider the spectral fingerprint as pattern of positive-valued finite measures – a statistical distribution. In this case divergences can be applied to characterize and distinguish spectral fingerprints [6, 2].

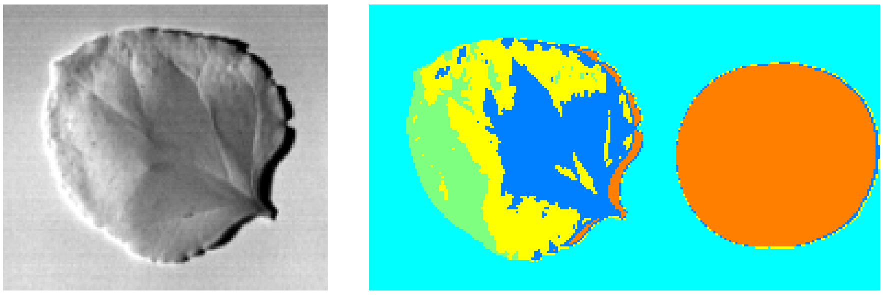


Figure 6: Demonstration of spectral properties: Sample leaf as gray-scale intensity image at an arbitrarily chosen wavelength and in false color coding after clustering the spectral fingerprints of all pixels. The circular structure on the right is the calibration pad (used for spectral calibration), that obviously forms an intrinsically separate cluster (orange color). Some pixels in the shade of the leaf (right edge) share the same cluster as the calibration pad, since this shaded background has similar neutral spectral properties. In contrast, the illuminated background has its own molecular structure that forms another separate cluster (cyan color).

Although this needs to be further developed in detail, parametric gamma-divergences seem to be a suitable option – not only due to their robustness in terms of outliers. Since these divergences are set in a machine learning environment now, the adaptation (training) of the gamma parameter itself in addition to the regular neural network training seems to be a promising approach.

3 Conclusions and outlook

The described work in progress is to develop specific application-adaptive dissimilarity measures that respect the particular properties of spectral signatures much better than standard options, such as Euclidean and correlation-based ones. The availability and implementation of these measures will further advance hyperspectral imaging to monitor biochemical compounds of plant leaves and beyond.

References

[1] Udo Seiffert, Felix Bollenbeck, Hans-Peter Mock, and Andrea Matros. Clustering of crop phenotypes by means of hyperspectral signatures using artificial neural networks. In *Proceedings of the 2nd IEEE Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing WHISPERS 2010*, pages 31–34. IEEE Press, 2010.

[2] Thomas Villmann, Sven Haase, Frank-Michael Schleif, and Barbara Hammer. Divergence based online learning in vector quantization. In *Artificial Intelligence and Soft Computing*, volume 6113 of *Lecture Notes in Computer Science*, pages 479–486. Springer-Verlag, Berlin, 2010.

[3] Thomas Villmann, Erzsébet Merényi, and Udo Seiffert. Machine learning approaches and pattern recognition for spectral data. In Michel Verleysen, editor, *Proceedings of the 16th European Symposium on Artificial Neural Networks ESANN 2008*, pages 433–444, Evere, Belgium, 2008. D-Side Publications.

[4] Gamal ElMasry, Ning Wang, Adel ElSayed, and Michael Ngadi. Hyperspectral imaging for nondestructive determination of some quality attributes for strawberry. *Journal of Food Engineering*, 81(1):98–107, 2007.

[5] Hiroshi Okamoto, Tetsuro Murata, Takashi Kataoka, and Shun-Ichi Hata. Plant classification for weed detection using hyperspectral imaging with wavelet analysis. *Weed Biology and Management*, 7(1):31–37, 2007.

[6] Dong-Chul Park, Chung Nguyen, and Yunsik Lee. Content-based classification of images using centroid neural network with divergence measure. In *AI 2006: Advances in Artificial Intelligence*, volume 4304 of *Lecture Notes in Computer Science*, pages 729–738. Springer-Verlag, Berlin, 2006.

[7] Udo Seiffert, Barbara Hammer, Samuel Kaski, and Thomas Villmann. Neural networks and machine learning in bioinformatics – theory and applications. In Michel Verleysen, editor, *Proceedings of the 14th European Symposium on Artificial Neural Networks ESANN 2006*, pages 521–532, Evere, Belgium, 2006. D-Side Publications.

[8] Yingzi Du, Chein-I Chang, Hsuan Ren, Chein-Chi Chang, James O. Jensen, and Francis M. D'Amico. New hyperspectral discrimination measure for spectral characterization. *Optical Engineering*, 43(8):1777–1786, 2004.

[9] Thomas Villmann, Erzsébet Merényi, and Barbara Hammer. Neural maps in remote sensing image analysis. *Neural Networks*, 16(3–4):389–403, 2003.

[10] Heesung Kwon, Sandor Z. Der, and Nasser M. Nasrabadi. Unsupervised segmentation algorithm based on an iterative spectral dissimilarity measure for hyperspectral imagery. volume 4310, pages 144–152. SPIE, 2000.

[11] Camille Lelong, Patrick Pinet, and Herve Poilve. Hyperspectral imaging and stress mapping in agriculture: A case study on wheat in Beauce (France). *Remote Sensing of Environment*, 66(2):179–191, 1998.

[12] Richard P. Lippmann. An introduction to computing with neural nets. *IEEE ASSP Magazine*, 4(87):4–23, 1987.

[13] Ray F. Severson, Richard F. Arrendale, Orestes T. Chortyk, Albert W. Johnson, D. Michael Jackson, G. Richard Gwynn, James F. Chaplin, and Michael G. Stephenson. Quantitation of the major cuticular components from green leaf of different tobacco types. *Journal of Agriculture and Food Chemistry*, 32:566–570, 1984.

[14] Joseph T. Woolley. Reflectance and transmittance of light by leaves. *Plant Physiology*, 47:656–662, 1971.