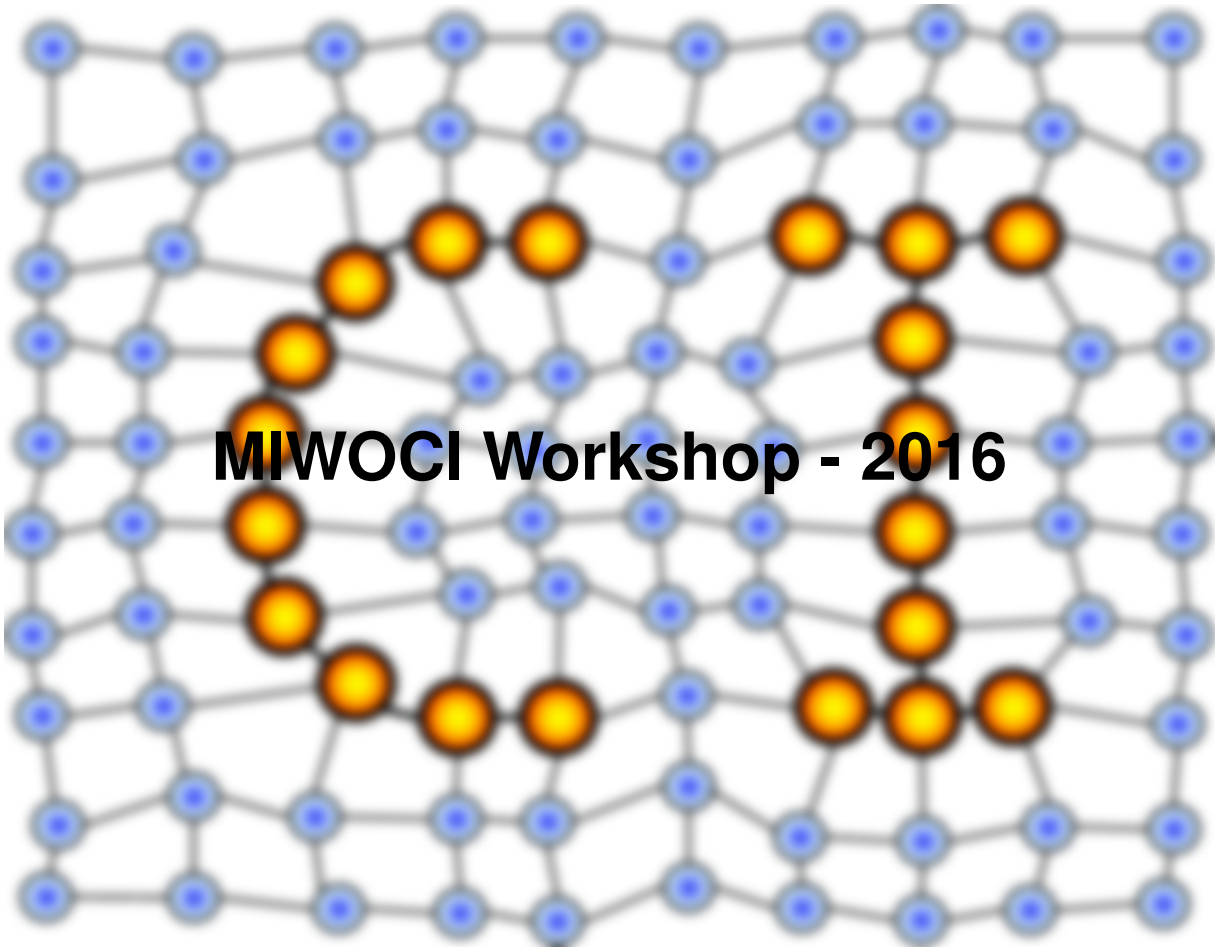
MACHINE LEARNING REPORTS



MIWOCI Workshop - 2016

Report 03/2016
Submitted: 10.07.2016
Published: 15.07.2016

Frank-Michael Schleif[1], Thomas Villmann[2] (Eds.)
(1) University of Bielefeld, Dept. of Technology CITEC - AG Computational Intelligence, Universitätsstrasse 21-23, 33615 Bielefeld
(2) University of Applied Sciences Mittweida, Technikumplatz 17, 09648 Mittweida, Germany

Figure 1: MiWoCi 2016

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:	fschleif@techfak.uni-bielefeld.de
URL:	http://techfak.uni-bielefeld.de/∼ fschleif/mlr/mlr.html
ISSN:	1865-3960

Abstracts of the 8^{th} Mittweida Workshop on Computational Intelligence - MiWoCI 2016 -

Frank-Michael Schleif and Thomas Villmann

Machine Learning Report 03/2016

Preface

The 8^{th} international *Mittweida Workshop on Computational Intelligence* (MiWoCI) gathering together twenty scientists from different universities including Bielefeld, Groningen, University of Applied Sciences Mittweida, University of Applied Sciences Dresden, and research facilities including Honda Research in Offenbach, Robert Bosch GmbH and IFF Fraunhofer in Magdeburg. The workshop took place in Mittweida, Germany, from 6.7. - 8.7.2016 and continued the tradition of scientific presentations, vivid discussions, and exchange of novel ideas at the cutting edge of research connected to diverse topics in computer science, automotive industry, and machine learning.

This report is a collection of abstracts and short contributions about the given presentations and discussions, which cover theoretical aspects, applications, as well as strategic developments in the fields.

Contents

Batch Neural Gas for Interval Data

Tina Geweniger

Computational Intelligence Group,
Univ. Applied Sciences Mittweida, DE
tgewenig@hs-mittweida.de

Abstract

For many real measurements taken by some technical device an interval describing the accuracy of the measured values is provided. Yet almost all clustering and classification methods just discard this additional information and work with the measured values only. We are going to investigate how well known machine learning methods especially those from the field of prototype based clustering can be modified to work with this kind of data. Some work has been done before by de Carvalho for the Fuzzy c-Means. In the following a variant of the Batch Neural Gas is provided. This modified version works with interval data and results in a partitioning described by interval data prototypes. Additionally to the derivation of the new method some insights into interval arithmetic are provided.

Feature Relevance Bounds for Linear Classification

Christina Göpfert Lukas Pfannschmidt
Barbara Hammer

Linear classifiers are widely used in medical and biological applications. The magnitudes of the weights in the normal vector are often used as an indicator for the relevance of individual features for the classification task. This principle can be misleading when features are high-dimensional or correlated. For example, correlated or even identical features share their weights, giving the impression that they are all equally unimportant, even though each single feature may be critical to the classification. We propose a formalization of this problem as maximal and minimal feature relevance bounds for linear classification and show how to efficiently calculate the bounds using linear programs. We illustrate the results using several toy datasets.

Is Deep Learning really worth it? – Comparing an NMF-GLVQ-based approach with Deep Learning

Sven Hellbach, Frank Bahrmann, Hans-Joachim Böhme

Semi-automatic semantic labelling of occupancy grid maps has numerous applications for assistance robotics. The Artificial Intelligence Lab at HTW Dresden has already proposed a method for representing local as well as global environment captured by 2D range scans using non-negative matrix factorisation (NMF). Unlike other approaches, no predefined features or geometric primitives, but, in contrast, extracted environment specific basis primitives from occupancy grid maps were used [4, 3].

The NMF also computes a description about where on the map these features need to be applied. This description is used after certain pre-processing steps as an input for generalised learning vector quantisation (GLVQ) to achieve the classification or labelling of the grid cells. For the supervised training of the GLVQ the assigned label is propagated to all grid cells of a semantic unit using a simple, yet effective segmentation algorithm. [2, 1]

For the implementation a sparse, transformation invariant version of NMF is used [5], which can be written as a convolution. As it has also been argumented in [6] similar ideas can be applied for deep learning, which makes both methods directly comparable.

The commencing research work aims at gathering advantages and disadvantages of both methods. With the application in mind, aspects like classification accuracy, number of necessary training data, and run time need to be evaluated. One of the major drawbacks with NMF for example, is the necessity to predefine the number of basis primitives to avoid redundant coding. How severe is this effect in deep learning? How well can both methods be used with the practical application? How interpretable and accessible are intermediate results?

References

[1] S. Hellbach, M. Himstedt, F. Bahrmann, M. Riedel, T. Villmann, and H.-J. Böhme. Find rooms for improvement: Towards semi-automatic labeling of occupancy grid maps. In *21th International Conference on Neural Information Processing*, 2014. Accepted for publication.

[2] S. Hellbach, M. Himstedt, F. Bahrmann, M. Riedel, T. Villmann, and H.-J. Böhme. Some Room for GLVQ: Semantic Labeling of Occupancy Grid Maps. In *Advances in Self-Organizing Maps and Learning Vector Quantization*,

volume 295 of *Advances in Intelligent Systems and Computing*, pages 133–143. Springer International Publishing, 2014.

[3] S. Hellbach, M. Himstedt, and H.-J. Böhme. Towards Non-negative Matrix Factorization based Localization. In *Proceedings of the European Conference on Mobile Robots (ECMR)*, 2013.

[4] M. Himstedt, S. Hellbach, and H.-J. Böhme. Feature extraction from Occupancy Grid Maps using Non-negative Matrix Factorization. In *Proceedings of the Workshop - New Challenges in Neural Computation 2012 (NC2 2012)*, pages 97–100, Graz (AT), 2012. Machine Learning Reports.

[5] C. Vollmer, S. Hellbach, J. Eggert, and H.-M. Gross. Sparse Coding of Human Motion Trajectories with Non-negative Matrix Factorization. *Neurocomputing, Elsevier*, 124:22–32, January 2014.

[6] B. Xin, Y. Wang, W. Gao, and D. P. Wipf. Maximal Sparsity with Deep Networks? *CoRR*, abs/1605.01636, 2016.

None Negative Sparse Coding for Analysis of Motion Data

Babak Hosseini Barbara Hammer *†

One of the current challenges in the area of machine learning and data analysis is to construct automate semantic search algorithms in motion data bases, specially where the data is not manually labeled. We hypothesize that semantics is mirrored by recurring signals, which are present in semantically similar motion data. In that scope, we investigate in how far natural priors such as sparsity allow an automatic extraction of semantically meaningful entities based on the given data alone.

To that aim, We propose an approach for decomposition of motion data into a sparse linear combination of base functions which enable efficient data processing. We combine two prominent frameworks: dynamic time warping (DTW), which offers particularly successful pairwise motion data comparison, and sparse coding (SC), which enables an automatic decomposition of vectorial data into a sparse linear combination of base vectors. We furthermore enhance SC via efficient kernelization which extends its application domain to general similarity data such as offered by DTW. In addition, we restrict the framework to provide non-negative linear representations of signals and base vectors in order to guarantee a meaningful dictionary as the outcome. We also implement the proposed method in a classification framework and evaluate its performance on various motion capture benchmark data sets to illustrate its effectiveness in semantic analysis of motion data.

*This research was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

†Babak Hosseini (bhosseini@techfak.uni-bielefeld.de) and Barbara Hammer are with the Theoretical Computer Science group of CITEC Bielefeld, Germany.

New insights into the Generalized Learning Vector Quantization cost function – *Non-linear shifting of the decision boundery* –

M. Kaden[2], D. Nebel[2], M. Biehl[1], and T. Villmann[2]

[1] Johann Bernoulli Institute for Mathematics and Computer Science
Intelligent Systems Group, Groningen, The Netherlands
[2] Computational Intelligence Group, Univ. Applied Sciences Mittweida, DE

Abstract

One classifier brings out one point in the false-positive rate vs true-positive rate diagram. To generate a ROC-curve either a lot of classifier are needed or a parameter can be introduced to change the decision border.

The Generalized Learning Vector Quantization (GLVQ) cost function for a given data set $\mathbf{v} \in V \subseteq \mathbb{R}^n$ with a the class labels $c(\mathbf{v}) \in \mathcal{C} \subset \mathbb{N}$ and a to adapted prototype set $\mathbf{w} \in W \subseteq \mathbb{R}^n$ with the class information $y(\mathbf{w}) \in \mathcal{C} \subset \mathbb{N}$ is defined as

$$E(V, W) = \sum_{\mathbf{v} \in V} f\left(\mu_W(\mathbf{v})\right)$$

whereby f is a monotone increasing function and

$$\mu_W(\mathbf{v}) = \frac{d(\mathbf{v}, \mathbf{w}^+) - d(\mathbf{v}, \mathbf{w}^-)}{d(\mathbf{v}, \mathbf{w}^+) + d(\mathbf{v}, \mathbf{w}^-)}$$

is called classifier function. The term $d(\mathbf{v}, \mathbf{w}^+)$ describes the dissimilarity between a data point $\mathbf{v}$ and the closest prototype $\mathbf{w}^+$ belonging to the same class. On the other hand, $d(\mathbf{v}, \mathbf{w}^-)$ is the dissimilarity of

a data point $\mathbf{v}$ and the next prototype belonging to any other class, i. e. $c(\mathbf{v}) \neq y(\mathbf{w})$.

The classifier function $\mu_W(\mathbf{v}) \in [-1, 1]$ is negative, iff the data point is correct classified, i. e. the nearest prototype, also called winning prototype $\mathbf{w}_{s(\mathbf{v})}$, has the same label like the data point. Thus, data points with $\mu_W(\mathbf{v}) = 0$ lying on the decision border. If we want to change the decision border by parameter we can easily define

$$\mu_W(\mathbf{v}) - \Theta = 0$$

with $\Theta \in [-1, 1]$.

In the presentation we show that the introduction of this parameter leads to a non-linear changing of the decision border. To obtain a linear change, only $d(\mathbf{v}, \mathbf{w}^+) - d(\mathbf{v}, \mathbf{w}^-) - \Theta$ can be analyzed. Yet, this term is not cost function adequate. We discuss this aspects and demonstrate the results on a two-dimensional data set.

Further, we argued about simplification of the original cost function and the behavoir concerning unique solution and boundedness. This impact is shown on a simple 1D data set example.

Choosing the Best Algorithm for an Incremental On-line Learning Task

Viktor Losing Barbara Hammer Heiko Wersing *

Recently, incremental and on-line learning gained more attention especially in the context of big data and learning from data streams, conflicting with the traditional assumption of complete data availability. Even though a variety of different methods are available, it often remains unclear which of them is suitable for a specific task and how they perform in comparison to each other.

We analyze the key properties of seven incremental methods representing different algorithm classes. Our extensive evaluation on data sets with different characteristics gives an overview of the performance with respect to accuracy as well as model complexity, facilitating the choice of the best method for a given application.

*Viktor Losing (vlosing@techfak.uni-bielefeld.de) does his PhD within a cooperation between CoreLab, Bielefeld and Honda Reasearch Institute EU, Offenbach. Barbara Hammer holds a professorship for Theoretical Computer Science at the CITEC in Bielefeld. Heiko Wersing is chief scientist at the Honda Reasearch Institute EU in Offenbach.

Metric Learning and its Caused Pitfall in Certainty Estimation

Lydia Fischer

Abstract

Certainty estimation is an upcoming topic in machine learning, usable in various topics: e. g. in rejection strategies, incremental online learning approaches, or in classifier selection mechanisms of ensembles. One category of certainty measures is based on distances, e. g. distance to the closest decision border. I will point out the difficulty of comparing distance-based certainty estimations of different classifiers with metric adaptation which is especially important for classifier selection mechanisms of ensembles. At an exemplary architecture with two classifiers, I will show the caused pitfall by metric learning, and how to deal with it. The full information can be obtained from:
Fischer et al.: Online Metric Learning for an Adaptation to Confidence Drift. In IJCNN, accepted, 2016.

The sugar dataset - A multimodal hyperspectral dataset for classification and research

Friedrich Melchert[1,2], Andrea Matros[3], Michael Biehl[1], and Udo Seiffert[2]

[1] *University of Groningen, Johann Bernoulli Institute for Mathematics and Computer Science, Groningen, The Netherlands*

[2] *Fraunhofer Institute for Factory Operation and Automation IFF, Magdeburg, Germany*

[3] *Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany*

We present a multi modal hyperspectral dataset (available online at [5]) that cannot only be used to evaluate and compare classification performance, but also enables research on new topics.

In the development of algorithms for hyperspectral data classification several benchmark datasets became common, e.g. the Tecator dataset [7] and the Wine dataset [6], to name two examples. These datasets are mainly used as benchmark problems for different algorithms and classification systems as in [1, 4], although they just compile a set of labeled spectra. In opposite to these well established datasets the *Sugar* dataset offers multiple sets of spectra for each of the available classes. Using a variety of different sensors and hyperspectral cameras the spectral information within the dataset is given over different wavelength ranges. An overview over the sensors and their corresponding wavelength ranges is given in Table 1.

As a training set we selected nine sugar and sugar related compounds with common optical appearance, which are not to be distinguishable by conventional optical imaging. We included were three monomeric sugars (D-galactose, D-glucose, and D-fructose), two sugar alcohols (D-sorbitol and D-mannitol), as well as four sugar esters (S170, S770, S1570, and P1570). Monomeric sugars containing six carbon atoms are also referred to as hexoses, having the chemical formula $C_6H_{12}O_6$. Hexoses occur in many stereoisomers and are classified into aldohexoses, having an aldehyde at position 1 (e.g. D-galactose and D-glucose), and ketohexoses having a ketone at position 2 (e.g. D-fructose). Sugar alcohols are typically derived from sugars by a reduction reaction, changing the aldehyde group to a hydroxyl group. We selected two hexose-derived sugar alcohols with the molecular formula $C_6H_{14}O_6$. Sugar esters, also called sucrose fatty acid

Table 1: Key properties of the different sensors used to record the sugar dataset.

Sensor name	Manufacturer	wavelength range [nm]	sampling points
EOS 70D	Canon[1]	RGB	3
Fieldspec	ASD[2]	350 - 2500	2151
VNIR-1600	NEO[3]	400 - 1000	160
VNIR-1800	NEO[3]	400 - 1000	186
Nuance EX	Nuance	520 - 880	37
SWIR-320m-e	NEO[3]	1000 - 2500	256
SWIR-384	NEO[3]	1000 - 2500	288

[1] `http://www.canon.com` [2] `http://www.asdi.com` [3] `http://www.neo.no`

esters, are nonionic surfactants consisting of sucrose as hydrophilic group and fatty acid as lipophilic group. Sugar esters can vary in the nature of the attached fatty acid, such as palmitate (P1570) or stearate (S170, S770, S1570) as well as in the number of attached fatty acids (called mono-, di-, tri-, tetraester). In our case, compounds with variation of both parameters have been chosen: S-170 (sucrose stearate, ratio 1% monoester, 99% di-, tri-, and polyester), S-770 (sucrose stearate, ratio 40% monoester, 60% di-, tri-, and polyester), S-1570 (sucrose stearate, ratio 70% monoester, 30% di-, tri-, and polyester), and P-1570 (sucrose palmitate, ratio 70% monoester, 30% di-, tri-, and polyester). According to the high variation in stereochemistry and composition we expected a high degree of diversity in our data set. All compounds appear as white powder, whereas D-fructose looked more crystalline. Spectral profiles were acquired using a variety of different sensors and hyperspectral cameras (Table 1). Given the nine different compounds it is possible to define five classification problems. The mapping of the compounds to the different classification problems is given in Table 2.

Besides the use as a benchmark dataset, the unique structure of the *Sugar* dataset offers the opportunity to encourage research on further topics. Three of the possible research questions are briefly discussed in the following

Dimensionality reduction The data within this dataset is compiled from high dimensional feature vectors. Various machine learning algorithms suffer from the presence of high dimensional inputs, which imply a high number of adaptive parameters, leading to convergence problems, overfitting effects and suboptimal results [2].

Taking into account the functional characteristics of spectral data, the high number of input dimensions is not justifiable. For functional data, such as the hyperspectral data in this dataset, a high correlation of neighbored features is expected. Thus the dataset can serve as a basis for the development and evaluation of dimension reduction algorithms. The varying number of input dimensions ($37 - 2151$, cf. Table 1) within the dataset facilitates a solid benchmarking of the performance and scaling of novel approaches.

Table 2: Definition of the different classification problems. The numbers in the table represent the class index of the substance with regard to the classification problem. Empty cells indicate, that this substance is not used within the concrete classification problem (row). Borders are used to illustrate the pooling of multiple substances to a single class.

	number of classes	Sugar ester S170	Sugar ester S770	Sugar ester S1570	Sugar ester P1570	D-Mannitol	D-Sorbitol	D-Glucose	D-Galactose	D-Fructose
problem 0	9	1	2	3	4	5	6	7	8	9
problem 1	3	1				2		3		
problem 2	4	1	2	3	4					
problem 3	2					1	2			
problem 4	3							1	2	3

Sensor invariant classification models In the design of classification models the structural properties of the input data plays a major role. In most cases a change of the input data properties is simply not possible. Furthermore trained classification models often implicitly incorporate sensor specific properties during optimization. So the change of measurement equipment can lead to a loss in classification performance. Since the training of classification models is usually time consuming the generation of a new classifier after a hardware change is costly.

For the composition of this dataset the spectral information of certain substances were recorded using multiple different sensors. Given the overlapping wavelength ranges (again cf. Table 1) the dataset contains data, that represents the same spectral information recorded with different sensors. This data can be used for the development and validation of algorithms, which are capable of handling different input formats such as variable sized feature vectors and slight shifts in the positions of the spectral sampling points, as well as sensor specific patterns and fragments within the data.

High dimensional data exploration For the generation of industrial classification systems based on spectral information the selection of a suitable sensor system is one of the key issues. In most of the cases the wavelengths which are relevant for the classification task are not known in advance, so the selection of a sensor system usually follows an educated guess or is guided by financial issues.

Having wide and limited wavelength range sensor data within the presented dataset, the question of proper sensors selection may be tackled in a more sys-

tematic way. Using the provided data it is possible to develop relevance learning schemes, which quantify the importance of wavelengths. Relevances emerging from the classification of wide spectral bandwidth data (which may have a low number of samples) can be used for a proper sensor selection, on which classification models may be tuned afterwards.

Apart from the sensor selection the dataset provides also opportunities for the challenging visualization of high dimensional data, which is a key issue for data exploration [3].

These questions may serve as a starting point for further research. Nevertheless this list is not complete (neither it is meant to be). The presented *Sugar* dataset is unique in terms of its structure and extent, and hopefully serves as a basis for future improvements in the classification of hyperspectral data, as well as the outlined research topics.

Acknowledgments

F. Melchert thanks for support through an Ubbo-Emmius Sandwich Scholarship by the Faculty of Mathematics and Natural Sciences, University of Groningen.

References

[1] Marika Kästner, Barbara Hammer, Michael Biehl, and Thomas Villmann. Functional relevance learning in generalized learning vector quantization. *Neurocomputing*, 90:85–95, 2012.

[2] Yann LeCun, J. S. Denker, S. Solla, R. E. Howard, and L. D. Jackel. Optimal brain damage. In David Touretzky, editor, *Advances in Neural Information Processing Systems (NIPS 1989)*, volume 2, Denver, CO, 1990. Morgan Kaufman.

[3] Allen R Martin and Matthew O Ward. High dimensional brushing for interactive exploration of multivariate data. In *Proceedings of the 6th Conference on Visualization'95*, page 271. IEEE Computer Society, 1995.

[4] Belen Martin-Barragan, Rosa Lillo, and Juan Romo. Interpretable support vector machines for functional data. *European Journal of Operational Research*, 232(1):146–155, 2014.

[5] Friedrich Melchert, Andrea Matros, Micheal Biehl, and Udo Seiffert. The sugar dataset. University of Groningen, `http://dx.doi.org/10.4121/uuid:fad486b6-0dfb-4d23-8025-b24407a08698`, 2016. Dataset.

[6] Marc Meurens. Wine mean infrared spectra dataset. University of Louvain, `http://mlg.info.ucl.ac.be/index.php?page=DataBases`. Dataset.

[7] Hans Henrik Thodberg. Tecator meat sample dataset. StatLib Datasets Archive, `http://lib.stat.cmu.edu/datasets/tecator`, 1995. Dataset.

Transferring machine learning models within a soft sensor system to achieve constant task performance under changing sensor hardware

Patrick Menz[1,2], Andreas Backhaus[1], and Udo Seiffert[1,2]

[1] *Biosystems Engineering, Fraunhofer Institute for Factory Operation and Automation IFF, Sandtorstr. 22, 39106 Magdeburg, Germany.*
[2] *Otto-von-Guericke Universität Magdeburg, Fakultät für Elektro- und Informationstechnik, Germany.*

1 Introduction

The combination of a hardware sensor system which generates high-dimensional data points, for example a hyperspectral camera, and machine learned method for data classification or prediction based on learning reference data is termed "soft sensor approach" and is widely used in application fields like precision farming or plant breeding. Soft sensors are calibrated from large datasets that are often collected over multiple years and field seasons. Calibration and training is also a numerically demanding process requiring computational resources. Therefore Obtaining datasets and providing processing power are associated with considerable costs. In such a system, the sensor component is subject to aging processes or needs to be replaced with a new sensor hardware or hardware from a different manufacturer. Here arises the motivation that machine learning models already carefully trained and validated are reused with changing sensor hardware.

Every sensor hardware is unique, so every sensor has different sampling rates. Because of this, it is nearly impossible for the model to accurately replicate the results of the original trained sensor. So the soft sensor cannot handle native data from different sensor hardware, because differences in the relevant properties are so subtle that they lie within their variation range of the hardware sensors. This work aims at creating a method to build a transfer function for mapping a changed sensor to the existing machine learning model (see Figure 1).

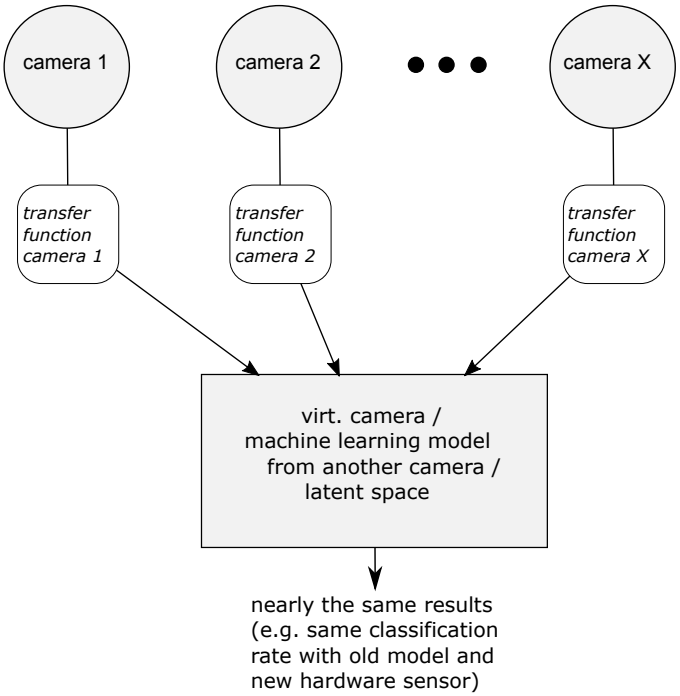


Figure 1: The main idea and aim of this work

2 Previous results

For model training, two different datasets were measured with several hyperspectral camera systems, where one consists of three different coffee varieties (Arabica, Robusta and Immature) and the other of nine different sugar varieties, partially with nearly identical chemical structure. A machine learning model was trained to classify the varieties within these datasets. As mentioned in the introduction, the aim is to find a transfer-function to replicate the results of the classifier accurately. A simple approach is the interpolation of the changed sensor hardware's output data to the model wavelength. Several interpolation methods were used to create enough representing data. Methods coming with MATLAB and are simple linear, next neighbour, nearest neighbour, previous neighbour, piecewise cubic, and spline interpolation. To create multi-variety, various training algorithms were used as can be seen in Figure 2. A drawback of simple interpolation is a significant loss of classification performance.

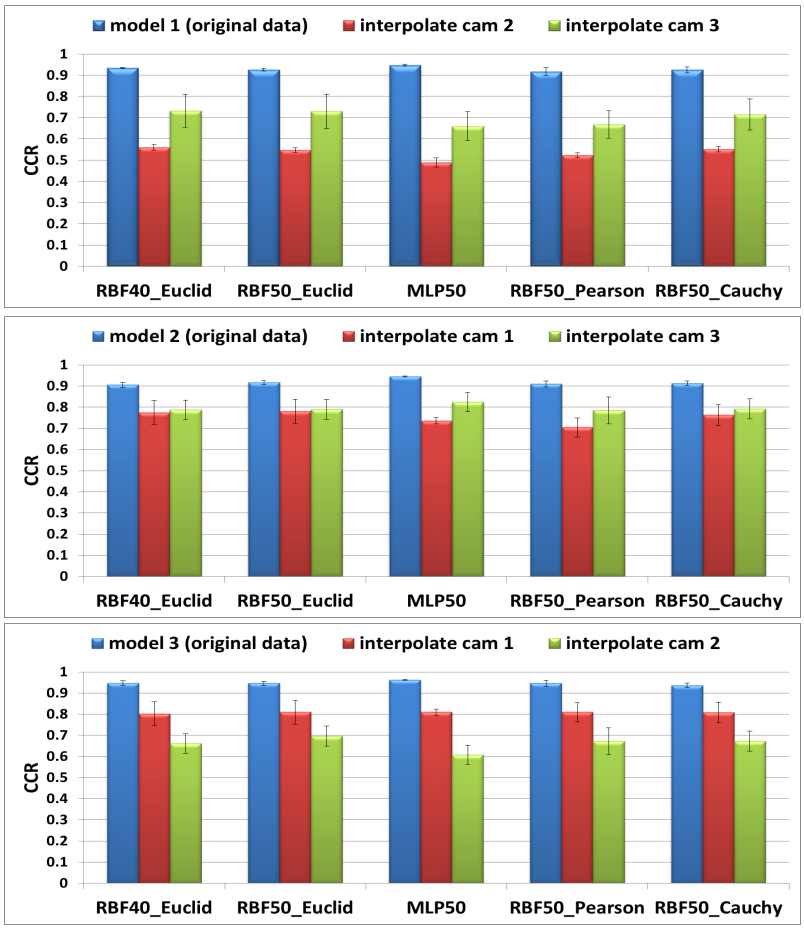


Figure 2: Coffee dataset correct classification rates of original data and transferred data over all interpolation methods with standard derivation. Different machine learning algorithms: RBF40_Euclid is a radial basis function network with 40 prototypes and a Euclidean metric, MLP50 is a multi-layer perceptron with 50 hidden layers, RBF50_Pearson same as before but with a Pearson Correlation metric, RBF50_Cauchy is with a Cauchy-Schwarz metric.

A more promising idea is to derive a more accurate model by using another calibration method. The standard method is a two-point calibration with white reference and dark current. In [1] it is being proposed to use more calibration standards (e.g. reflectance of 99%, 75%, 50%, 25% and 0%) to create a multi-calibration model for mapping the true reflectance in a greater detail. The machine learning algorithm chosen here exceed the classification performance in comparison with early approaches. However, when compared to simple interpolation, the model transformation did not show any improved results.

3 Conclusion and Outlook

The high sensitivity of machine learning models to changing sensor hardware is a more difficult issue. Interpolation, even with better calibration methods applied, does not show any benefits in transferring machine learning models. Future work concentrates on finding and correcting of nonlinearities, for example with gamma correction known from the image processing area. Furthermore, latent space modelling for spectral alignment (cf. [4]) is a promising method for this issue. Also an improvement of the machine learning model can be done by transfer learning (see [3]), in order to keep the knowledge of the already trained model. In addition to the existing calibration methods, a calibration model transfer based on alternating triliniear decomposition can be applied, explained in [2].

References

[1] J. Burger and P. Geladi. Hyperspectral nir image regression part 1: Calibration and correction. *J. Chemometrics*, (19):355–368, 2005.

[2] Yan Liu, Wensheng Cai, and Xueguang Shao. Standardization of near infrared spectra measured on multi-instrument. *Analytica Chimica Acta*, (836):18–23, 2014.

[3] L. Torrey and J. Shavlik. Transfer learning. *Handbook of Research on Machine Learning Applications*, 2009.

[4] M. Volpi, G. Camps-Valls, and D. Tuia. Spectral alignment of multi-temporal cross-sensor images with automated kernel canonical correlation analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, (Preprint), 2015.

(Dis-)Similarities Measures vs Inner Products – *Pitfalls and Properties* –

D. Nebel, M. Kaden, A. Bohnsack, and T. Villmann

Computational Intelligence Group,
University of Applied Sciences Mittweida, DE

Abstract

In machine learning the terms of (dis-)similarity measures (D/S) and inner product (IP) were mixed in many articles. Yet, the properties of D/S and IP are diverse. In general, it is almost impossible to interpret an IP as similarity and vise versa. We give a mathematical characterization and classification of D/S based on structural properties. Moreover, we introduce a rank measures to compare different D/S in prototype based learning. Further, the introduced rank measures can be used to show that the preparation of proximity data, e. g. to make them Euclidean, leads to substantial changes of information.

Supervised Linear Transfer Learning for GMLVQ

Benjamin Paassen Alexander Schulz Barbara Hammer *†

In practical applications of machine learning models, input data is subject to transformations, which result from sensor disturbances. For example, machine learning models used for controlling bionic arm prostheses typically receive input from myoelectric sensors, which are disturbed by shifts in electrode placement, sweat, posture changes and user fatigue. Under such conditions, the models predictive performance is likely to break down as the model does not fit the data anymore. We propose to alleviate such problems by applying transfer learning, that is, to project the disturbed input data into a space where the model fits again, such that the original performance is regained.

We model the disturbing transformation as a linear transformation, and learn its inverse directly using stochastic gradient descent on the G(M)LVQ cost function, thereby identifying a data transformation that optimizes the performance of a GMLVQ classifier. We demonstrate the effectiveness of the approach on a practical example, namely the classification of motion intend to control a bionic arm prosthesis after disturbance by electrode shift.

*Funding by the DFG under grant numbers HA2719/6-2 and HA2719/7-1 and the CITEC center of excellence (EXC 277) is gratefully acknowledged.

†Benjamin Paassen (bpaassen@techfak.uni-bielefeld.de), Alexander Schulz and Barbara Hammer are with the Theoretical Computer Science group of CITEC Bielefeld, Germany.

Secure Classification and Reject Options in LVQ

T. Villmann[1,2]*, A. Bohnsack[1,3], and M. Kaden[1,2]

[1]Computational Intelligence Group,
University of Applied Sciences Mittweida, Germany

[2]Inst. for Computational Intelligence and
Intelligent Data Analysis e.V. (CIID),
Mittweida, Germany

[3]Berufliches Schulzentrum Döbeln-Mittweida, Germany

Currently, the aspect of *secure* classification, i.e. the estimation of the evidence of a classification decision for an unknown data object, is in the focus of ongoing research. For support vector machines (SVM), the separation margin determines the certainty of a classification decision, which is maximized during the model learning [1]. For GLVQ [2], the hypothesis margin is optimized describing the robustness of the GLVQ regarding model shifts [3]. Yet, these quantities cannot be used to estimate the decision certainty for unknown data.

If classification is related to costs, classifier based on Bayes decisions regarding optimum costs come into play, which allow efficient reject options to increase the classification security [4]. Yet, the methods require the precise determination of the class distributions, which might be difficult [5]. GLVQ provides a robust model approach to estimate class distributions and, hence, it may serve as an approximated Bayes classifier with reject option in the working phase [6], whereas adaptive classification reject was considered in [7]. Cost based outlier detection are introduced [8].

In this contribution, we provide an alternative outlier detection strategy for GLVQ, which takes explicitly the hypothesis margin into account. According to this idea, a prototype rejects a data vector because of uncertainty, if the distance to the best matching prototype is greater than the hypothesis margin. We denote this strategy as an exploration horizon based reject option (EHBRO). Particularly, we show that the knowledge about this post-learning reject option can be integrated

* *corresponding author - email: thomas.villmann@hs-mittweida.de*

into GLVQ adaptation, During recall EHBRO can be used to indicate classification decisions with high uncertainty or to reject those outliers.

References

[1] B. Schlkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, 2002.

[2] A. Sato and K. Yamada. Generalized learning vector quantization. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference*, pages 423–9. MIT Press, Cambridge, MA, USA, 1996.

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

[4] C.K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions in Information Theory*, 16(1):41–46, 1970.

[5] L. Fischer and T. Villmann. A probabilistic classifier model with adaptive rejection option. *Machine Learning Reports*, 10(MLR-01-2016):1–16, 2016. ISSN:1865-3960, http://www.techfak.uni-bielefeld.de/ ~ fschleif/mlr/mlr_01_2016.pdf.

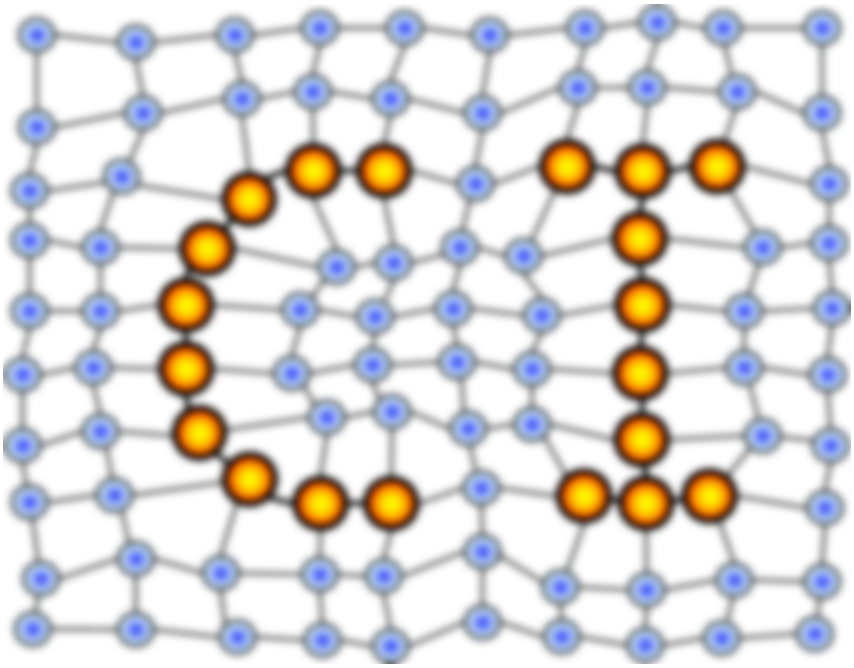
[6] L. Fischer, B. Hammer, and H. Wersing. Efficient rejection strategies for prototype-based classification. *Neurocomputing*, 169:334–342, 2015.

[7] T. Villmann, M. Kaden, A. Bohnsack, S. Saralajew, J.-M. Villmann, T. Drogies, and B. Hammer. Self-adjusting reject options in prototype based classification. In E. Merényi, M.J. Mendenhall, and P. O'Driscoll, editors, *Advances in Self-Organizing Maps and Learning Vector Quantization: Proceedings of 11th International Workshop WSOM 2016*, volume 428 of *Advances in Intelligent Systems and Computing*, pages 269–279, Berlin-Heidelberg, 2016. Springer.

[8] T. Villmann, M. Kaden, D. Nebel, and M. Biehl. Learning vector quantization with adaptive cost-based outlier-rejection. In G. Azzopardi and N. Petkov, editors, *Proceedings of 16th International Conference on Computer Analysis of Images and Pattern, CAIP 2015, Valetta - Malta*, volume Part II of *LNCS 9257*, pages 772 – 782, Berlin-Heidelberg, 2015. Springer.

MACHINE LEARNING REPORTS

Report 03/2016



Impressum

Machine Learning Reports ISSN: 1865-3960

▽ Publisher/Editors

Prof. Dr. rer. nat. Thomas Villmann
University of Applied Sciences Mittweida
Technikumplatz 17, 09648 Mittweida, Germany
• http://www.mni.hs-mittweida.de/

Dr. rer. nat. Frank-Michael Schleif
University of Bielefeld
Universitätsstrasse 21-23, 33615 Bielefeld, Germany
• http://www.cit-ec.de/tcs/about

▽ Acknowledgments

We would like to thank the reviewers for their time and patience.