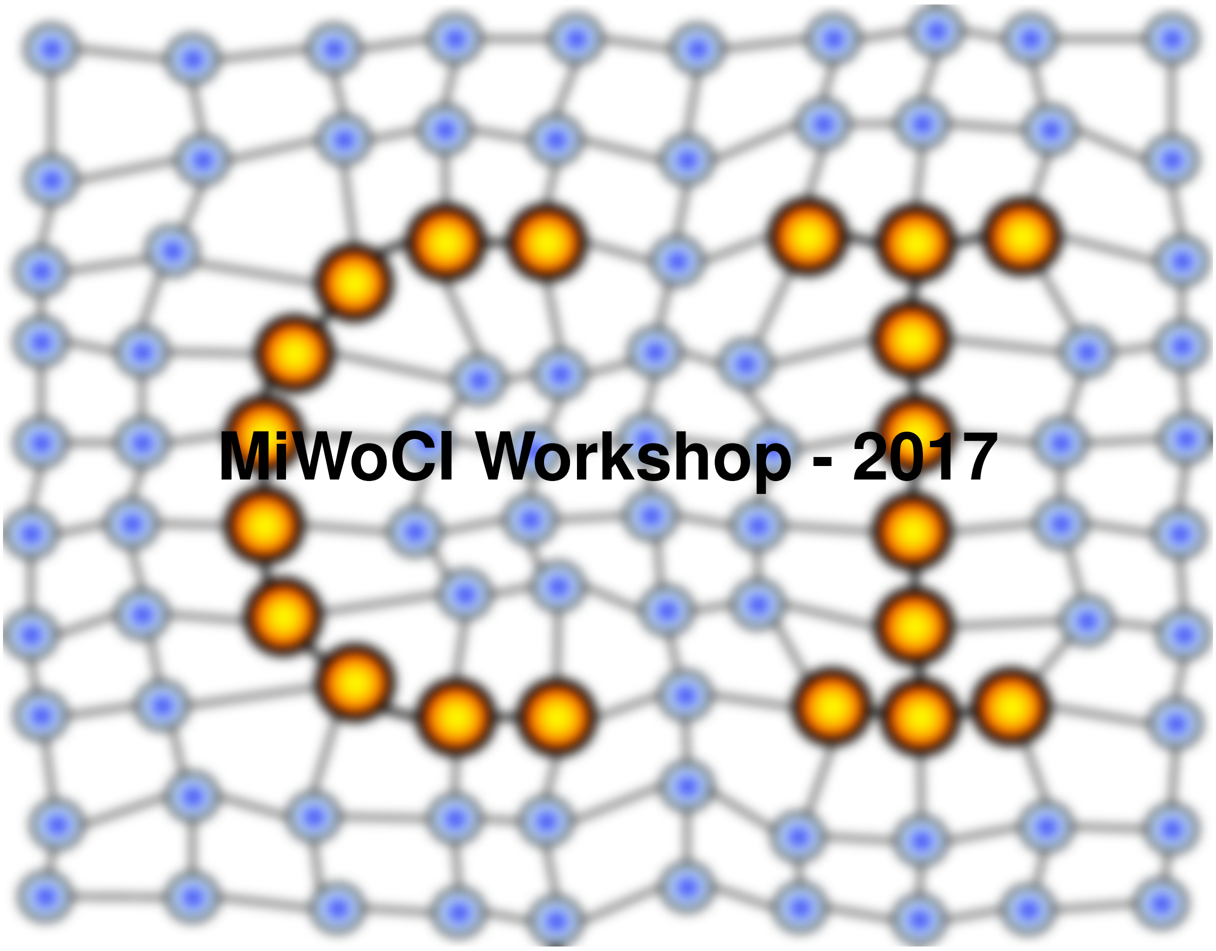
Machine Learning Reports



MiWoCI Workshop - 2017

Report 01/2017
Submitted: 23.07.2017
Published: 24.07.2017

Frank-Michael Schleif[1,2*,3*], Thomas Villmann[2] (Eds.)
(1) University of Applied Sciences Wuerzburg-Schweifurt, Sanderheinrichsleitenweg 20, 97074 Wuerzburg, Germany (2) University of Applied Sciences Mittweida, Technikumplatz 17, 09648 Mittweida, Germany (3) University of Birmingham, School of Computer Science, Edgbaston, B15 2TT Birmingham, UK

Machine Learning Reports
http://www.techfak.uni-bielefeld.de/∼fschleif/mlr/mlr.html

Contents

Impressum

Publisher:	University of Applied Sciences Mittweida Technikumplatz 17, 09648 Mittweida, Germany
Editor:	Prof. Dr. Thomas Villmann Prof. Dr. Frank-Michael Schleif
Techical-Editor:	Prof. Dr. rer. nat. habil. Frank-Michael Schleif
Contact:	f.schleif@cs.bham.ac.uk
URL:	http://techfak.uni-bielefeld.de/∼ fschleif/mlr/mlr.html
ISSN:	1865-3960

1 Ninth Mittweida Workshop on Computational Intelligence

From 24 Juli to 26 Juli 2017 we had the pleasure to organize and attend the ninth Mittweida Workshop on Computational Intelligence (MiWoCi 2017). Multiple scientists from the University of Bielefeld, HTW Dresden, the University of Groningen (NL), the University of Birmingham (UK), the University of Applied Sciences Mittweida,Hefei University (China), Honda Research Offenbach, Porsche AG met in Mittweida, Germany, to continue the tradition of the Mittweida Workshops on Computational Intelligence - *MiWoCi'2017*.

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 abstracts which accompany some of the discussions and presented work of the MiWoCi Workshop.

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

Mittweida, July, 2017
Frank-M. Schleif

[1] E-mail: frank-michael.schleif@fhws.de

[2] University of Appl. Sc. Wuerzburg-Schweinfurt, Wuerzburg, Germany

Learning Pharmakokinetic Models for Prednisone Absoption

Kerstin Bunte[1], David Smith[2], Michael Chappell[3], Zaki K. Hassan-Smith[2], Jon[2], Wiebke Arlt[2], and Peter Tiño[2]

[1]University of Groningen, Groningen, NL
[2]The University of Birmingham, Birmingham, UK
[3]University of Warwick, Coventry, Warwick, UK

Abstract

We propose a method for learning clusters of pharmacokinetic models demonstrated on a clinical data set investigating the 11β-HSD1 activity in healthy adults. Prednisone has an identical affinity for 11β-HSD1 as cortisone and the interconversion of oral prednisone to prednisolone has been used as a marker of the enzyme activity. The parameters of the multi-compartment ordinary differential equation model are studied via identifiability analysis and the observable measurements, which is used to interpret the learned clusters. We approximate the model using the pertubation method, which enables very efficient training of the proposed Guassian mixture clustering technique optimized by Estimation Maximization (EM). The training on the clinical data results in 4 clusters resembling the prednisone conversion rate in a period of 4 hours based on venous blood samples taken at 20-minute intervals. The learned clusters differ in prednisone absorption as well as prednisone/prednisolone conversion rate, which can be seen from the analysis of the learned parameter relationships. Consultation of further satellite data for each person not used for training reveals a correlation of cluster membership and total fat mass.

Computer aided diagnosis of inborn steroidogenic disorders

Sreejita Ghosh[1], Elizabeth Sarah Baranowski[2], Michael Biehl[1],
Wiebke Arlt[2], Peter Tino[3], Kerstin Bunte[1]

1- University of Groningen - JBI of Mathematics and Computer Science, NL

2- University of Birmingham - IMSR, UK

3- University of Birmingham - School of Computer Science, UK

Abstract: Due to improved biochemical sensor technology, there is increase in both amount of complex biomedical data, and the demand for automated interpretable interdisciplinary analysis techniques. However biomedical data have the challenges of 1) heterogenous measures, 2) missingness, and 3) imbalanced classes. The problem of imbalanced class becomes prominent especially for patients with rare diseases. For such datasets, even if all the patients are misclassififield as healthy the overall class accuracy might still be close to ninety percent. Thus optimizing overall class accuracy of the classification technique is not enough. It is the high detection rate of the minority classes which is particularly desirable. We have dataset of rare inborn steroidogenic disorsders which are caused by specific genetic mutation, and lead to defective production of any of the enzymes or a cofactor responsible for catalysing salt and glucose homeostasis, sex differentiation and sex specific development. Inborn steroidogenic disorders need to be diagnosed as early as possible, to avoid delay of lifesaving glucocorticoid therapy for adrenal insufficiency, and to facilitate gender allocation and surgical planning in patients with disordered sex development. Our dataset consist of urine GC/MS measurements from 829 healthy controls (305 under 1 year of age) and 118 genetically confirmed patients with steroidogenic disorders. Data samples are presented as 165 dimensional ratio vectors of 34 distinct steroid metabolite concentrations constructed using domain knowledge [3]. Bunte et al. [1] introduced an approach for the computer-aided diagnosis of the most prevalent condition, 21-hydroxylase deficiency (CYP21A2), and two other representative, 5α-reductase type 2 deficiency (SRD5A2) and P450 oxidorectase deficiency (PORD), and simultaneously handling missing and heterogenous measurements in the urine data. In Ghosh et al. [3] we investigated two main strategies for learning from imbalanced data: 1) penalizing misclassification of disease to healthy more severely than of misclassification within-diseases. 2) re-sampling the original dataset by either under-sampling the majority class and/or over-sampling the minority classes according to Chawla et al. [2]. We used two variants of Learning vector quantization(LVQ) which are capable of dealing with missingness, NaNLVQ and Angle-LVQ, as classifiers. In Ghosh et al. [3] we had just used the relevance vectors. As next steps we investigated 2 and 3 dimension global matrices in the Angle-LVQ classifier, followed by the corresponding local matrices, to see if we could gain further insights from these higher dimensions and more complex models. From the 2 and 3 dimension global matrices we obtained *Disease Maps* and *Disease Globes* which are the projection of the samples from 2D and 3D global matrices of AngleLVQ respectively. The *Disease Globes* were then flattened out into maps using Mollweide projection. Comparison between the relevance profiles obtained from local and global matrices gave us better idea about the disease specific blokages in the steroidogenic pathway (extraction of important decision boundares). Such an understanding will help us create a system for personalized medicine and individual treatment titration. In this workshop we would like to discuss the results from the above experiments and the issues we are trying to solve.

References

[1] K. Bunte, E. S. Baranowski, W. Arlt, and P. Tino. Relevance learning vector quantization in variable dimensional spaces. Workshop of the GI-Fachgruppe Neuronale Netze and the German Neural Networks Society in connection to GCPR 2016, pages 20–23, Hannover, Germany, August 2016. LNCS. URL `https://www.techfak.uni-bielefeld.de/ fschleif/mlr/mlr`$_0$4$_2$016*.pdf*.

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[3] S. Ghosh, E. Baranowski, R. van Veen, G. de Vries, M. Biehl, W. Arlt, P. Tino, and K. Bunte. Comparison of strategies to learn from imbalanced classes for computer aided diagnosis of inborn steroidogenic disorders. In *Proc. of the European Symposium on Artificial Neural Networks*, 2017.

Fast and precise identification of flow cytometry cell populations

Markus Lux

Identification of cell populations is a critical part of flow cytometry data analysis and lays the groundwork for both clinical diagnostics and research discovery. The current paradigm of manual analysis is time consuming and subjective. For automated gating, supervised tools provide the best performance, however they require fine parameterization to obtain the best results. In this ongoing work, we present a semi-supervised approach for the identification of cell populations. Using as few as one manually gated sample, it is able to predict gates on other samples with high accuracy and speed.

Feature Relevance

Christina Göpfert Barbara Hammer

Bielefeld University, CITEC - Center of Excellence, Germany

cgoepfert@techfak.uni-bielefeld.de

While the relevance of features or sets of features for a machine learning task is frequently analyzed, the results can be difficult to interpret. This is in part due to varying definitions of feature relevance and the feature selection problem, some of which can show counter-intuitive behavior. In my talk, I give an introduction into basic concepts for formalizing feature relevance and feature selection problems and their pitfalls. Furthermore, I introduce a general concept for formalizing the all-relevant problem in terms of generalization error bounds.

Using RSLVQ for prior distribution

Mohammad Mohammadi*

*University of Groningen, Groningen, NL

Abstract

Fitting a probabilistic model for a given data set means we are trying to find the best parameters for the model. In many application, it is necessary to know about the uncertainty of the resulting parameters. Bayesian approaches use posterior distributions to describe the uncertainty of parameters. The posterior distribution depends on two factor: prior distribution (what we expect about the parameters), and likelihood function (what data tells us).

The prior distributions encode our knowledge or guess about the parameters. However, in many situations we do not have any knowledge about it. In this scenario, data can give us a clue about parameters. One option is to use prototypes' distribution as a prior distribution. RSLVQ can provide an approximation for prototypes' distribution.

Soft-LVQ and Dependent Prototypes

M. Mohannazadeh and T. Villmann

Computational Intelligence Group,

University of Applied Sciences Mittweida, Germany

Soft LVQ (SLVQ) was introduced in [1] as a probabilistic variant for learning vector quantization (LVQ) networks. Yet, we show in this contribution that the formulation of SLVQ does not describe a complete probabilistic model in the mathematical sense. Therefore, a modification of the original SLVQ is proposed to overcome this lack.

SLVQ assumes a data density $p(\mathbf{x})$, which is a mixture of the class conditional probabilities $p(\mathbf{x}|k)$, i.e.

$$p(\mathbf{x}) = \sum_{k=1}^{C} p(\mathbf{x}|c) P(c)$$

with $P(c)$ is the prior of class c. The class conditional probabilities $p(\mathbf{x}|c)$ are determined in SLVQ based on the prototype set $W = \{\mathbf{w}_1, \ldots, \mathbf{w}_M\}$ with class labels $c(\mathbf{w}_j)$, i.e. we have the estimators $p(\mathbf{x}, c|W)$ for $p(\mathbf{x}|c)$ as

$$p(\mathbf{x}, c|W) = \sum_{j=1}^{M} \delta_{c,c(\mathbf{w}_j)} p(\mathbf{x}|\mathbf{w}_j) P_W(\mathbf{w}_j) \tag{1}$$

where $P_W(\mathbf{w}_j)$ is the prior for the prototype $\mathbf{w}_j$ and $p(\mathbf{x}|\mathbf{w}_j)$ is the probability that prototype $\mathbf{w}_j$ has generated the data point $\mathbf{x}$. The function

$$\delta_{c,c(\mathbf{w}_j)} = \begin{cases} 1 & \text{if } c = c(\mathbf{w}_j) \\ \\ 0 & \text{if } c \neq c(\mathbf{w}_j) \end{cases}$$

is known as the Kronecker symbol. Analogously, we have

$$p\left(\mathbf{x},\bar{c}|W\right)=\sum_{j=1}^{M}\left(1-\delta_{c,c(\mathbf{w}_j)}\right)p\left(\mathbf{x}|\mathbf{w}_j\right)P_W\left(\mathbf{w}_j\right)\tag{2}$$

as the probability that $\mathbf{x}$ is not generated by the class c according to the SLVQ model. Learning in SLVQ takes place as an optimization of the log-likelihood

$$L_{SLVQ}\left(W\right)=\sum_{k=1}^{N}\log\left(\frac{p\left(\mathbf{x},c|W\right)}{p\left(\mathbf{x},\bar{c}|W\right)}\right)\tag{3}$$

by stochastic gradient learning. However, the probabilities $p\left(\mathbf{x},c|W\right)$ and $p\left(\mathbf{x},\bar{c}|W\right)$ do note generate a complete probabilistic model because the equality

$$p\left(\mathbf{x},c|W\right)+p\left(\mathbf{x},\bar{c}|W\right)=p\left(\mathbf{x}|W\right)$$

holds and generally the inequality $p\left(\mathbf{x}|W\right)=p\left(\mathbf{x}\right)\neq 1$ is valid. To obtain a complete probabilistic model, we consider the normalized probabilities

$$p_W\left(\mathbf{x},c\right)=\frac{p\left(\mathbf{x},c|W\right)}{p\left(\mathbf{x}\right)}\text{ and }p_W\left(\mathbf{x},\bar{c}\right)=\frac{p\left(\mathbf{x},\bar{c}|W\right)}{p\left(\mathbf{x}\right)}\tag{4}$$

with

$$p_W\left(\mathbf{x},c\right)+p_W\left(\mathbf{x},\bar{c}\right)=1\tag{5}$$

and result the new log-likelihood

$$\mathcal{L}_{SLVQ}\left(W\right)=\sum_{k=1}^{N}\log\left(\frac{p_W\left(\mathbf{x},c\right)}{p_W\left(\mathbf{x},\bar{c}\right)}\right)\tag{6}$$

for optimization with the constraint (5). In this model the prototypes do not longer act independently because they are mutually interacting via the condition (5). Surprisingly, one easily verifies that $\mathcal{L}_{SLVQ}\left(W\right)=L_{SLVQ}\left(W\right)$ is valid. However, the new cost function allows the reformulation

$$\mathcal{L}_{SLVQ}\left(W\right)=\sum_{k=1}^{N}\log\left(\frac{p_W\left(\mathbf{x},c\right)}{1-p_W\left(\mathbf{x},c\right)}\right)\tag{7}$$

according to the equality (5).

The consequence of this reformulation for optimization is a new stochastic gradient learning scheme for SLVQ containing only an attraction term for correctly classifying prototypes whereas the repulsion term known from standard SLVQ is vanished. We denote this scheme as *Attraction SLVQ* (ASLVQ). The other possibility is to treat (6) as an optimization problem with constraints, which requires an Lagrangian approach for optimization (L-SLVQ).

We will explain both approaches in detail during the talk at the MiWoCI-workshop.

References

[1] S. Seo and K. Obermayer. Soft learning vector quantization. *Neural Computation*, 15:1589–1604, 2003.

Task-Driven Sparse Coding for Classification of Motion Data

Babak Hosseini Barbara Hammer [*][†]

One of the current trends in the areas of image processing and motion analysis is to extract semantic entities in order to facilitate doing semantic search in the data bases and also to improve performance of high level approaches such as classification and clustering. 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.

To that aim, We utilize a non-negative variant of sparse coding (SC) based on a similarity kernel. The kernel is formed by using Dynamic time warping (DTW) which offers particularly successful pairwise motion data comparison. This combination leads to decomposition of motion data into a sparse linear composition of base functions which enables efficient data processing.

The other concern is having the mentioned decomposition and extraction in such a way that increases the classification performance of the motion data. As the approach we choose the linear classifier based on the generated sparse codes. And to formulate the joint ”sparse coding-classification” framework we use the task-driven optimization which relates one of the optimization parts (classifier) to the closed form solution of the other part (SC).

Although the coupled optimization framework converges to an optimal point, there are challenges regarding the quality of this optimal point. The constrained optimization framework is consist of two different objective function and being solved in an alternating fashion and is prone to be trapped in the local minimum points. Therefore we are looking for optimization techniques and conditions to overcome this problem and converges to a global optimum point or to an optimum point with satisfactory optimality.

[*] 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.

Extending RFSOM with DeepFeatures

Sven Hellbach, Thomas Neumann, Mathias Klingner,
Hans-Joachim Böhme *

*University of Applied Sciences Dresden

Klingner et al. [3] proposes an algorithm for posture estimation of a human body. The algorithm takes the approach from Haker et al. [2] and extends it using Generalized Matrix Learning Vector Quantization (GMLVQ).

The original approach [2] uses only a three dimensional space, i. e. direct spatial coordinates, as feature space to fit a self-organizing feature map (SOM) with a body-like topology. This leads to problems when individual regions of the person are in close proximity.

Hence, [3] decided to add textural information by training prototypical description of the body parts texture using GMLVQ. A bunch of typical texture descriptors, like RGB, HSV, HOG, LBP, GLCM, are precomputed. Interpreting the matrix Ω describing the adaptive metric in GMLVQ as relevance matrix gives a weighting of feature combination to discriminate the individual body regions. The learning metric together with the class prototypes are then used in the SOM.

Deep learning offers a possibility due to the deep architecture to learn the necessary features for a given problem. Hence, we decided to exchange the mentioned predefined textural features with the activities of earlier layers of a pre-trained deep network. We use VGG16 [4] as a convolutional neural network (CNN) specifically trained for image classification problems.

A similar approach to train the features was suggested by Giotis et al. [1]. However, they are focusing on feature with an analytic description, like Gabor wavelets, while the use of deep learning can be regarded as a more generalized formulation.

References

[1] Ioannis Giotis, Kerstin Bunte, Nicolai Petkov, and Michael Biehl. Adaptive matrices and filters for color texture classification. *Journal of Mathematical Imaging and Vision*, 47(1):79–92, Sep 2013.

[2] Martin Haker, Martin Böhme, Thomas Martinetz, and Erhardt Barth. Self-organizing maps for pose estimation with a time-of-flight camera. In *Proceedings of the DAGM 2009 Workshop on Dynamic 3D Imaging*, Dyn3D '09, pages 142–153, Berlin, Heidelberg, 2009. Springer-Verlag.

[3] Mathias Klingner, Sven Hellbach, Martin Riedel, Marika Kaden, Thomas Villmann, and Hans-Joachim Böhme. RFSOM Extending Self-Organizing Feature Maps with Adaptive Metrics to Combine Spatial and Textural Features for Body Pose Estimation. In *Advances in Self-Organizing Maps and Learning Vector Quantization*, volume 295 of *Advances in Intelligent Systems and Computing*, pages 157–166. Springer International Publishing, 2014.

[4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

Deep Learning and LVQ
Some first Results in Image Classification

Thomas Neumann*, Sven Hellbach*, and Markus Wacker*

*University of Applied Sciences, Dresden, Germany

Since the influential work by Krizhevsky, Sutskever, and Hinton [3], deep convolutional neural networks (ConvNets) have become the predominant approach for image classification, showing excellent performance especially when large amounts of training data is available. Recent work in this area mostly focused on novel network architectures, new activation functions, or clever optimisation algorithms. However, the final classification module on top of such a ConvNet essentially remained the same for years: a fully-connected layer with softmax activation, whose output is optimised with the cross-entropy loss. Different loss functions were studied only very recently [2].

Villmann et al. [4] propose another alternative: their theoretical arguments prove that generalised learning vector quantisation (GLVQ) can be combined with an (arbitrarily deep) neural network. With a similar idea in mind, Vries, Memisevic, and Courville [5] were already successful in providing first evidence for the practicability of this general idea - at least when the network is augmented with a supervised neural gas based loss. Our experiments show that modern ConvNets architectures such as pre-activation residual networks [1] and wide residual networks [7], architectures that contain over 10 million parameters, can be successfully trained directly with the classical GLVQ loss. These networks achieve up to 5.25% test error on the CIFAR-10 dataset, reaching almost the same accuracy as their softmax/cross-entropy counterparts (4.81% according to [7], 5.16% in our reimplementation of their paper). These first results show: deep learning quantizers seem to work in practice.

However, we also identified several problems with the feature representation that is discovered by these deep GLVQ networks. Their feature space seems unnecessarily sparse, especially when compared to the representations recovered by softmax/cross-entropy networks. Prototypes seem to latch onto individual, independent axes in feature space. Visualisation of the filter responses using the method by Yosinski et al. [6] reveal noisy and unintuitive prototypes.

We do not yet know if these observations actually hint at an inherent problem with our deep learning vector quantisers, or if they are just a manifestation of the intrinsic properties of one of the components, e.g. the GLVQ loss function itself or the metric used to measure distances to the prototypes. In any case, these observations provoke interesting questions:

- Is ℓ_2 the most suitable metric for deep GLVQ?
- Do we need completely different network architectures or training hyperparameters when training deep learning quantisers in order for them to converge to better solutions?
- Do we need to augment the loss function to further constrain the solution space - after all, this solution space is extremely large due to the immense number of degrees of freedom in a deep neural network?
- Do we need new visualisation techniques to uncover the feature representation of deep GLVQ networks?
- Is it hopeless to expect from deep learning vector quantisers that they yield intuitive feature dimensions and intuitive prototypes?
- How do we correctly recover and interpret the deep prototypes?

We aim to answer these questions in future work. Our hope is to establish the GLVQ scheme as a viable alternative to classical loss functions in the deep learning community, while at the same time expanding the capabilities of classical learning vector quantisation.

References

[1] Kaiming He et al. "Identity Mappings in Deep Residual Networks". In: (2016). arXiv: 1603.05027.

[2] Katarzyna Janocha and Wojciech Marian Czarnecki. "On Loss Functions for Deep Neural Networks in Classification". In: (2017). arXiv: 1702.05659.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25*. 2012, pp. 1097–1105.

[4] Thomas Villmann et al. "Combination of Deep Learning Architectures, Multilayer Feedforward Networks and Learning Vector Quantizers for Deep Classification Learning". In: *Advances in Self-Organizing Maps and Learning Vector Quantization: Proceedings of the 12th International Workshop WSOM 2017.* to appear.

[5] Harm de Vries, Roland Memisevic, and Aaron Courville. "Deep Learning Vector Quantization". In: *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning.* 2016.

[6] Jason Yosinski et al. "Understanding Neural Networks Through Deep Visualization". In: (2015). arXiv: 1506.06579.

[7] Sergey Zagoruyko and Nikos Komodakis. "Wide Residual Networks". In: (2016). arXiv: 1605.07146.

Differential Privacy for GLVQ

Johannes Brinkrolf Barbara Hammer

Bielefeld University, CITEC - Center of Excellence, Germany

jbrinkro@techfak.uni-bielefeld.de

Abstract

Digital information is collected daily in growing volumes. Mutual benefits drive the demand for the exchange and publication of data among parties. It is often unclear how to handle these data properly because the original data typically contains sensitive information. It is shown that simple anonymization of the data does not ensure de-identification, e.g., by linking the anonymized database with another one [2]. Differential privacy has become a powerful principle for privacy-preserving data analysis tasks in the last few years, which entails a formal privacy guarantee by separating the utility of the database and the risk due to individual participation. We briefly review the problem of statistical disclosure control under differential privacy model and show one example how the training of a GLVQ model can be change obey differential privacy. We first enhance the initialization by a simple differential private mechanism, and then use one differential private version of the stochastic gradient decent by Abadi et al. [1] for GLVQ training.

References

[1] Martín Abadi et al. "Deep Learning with Differential Privacy". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*. 2016, pp. 308–318.

[2] Arvind Narayanan and Vitaly Shmatikov. "Robust De-anonymization of Large Sparse Datasets". In: *2008 IEEE Symposium on Security and Privacy (S&P 2008), 18-21 May 2008, Oakland, California, USA*. 2008, pp. 111–125.

Going beyond global and static: Extending supervised linear transfer learning to more complex models

Alexander Schulz Benjamin Paaßen Barbara Hammer

June 29, 2017

The aim of transfer learning is to re-use knowledge from existing models in new domains and thereby avoid to train an entirely new model. This methodology is particularly promising if the trained model is complex but the relationship between the old and the new domain is simple, for example an approximately linear function. Recently, the framework of linear supervised transfer learning has been suggested which learns a mapping from target to source domain such that the original model combined with the mapping minimizes the model error on target space training data [2]. This framework has been applied successfully to counteract disturbances in bionic prosthesis control as well as transferring models trained on one hyper-spectral sensor to a different hyper-spectral sensor [2, 1]. However, in these cases the re-used model was a relatively simple GMLVQ model, which is fast to re-train. Linear supervised transfer learning promises even more added value for models which require substantially more data due to there inherent complexity. This contribution kicks off the journey toward transfer learning for more complex models.

References

[1] K. Berger, A. Schulz, B. Paaen, and B. Hammer. Linear supervised transfer learning for the large margin nearest neighbor classifier. In *submitted to the AIAI2017*, 2017.

[2] B. Paaßen, A. Schulz, J. Hahne, and B. Hammer. An EM transfer learning algorithm with applications in bionic hand prostheses. In M. Verleysen, editor, *Proceedings of the 25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2017)*, pages 129–134. i6doc.com, 2017.

Restricted Tangent Distances in Learning Vector Quantization

S. Saralajew[1] and T. Villmann[2]

[1] Electrical/ Electronics Driver Assistance Platform/ Systems,
Dr. Ing. h.c. F. Porsche AG Weissach, Germany

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

The classical tangent distance concept applied to the Learning Vector Quantization (LVQ,[1]) framework can be considered as an extension of the LVQ-prototype concept [2, 3, 4]. More precisely, in the original LVQ the prototypes are points in a vector space whereas in the tangent distance approach prototypes represent affine subspaces [5, 6].

In this contribution we present a modification of the tangent distance concept. Particularly, we restrict the affine subspaces to be only patches of affine subspaces, i.e. we consider a local representation. This idea leads to a modified tangent distance measure for LVQ learning, which is differentiable regarding the parameters of interest and, hence, can be adapted during training.

Obviously, the new distance measure can be plugged into an arbitrary distance-based machine learning framework and, therefore, is of general interest.

In the workshop contribution we demonstrate the working principles of the method for two toy data sets. Further, we address related open questions to stimulate ongoing research.

References

[1] Teuvo Kohonen. Learning Vector Quantization. *Neural Networks*, 1(Supplement 1):303, 1988.

[2] P. Simard, Y. LeCun, and J.S. Denker. Efficient pattern recognition using a new transformation distance. In S.J. Hanson, J.D. Cowan, and C.L. Giles, editors, *Advances in Neural Information Processing Systems 5*, pages 50–58. Morgan-Kaufmann, 1993.

[3] T. Hastie and P.Y. Simard. Metrics and models for handwritten character recognition. *Statistical Science*, 13(1):54–65, 1998.

[4] T. Hastie, P. Simard, and E. Säckinger. Learning prototype models for tangent distance. In G. Tesauro, D.S. Touretzky, and T.K. Leen, editors, *Advances in Neural Information Processing Systems 7*, pages 999–1006. MIT Press, 1995.

[5] S. Saralajew and T. Villmann. Adaptive tangent metrics in generalized learning vector quantization for transformation and distortion invariant classification learning. In *Proceedings of the International Joint Conference on Neural networks (IJCNN) , Vancover*, pages 2672–2679. IEEE Computer Society Press, 2016.

[6] S. Saralajew and T. Villmann. Transfer learning in classification based on manifold models and its relation to tangent metric learning. In *Proceedings of the International Joint Conference on Neural networks (IJCNN) , Anchorage*, pages 1756–1765. IEEE Computer Society Press, 2017.

Grassmann Manifolds, Hankel Matrices and Tangent Metric Models in Classification Learning

T. Villmann

Computational Intelligence Group,

University of Applied Sciences Mittweida, Germany

Pattern recognition frequently has to deal with noisy data describing objects or with different representations of objects, for example different illumination or rotations in image processing. Those data can be seen as particular sample vectors $\mathbf{x} \in \mathbb{R}^n$ belonging to a data space describing the object and its variations. A mathematical frame work for robust and adequate data processing is the concept of Grassmann manifolds equipped with the Riemannian geometry [1].

Supposing a k-frame of (orthogonal) sample vectors assigned to an object which are collected into a matrix $\mathbf{X} \in \mathbb{R}^{n \times k}$ with $0 < k \leq n$. Then the matrix $\mathbf{X}$ generates a linear subspace $\mathfrak{H}_k(\mathbf{X})$. The Grassmann manifold $\mathcal{G}_k^n$ is the space of k-dimensional linear subspaces (hyperplanes) $\mathfrak{H}_k$, i.e. the matrix $\mathbf{X}$ determines a certain point in the Grassmann manifold $\mathcal{G}_k^n$, see Fig. (1). Comparing of object representations $\mathbf{X}$ and $\mathbf{Y}$ is done as the calculation of distances between the linear subspaces $\mathfrak{H}_k(\mathbf{X})$ and $\mathfrak{H}_k(\mathbf{Y})$. For this purpose, several dissimilarity measures are known, most of them based on subspace angles $\theta_1, \ldots, \theta_k$ between the subspaces [2]. For example,

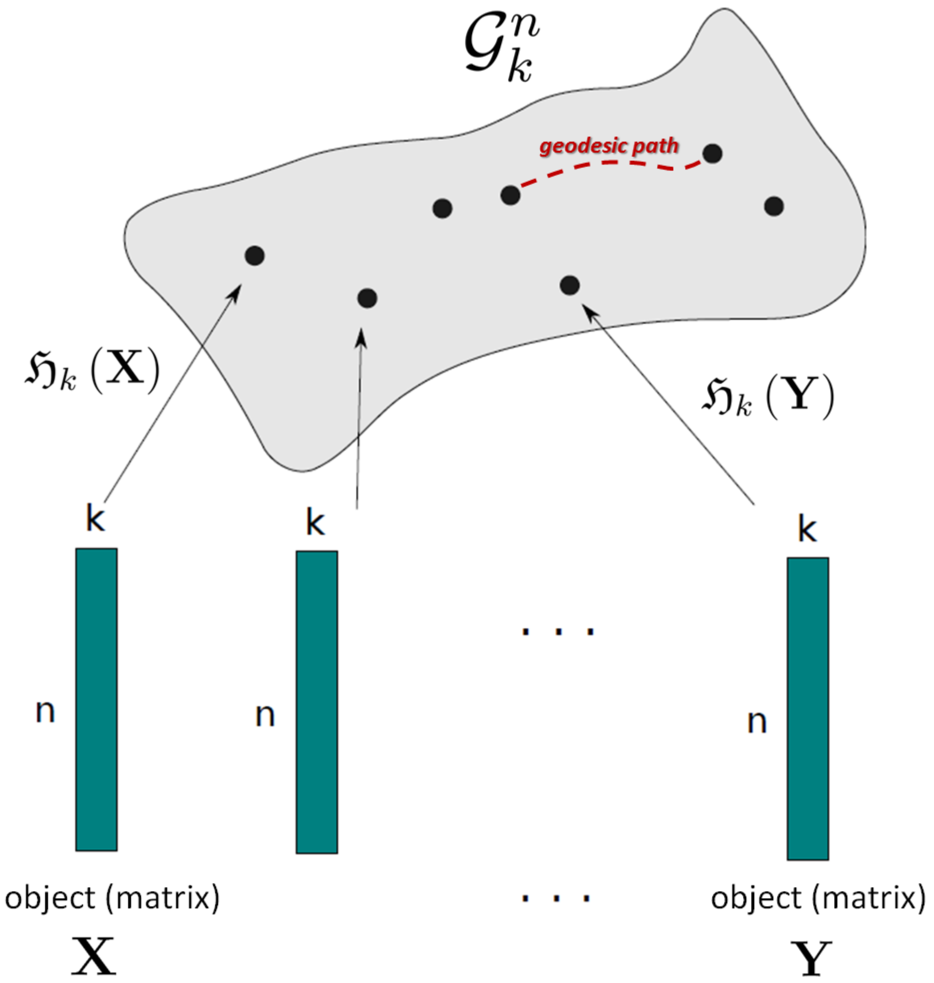


Figure 1: Illustration of a Grassmann manifold $\mathcal{G}_k^n$. Objects variations collected in a matrix generate linear subspaces $\mathfrak{H}_k$, which are points at the manifold. Distances between points are measured in terms of manifold distances. The geodesic distance $d_g(\mathfrak{H}_k(\mathbf{X}), \mathfrak{H}_k(\mathbf{Y}))$ from (1) is the path length along the geodesic path within the manifold.

the *geodesic distance* along the geodesic path in the manifold (see Fig.(1)) is

$$d_g\left(\mathfrak{H}_k\left(\mathbf{X}\right), \mathfrak{H}_k\left(\mathbf{Y}\right)\right) = \sqrt{\sum_{j=1}^{k} \theta_j^2} \tag{1}$$

whereas

$$d_c\left(\mathfrak{H}_k\left(\mathbf{X}\right), \mathfrak{H}_k\left(\mathbf{Y}\right)\right) = \sqrt{\sum_{j=1}^{k} \sin^2\left(\theta_j\right)} \tag{2}$$

is the *chordal distance* [3]. For the latter distance exists an isometrically embedding into the Euclidean space [4]. The geodesic distance realizes a non-Euclidean embedding. Both metrics can be seen also as examples to compare sets of vectors stored in the matrices $\mathbf{X}$ and $\mathbf{Y}$, i.e. they are particular realizations of a Hausdorff-metric [5].

Both distances can be immediately used in median (geodesic) or relational variants (chordal) of learning vector quantization for classification learning [6, 7]. Further, the geodesic distance can be used also in online learning vector quantizers using the derivative representation proposed and explained in [4] based on the mathematical considerations provided by [8].

In the contribution at the MiWoCI-workshop we explain the mathematical foundations of the Grassmannian approach and discuss basic properties related to classification learning. Further we show that this framework can be used also to compare tangent metric models in transfer learning or Hankel matrices in sequence discrimination learning as proposed for learning vector quantization in [9, 10].

References

[1] J. Hamm and D.D. Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 376–388, 2008.

[2] P.A. Wedin. *On angles between subspaces of a finite dimensional inner product space*, pages 263–285. Number 973 in Lectur Notes in Mathematics. Springer Berlin Heidelberg, Berlin, Heidelberg, 1983.

[3] S. Chepushtanova and M. Kirby. Sparse Grassmannian embeddings of hyperspectral data representations and classification. *IEEE Geoscience and Remote Sensing Letters*, 14(3):434–438, 2017.

[4] M. Kirby and C. Peterson. Visualizing data sets on the Grassmannian using self-organizing maps. In *Proceedings of the 12th Workshop on Self-Organizing Maps and Learning Vector Quantization (WSOM+ 2017), Nancy, France*, pages 32–37, Los Alamitos, 2017. IEEE Press.

[5] S. Saralajew, D. Nebel, and T. Villmann. Adaptive Hausdorff distances and tangent distance adaptation for transformation invariant classification learning. In A. Hirose, editor, *Proceedings of the International Conference on Neural Information Processing (ICONIP) , Kyoto*, volume 9949 of *LNCS*, pages 362–371. Springer, 2016.

[6] D. Nebel, B. Hammer, K. Frohberg, and T. Villmann. Median variants of learning vector quantization for learning of dissimilarity data. *Neurocomputing*, 169:295–305, 2015.

[7] B. Hammer, D. Hofmann, F.-M. Schleif, and X. Zhu. Learning vector quantization for (dis-)similarities. *Neurocomputing*, 131:43–51, 2014.

[8] P.-A. Absil, R. Mahony, and R. Sepulchre. Riemann geometry of Grassmannian manifolds with a view on algorithmic computation. *Acta Applicacandae Mathematika*, 80(2):199–200, 2004.

[9] S. Saralajew and T. Villmann. Transfer learning in classification based on manifold models and its relation to tangent metric learning. In *Proceedings of the International Joint Conference on Neural networks (IJCNN) , Anchorage*, pages 1756–1765. IEEE Computer Society Press, 2017.

[10] M. Mohammadi, M. Biehl, A. Bohnsack, and T. Villmann. Sequence learning in unsupervised and supervised vector quantization using Hankel matrices. In L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L.A. Zadeh, and J.M. Zurada, editors, *Proceedings of the 16th International Conference Artificial Intelligence and Soft Computing - ICAISC, Zakopane*, LNAI, pages 131–142, Cham, 2017. Springer International Publishing, Switzerland.

Causal inference to mitigate redundancy in feature selection

Lukas Pfannschmidt

Computational Methods for the Analysis of the Diversity and Dynamic of Genomes (IGK 1906)

Bielefeld University, Bielefeld, Germany

Abstract

In a classical learning scenario, feature selection is used to find relevant parts of the input space. In the past we lookeda gct the problem of finding feature relevance bounds for the all-relevant feature set on biomedical data. While previously relying on the statistical dependency between features to find correlated feature subsets, we now want to extend this approach by incorporating causal inference to add an ordering of features. This would enable a focused analysis of the causal factors in biomedical experiments and reduce redundancies. In this talk we want to present the problem of causal inference, in the context of feature selection and its potential in biomedical analysis, to encourage the audience to contribute their own ideas into this ongoing project.

Lifelong (machine) learning of drifting concepts in prototype-based classifiers

Michael Biehl*, Fthi Abadi*, and Christina Göpfert, Barbara Hammer[1]

*University of Groningen, Groningen, NL
[1]University of Bielefeld, Bielefeld, Germany

Abstract

Most frequently, frameworks of machine learning comprise of two different stages: First, in a training phase, a given set of example data is analysed, information is extracted and a corresponding hypothesis is parameterized in terms of, say, a classifier or regression system. In a subsequent working phase, this hypothesis is then applied to novel data.

For many practical applications of machine learning this separation is convenient and appears natural. A - by now - classical example would be the automated classification of handwritten digits by means of a neural network that has previously been trained from a large number of labeled input examples.

Obviously, the conceptual and temporal separation of training and working phase is not a very plausible assumption for human and other biological learning processes. Moreover, it becomes inappropriate if the actual task of learning, e.g. the target rule in a classification problem, changes continuously in time. In such a situation, the learning system must be able to detect and track the concept drift, i.e. forget irrelevant, older information while continuously adapting to more recent inputs.

In this contribution we present a mathematical model of learning drifting concepts in prototype-based classifiers, which are trained from high-dimensional data. Methods borrowed from statistical physics allow for the study of the typical learning dynamics for different training strategies in the presence of various drift scenarios. The mathematical framework is outlined, first results are presented and open questions are discussed.

3D Head Reconstruction via Convolutional Neural Networks Trained on Synthetic Images

Jan Philip Göpfert

Bielefeld University, CITEC - Center of Excellence, Germany

jgoepfert@techfak.uni-bielefeld.de

Convolutional Neural Networks can learn complicated mappings on images that would otherwise be difficult to formulate. However, a lack of labeled data can preclude the training of such networks. This is the case in the reconstruction of 3-dimensional human heads from 2-dimensional photographs. Approaching the problem backwards, starting from 3-dimensional heads and using photo-realistic rendering, one can create any number of training data to tackle the problem. This way, fine control over the data allows for new insights into how a Convolutional Neural Network interprets data and how hidden variables affect its performance.

Short Summary of Relational and Median Variants of Possibilistic Fuzzy C-Means

Tina Geweniger
Faculty Applied Computer Sciences and Biosciences
University of Applied Sciences Mittweida
Mittweida, Germany
Email: tgewenig@hs-mittweida.de

***Abstract*—Possibilistic Fuzzy c-Means is a clustering technique introduced by Pal et al. in 2005. To position cluster representative prototypes the method takes probabilistic and possibilistic cluster assignments into account. We extend this method to handle non-vectorial data. Thereby, we first assume that a Euclidean data embedding is possible and derive a relational variant. The second proposed modification aims to perform a clustering based on abstract data objects. Here only their dissimilarities are known and representative data samples are selected as prototypes.**

I. Introduction

In [1] Pal et. al proposed a special kind of c-Means taking probabilistic and possibilistic cluster assignments into account. They combined both paradigms in one cost function balancing their influence by user-defined parameters. This way a soft clustering can be obtained, where each data point belongs to one or more clusters depending on the distance or similarity to all other clusters. Yet there is a difference between the interpretation of probabilistic and possibilistic assignments. If a data point is equidistant to two arbitrary clusters, the probabilistic membership values to both cluster centers are the same no matter of the actual distance to both clusters. Therefore, in this case the membership value does not hold much information about the similarity between data point and prototypes. To circumvent this problem possibilistic assignments are integrated into the algorithm. These assignments are treated differently and can also be interpreted as *typicalities*. The higher the similarity between data point and cluster center, the higher the typicality value. Therefore, while probabilistic restrictions force data points to belong to one ore more clusters, the possibilistic assignments allow to detect outliers [1]. Figure 1 illustrates the difference between probabilistic membership and typicality. Yet as for the common C-Means the data points have to be provided in vectorial form and a Euclidean embedding is assumed.

If only dissimilarity data exists, this algorithm has to be modified. For data which can be embedded into the Euclidean space we propose a relational variant. If this embedding is not possible we developed an appropriate median version by modifying the algorithm to work with given dissimilarities of the objects and to select representative data samples as prototypes.

In our article [2] presented at the WSOM 2017 we provided the mathematical framework, the modified algorithms and update rules, and the proofs of convergence for both median and relational data. Here we present a short survey of the most important features of the modifications to allow easy use and implementation.

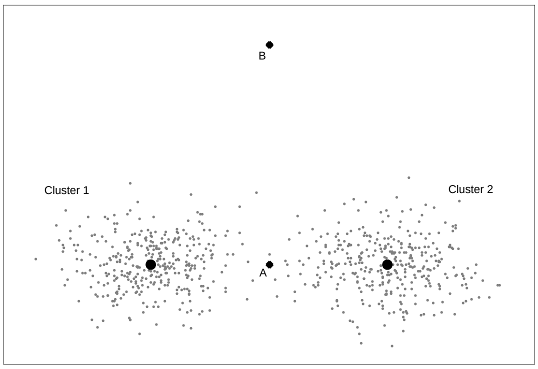


Fig. 1. Two normally distributed two-dimensional clusters with same variance and number of data samples. Data Point A and B are equidistant to both cluster centers. The probabilistic assignments (fuzzy memberships) of A and B to the cluster centers are identical $u_{A1} = u_{A2} = u_{B1} = u_{B2} = 0.5$. Yet the possibilistic assignments (typicalities) are different: The values of A are much higher than the values of B $t_{A1} = t_{A2} > t_{B1} = t_{B2}$ (notation see next section).

In the following you find three sections. The first one describes the Possibilistic Fuzzy C-Means (PFCM) as introduced by Pal et al. [1] and the following two sections describe our modifications concerning relational and median data respectively.

II. Possibilistic Fuzzy C-Means

The Possibilistic Fuzzy C-Means incorporating probabilistic (fuzzy) cluster assignments and possibilistic typicalities was proposed by Pal et al. [1]. The cost function is defined as

$$\begin{aligned} J_{PFCM}\left(\mathbf{U},\mathbf{T},\mathbf{W};\mathbf{X}\right) &= \sum_{k=1}^{n}\sum_{i=1}^{c}\left(a\cdot u_{ik}^{m}+b\cdot t_{ik}^{\eta}\right)\cdot d_{ik}^{2} \\ &\quad +\sum_{i=1}^{c}\gamma_{i}\sum_{k=1}^{n}\left(1-t_{ik}\right)^{\eta} \qquad (1) \end{aligned}$$

where we have a set $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ of n d-dimensional data points $\mathbf{x}_k \in \mathbf{R}^d$ and $d_{ik} = d\left(\mathbf{x}_k, \mathbf{w}_i\right)$ is a dissimilarity measure. The aim of PFCM is to cluster the data and to find representative prototypes $\mathbf{w}_i \in \mathbb{R}^d$. The set of all c prototypes is given by $\mathbf{W} = \{\mathbf{w}_1, ..., \mathbf{w}_c\}$ whereby $n \gg c$. The probabilities u_{ik} and the typicalities t_{ik} are subject to some restrictions: $u_{ik} \in [0,1]$ with $\sum_{i=1}^{c} u_{ik} = 1 \ \forall k$ and $t_{ik} \in [0,1]$. The exponents $m > 1$ and $\eta > 1$ control the degree of fuzziness and typicality. The matrices $\mathbf{U}$ and $\mathbf{T}$ are both of size $c \times n$ and hold the fuzzy assignments and typicality values respectively. The parameters $a \geq 0$, $b \geq 0$, and $\gamma_i \geq 0$ balance the probabilistic and possibilistic influence on the cost function. In [1] the authors Pal et al. explicitly mention that the condition $a + b = 1$ does not have to be fulfilled.

For the special case of $b = 0$ and $\gamma_i = 0 \ \forall i$ the cost function in eq. (1) reduces to the objective function of the common Fuzzy C-Means [3]. Setting $a = 0$ the cost function of the Possibilistic C-Means [4] is obtained.

The second term of the cost function 1 was introduced to put a constrained on the typicalities to avoid the problem of very small typicality values for large data sets.

The algorithm following an alternating optimization scheme is given in alg. 1.

Algorithm 1 Possibilistic Fuzzy C-Means (PFCM) [1]

```
set number $c$ of prototypes
initialize all parameters
initialize prototypes randomly within the data space
repeat
    update probabilistic assignments followed by normalization
        $u_{ik} = \left( \sum_{j=1}^{c} \left( \frac{d_{ik}}{d_{jk}} \right)^{2/(m-1)} \right)^{-1}$
    update typicalities (possibilistic assignments)
        $t_{ik} = \left( 1 + \left( \frac{b}{\gamma_i} d_{ik}^2 \right)^{1/(\eta-1)} \right)^{-1}$
    update prototypes
        $\mathbf{w}_i = \frac{\sum_{k=1}^{n} (a \cdot u_{ik}^m + b \cdot t_{ik}^\eta) \mathbf{x}_k}{\sum_{k=1}^{n} (a \cdot u_{ik}^m + b \cdot t_{ik}^\eta)}$
    optionally adapt $a$ and $b$ to recalibrate the influence of probabilities and typicalities
until convergence or manual stop
```

The update rules are obtained by applying the Lagrange multiplier theorem to minimize the cost function (1) [1]. Please note, that the update rule for the prototypes $\mathbf{w}_i$ is only valid if the squared Euclidean distance is used as dissimilarity measure $d(\mathbf{x}_k, \mathbf{w}_i) = \|\mathbf{x}_k - \mathbf{w}_i\|$. Further consideration concerning parameter initialization and in-depth hints are also provided in [1].

III. Relational PFCM

When dealing with abstract or non-vectorial data objects it is not possible to use the data samples x_k themselves for clustering. However, assume that the distance matrix $\mathbf{D} \in \mathbb{R}_+^{n \times n}$ containing the dissimilarities $D_{ij} = d(x_i, x_j)$ of the n objects is provided. If further it can be assumed that there exists a (possibly non-linear) mapping

$$g(x_k) = \mathbf{v}_k \tag{2}$$

with $\mathbf{v}_k \in V$ projecting the data objects into a possibly high-dimensional Euclidean embedding space V such that $D_{ij} = d_V^2(\mathbf{v}_i, \mathbf{v}_j)$ is the squared Euclidean distance in V then the prototypes $\mathbf{w}_i inV$ can be defined as convex linear combinations of the data and can be described as

$$\mathbf{w}_i = \sum_{j=1}^{n} \alpha_{ij} \mathbf{v}_j \tag{3}$$

with $\alpha_{ij} \geq 0$ and $\sum_{j=1}^{n} \alpha_{ij} = 1$.

We can write the distance between data and the weight vectors as

$$d_V^2(\mathbf{v}_k, \mathbf{w}_i) = \sum_j \alpha_{ij} \cdot d_V^2(\mathbf{v}_k, \mathbf{v}_j) - \frac{1}{2} \boldsymbol{\alpha}_i^T \cdot \mathbf{D} \cdot \boldsymbol{\alpha}_i \tag{4}$$

with $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{in})^T$ being the vector of the embedding coefficients [5] and $\mathbf{D}$ as matrix of given data dissimilarities D_{ij} [6].

The cost function is structurally identical with (1). Yet now, instead of calculating distance values the values provided in $\mathbf{D}$ are used.

The update of prototypes, probabilities, and typicalities takes place by Stochastic Gradient Descent Learning (SGDL) and considering a Lagrange function taking all restrictions into account [2]. The updated algorithm for relational data is given in alg. 2.

Note that no real prototypes are obtained. Instead they are described indirectly by the coefficient vectors $\boldsymbol{\alpha}_i$, i. e. by setting $\boldsymbol{\alpha}_i$ virtual prototypes are generated.

Further details, hints, derivation of update rules, and proof of convergence can be found in [2].

IV. Median PFCM

Again we assume that only the distance matrix $\mathbf{D} \in \mathbb{R}_+^{n \times n}$ of all objects of the data set is available. Yet we drop all restrictions regarding triangle inequality and symmetry, i. e. we no longer assume that there is an underlying Euclidean metric. Therefore, we cannot use the relational ansatz of R-PFCM. Instead we have to select representative data samples to act as prototypes for emerging clusters. As before, the overall goal is to minimize a cost function structurally equivalent to (1)

$$\begin{aligned} J_{M-PFCM} &= \sum_{k=1}^{n} \sum_{i=1}^{c} (a \cdot u_{ik}^m + b \cdot t_{ik}^\eta) \cdot D_{J(i)k}^2 \\ &\quad + \sum_{i=1}^{c} \gamma_i \sum_{k=1}^{n} (1 - t_{ik})^\eta \end{aligned} \tag{5}$$

where $J(i) = l$ is a mapping function which takes the index i of prototype w_i as parameter and refers to index l of the respective identical data sample x_l. The dissimilarities

Algorithm 2 Relational Poss. Fuzzy C-Means (R-PFCM)

```
set number $c$ of prototypes
initialize all parameters
initialize coefficient vectors $\boldsymbol{\alpha}_i$ taking the restrictions $\alpha_{ij} > 0$ and $\sum_{j=1}^{n} \alpha_{ij} = 1$ into consideration
repeat
    calculate distances $d_V^2(\mathbf{v}_k, \mathbf{w}_i)$ according eq. (4)
    update probabilistic assignments followed by normalization
        $u_{ik} = \left( \sum_{j=1}^{c} \left( \frac{d_{ik}}{d_{jk}} \right)^{2/(m-1)} \right)^{-1}$
    update typicalities (possibilistic assignments)
        $t_{ik} = \left( 1 + \left( \frac{b}{\gamma_i} d_{ik}^2 \right)^{1/(\eta-1)} \right)^{-1}$
    update coefficient vectors $\boldsymbol{\alpha}_i$
        $\Delta\boldsymbol{\alpha}_i = (a \cdot u_{ik}^m + b \cdot t_{ik}^\eta) \left( \mathbf{D}_k - \mathbf{D}_i \sum_{j=1}^{n} \alpha_{ij} \right)$
    normalize coefficient vectors $\boldsymbol{\alpha}_i$
        $\alpha_{ij} = \frac{\alpha_{ij}}{\sum_{l=1}^{c} a_{il}}$
    optionally adapt $a$ and $b$ to recalibrate the influence of probabilities and typicalities
until convergence or manual stop
```

$D_{J(i)k} = d(x_k, w_i) = d(x_k, x_l)$ are taken directly from the given dissimilarity matrix $\mathbf{D} \in \mathbb{R}^{n \times n}$.

In alg. 3 all steps of the algorithm following an alternating optimization scheme are summarized.

As before, further details and the proof of convergence can be found in our paper presented at the WSOM 2017 [2].

References

[1] N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek, "A possibilistic fuzzy c-means clustering algorithm," *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 4, pp. 517–530, Aug 2005.

[2] T. Geweniger and T. Villmann, "Relational and median variants of possibilistic fuzzy c-means," in *WSOM+ 2017 12th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization*, 2017, pp. 207–213.

[3] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, 1973.

[4] R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering," *IEEE Transactions on Fuzzy Systems*, vol. 1, no. 2, pp. 98–110, May 1993.

[5] B.Hammer and A.Hasenfuss, "Relational neural gas," *Künstliche Intelligenz*, pp. 190–204, 2007.

[6] A. Hasenfuss, B. Hammer, F.-M. Schleif, and T. Villmann, "Neural gas clustering for dissimilarity data with continuous prototypes," in *Computational and Ambient Intelligence – Proceedings of the 9th Work-conference on Artificial Neural Networks (IWANN), San Sebastian (Spain)*, ser. LNCS 4507, F. Sandoval, A. Prieto, J. Cabestany, and M. Grana, Eds. Berlin: Springer, 2007, pp. 539–546.

Algorithm 3 Median Possibilistic Fuzzy C-Means (M-PFCM)

```
set number of $c$ prototypes
initialize all parameters
select randomly $c$ data samples as prototypes
repeat
    update probabilistic assignments followed by normalization
        $u_{ik} = \left( \sum_{j=1}^{c} \left( \frac{d(x_k, x_{J(i)})}{d(x_k, x_{J(j)})} \right)^{2/(m-1)} \right)^{-1}$
    update possibilistic typicalities
        $t_{ik} = \left( 1 + \left( \frac{b}{\gamma_i} d(x_k, x_{J(i)})^2 \right)^{1/(\eta-1)} \right)^{-1}$
    select new data samples as prototypes
        $l = \underset{l'}{argmin} \left( \sum_{k=1}^{n} (a \cdot u_{ik}^m + b \cdot t_{ik}^\eta) d_{kl'}^2 + \gamma_i \sum_{k=1}^{n} (1 - t_{ik})^\eta \right)$
    optionally adapt $a$ and $b$ to recalibrate the influence of probabilities and typicalities
until convergence or manual stop
```

Two or three things we do (not) know about distances

Benjamin Paassen
bpaassen@techfak.uni-bielefeld.de
Machine Learning Group
CITEC Center of Excellence
Bielefeld University

Some of the earliest approaches in pattern recognition and machine learning rely on distances between data [1]. This representation of data in terms of distance or similarity is motivated by cognitive science, which tries to explain cognitive skills such as categorization, memory retrieval, reasoning and induction by judgments of similarity [2, 3]. The earliest work attempted to identify an underlying Euclidean space corresponding to the similarities reported by human subjects [7]. However, critics have pointed out that similarities in a cognitive sense do not adhere to classic mathematical axioms, such as symmetry or the triangular inequality [8]. Further, recent work in cognitive science has aimed at describing distances in terms of the transformations required to turn one object into another [2]. Guided by these frameworks from cognitive science, this contribution will discuss two or three things we do (not) know about distances in the machine learning context, namely that and how they can be embedded in (pseudo-)Euclidean spaces [5], how we can deal with non-Euclidean in indefinite distances [6] and how we can efficiently compute transformation-based distances [4].

References

[1] Thomas M. Cover and Peter E. Hart. “Nearest neighbor pattern classification”. In: *IEEE Transactions on Information Theory* 13.1 (1967), pp. 21–27. DOI: 10.1109/TIT.1967.1053964.

[2] Carl J. Hodgetts, Ulrike Hahn, and Nick Chater. “Transformation and alignment in similarity”. In: *Cognition* 113.1 (2009), pp. 62–79. DOI: 10.1016/j.cognition.2009.07.010.

[3] D. Nebel et al. “Types of (dis-)similarities and adaptive mixtures thereof for improved classification learning”. In: *Neurocomputing* (2017). in press. DOI: 10.1016/j.neucom.2016.12.091.

[4] Benjamin Paassen, Bassam Mokbel, and Barbara Hammer. “Adaptive structure metrics for automated feedback provision in intelligent tutoring systems”. In: *Neurocomputing* 192 (2016), pp. 3–13. DOI: 10.1016/j.neucom.2015.12.108.

[5] Elżbieta Pękalska and Robert P. W. Duin. *The dissimilarity representation for pattern recognition: foundations and applications*. Ed. by H. Bunke and P. S. P. Wang. Vol. 64. Series in Machine Perception and Artificial Intelligence. Singapore, 2005. ISBN: 981-256-530-2.

[6] Frank-Michael Schleif and Peter Tino. “Indefinite Proximity Learning: A Review”. In: *Neural Computation* 27.10 (2015), pp. 2039–2096. DOI: 10.1162/NECO_a_00770.

[7] Roger N. Shepard. “The analysis of proximities: Multidimensional scaling with an unknown distance function. I.” In: *Psychometrika* 27.2 (1962), pp. 125–140. DOI: 10.1007/BF02289630.

[8] Amos Tversky. “Features of similarity”. In: *Psychological Review* 84.4 (1977), pp. 327–352. DOI: 10.1037/0033-295X.84.4.327.

Learning in Krein Spaces
Challenges and Perspectives

Frank-Michael Schleif*

*University of Applied Sciences Wuerzburg-Schweinfurt,
Wuerzburg, Germany

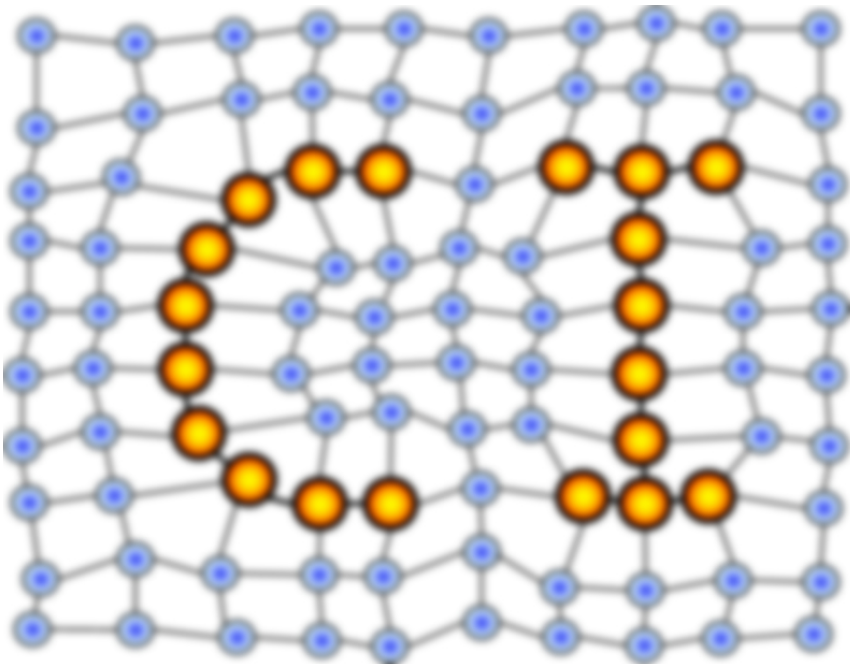
frank-michael.schleif@fhws.de

Abstract

Non-metric proximity measures got wide interest in various domains such as life sciences, robotics and image processing. The majority of learning algorithms for these data are focusing on classification problems and make use of heuristics or complicated optimization schemes to cope with an indefinite kernel matrix or non-metric distances. We discuss some basic concepts about non-metric measures and the respective learning in a Krein space.

MACHINE LEARNING REPORTS

Report 01/2017



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/

Prof. Dr. rer. nat. Frank-Michael Schleif
University of Birmingham
Edgbaston, B15 2TT Birmingham, UK,
• www.cs.bham.ac.uk/∼schleify/

▽ Acknowledgments

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