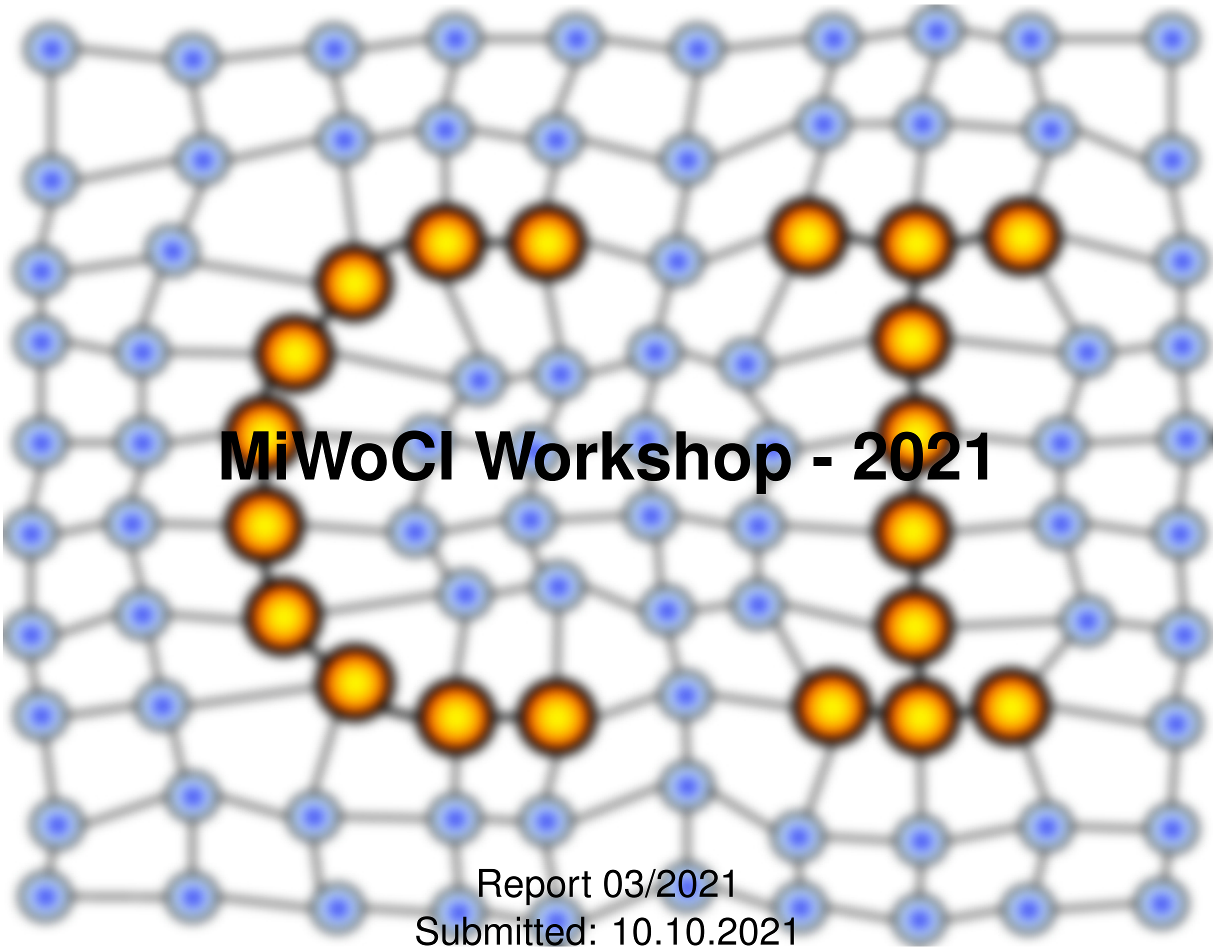
MACHINE LEARNING REPORTS



MiWoCI Workshop - 2021

Report 03/2021
Submitted: 10.10.2021
Published: 22.10.2021

Frank-Michael Schleif[1,2*,3*], Marika Kaden[2], Thomas Villmann[2] (Eds.)
(1) University of Applied Sciences Wuerzburg-Schweinfurt, 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

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. 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 13^{th} Mittweida Workshop on Computational Intelligence - MiWoCI 2021 -

Frank-Michael Schleif, Marika Kaden, and Thomas Villmann

Machine Learning Report 03/2021

Preface

The 13 th international *Mittweida Workshop on Computational Intelligence* (MiWoCI) gathering together more than 50 scientists from different universities including Bielefeld, Groningen, UAS Mittweida, UAS Würzburg-Schweinfurt, UAS Zwickau, Dr. Ing. h.c. F. Porsche AG in Weissach and IFF Fraunhofer in Magdeburg. This year we could again gathering together in Mittweida, Germany. For all who could join in person the workshop was hybrid. Thus, from 1.9- 3.9.2021 the tradition of scientific presentations, vivid discussions, and exchange of novel ideas at the cutting edge of research was continued. They were 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

Advances and new developments in the machine learning analysis of steroid metabolomics data

Michael Biehl

Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, The Netherlands

Abstract

In this presentation, recent results and novel developments in the analysis of steroid metabolomics data will be summarized. These studies are part of a long-standing collaboration of the Intelligent Systems Group at the University of Groningen with the Institute of Metabolism and Systems Research at the University of Birmingham/UK. Examples include applications of unsupervised and supervised machine learning for

- the differentiation of various types of benign adrenocortical tumors
- the discovery of potential subtypes of adrenocortical carcinoma
- the discrimination of different forms of Cushing's disease
- the analysis of high-dimensional "non-targeted" metabolomics data

Most recent, relevant publications:

- I. Bancos, A. Taylor, V. Chortis, A. Sitch et al. Urine steroid metabolomics for the differential diagnosis of adrenal incidentalomas in the EURINE-ACT study: a prospective test validation study The Lancet Diabetes and Endocrinology Vol. 8 (issue 9): 773-781, 2020 (open access)
- A. Moolla, J. de Boer, D. Pavlov, A. Amin et al. Accurate non-invasive diagnosis and staging of non-alcoholic fatty liver disease using the urinary steroid metabolome Alimentary Pharmacology and Therapeutics 51: 1188-1197, 2020 (open access

Machine Learning-Based Classification of Movement Disorders

Elina van den Brandhof

Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, The Netherlands

Abstract

Hyperkinetic movement disorders cause excessive, involuntary movements. A diagnosis is made by experts and is based on clinical definition. Despite their expertise, this subjective approach can lead to variable diagnoses among different experts or even be inaccurate.

The project 'Next Move in Movement Disorders (NEMO)' aims to develop a machine learning-based diagnostic system to support experts in the classification of hyperkinetic movement disorders. It involves the investigation of 3D camera images, muscle activity measured by electromyography (EMG), and motion measured by accelerometry (ACC). In this study, we focus on the analysis of EMG and ACC data.

Primarily, we selected three movement disorders, including tremor ($n = 37$), myoclonus ($n = 20$), and dystonia ($n = 19$) and healthy controls ($n = 28$), of the constantly increasing dataset. For each participant, 36 tasks, including static, dynamic, and distraction tasks, were recorded by sixteen combined EMG and ACC sensors placed on the upper body.

The literature defines a wide range of features extracted from EMG and ACC data that describe typical signs of movement disorders and help classify them. Our approach builds on the clinical approach to make the decision comprehensible and transparent. We build a set of features extracted from the frequency domain and from the time domain, including both features defined in literature and self-defined features that might contribute to the decision-making.

Currently, we are defining features and extracting them from our dataset. Subsequently, we will use relevance learning techniques, such as GMLVQ and Random Forest, possibly in combination with a principal component analysis to reduce the dimension of the feature space. In addition, relevance learning could provide new insights into characteristic features of the three hyperkinetic movement disorders and expand the neurological toolbox.

Orthogonal Learning Correction

Rick van Veen, Neha Rajendra Bari Tamboli, Michael Biehl

Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, The Netherlands

Abstract

Frequently, data comes from different sources and must be combined in a single machine-learning system. Multi-source data can lead to systems using variation not intrinsic to the classes one wishes to distinguish. Although multiple sources of variation do not necessarily degrade classification performance, it is detrimental for interpreting the resulting machine-learning systems. In order to gather enough FDG-PET data for the classification of neurodegenerative disease, collection needs to happen at more than one neuroimaging center. This multi-center data contains unwanted sources of variation explained by factors, such as different scanners, differences in scanning protocols, and processing methods. First, using Generalized Matrix Learning Vector Quantization (GMLVQ), we can find a 'relevance space' explaining only the difference between scans taken at distinct centers. Second, a correction matrix can be constructed based on this 'relevance space.' Third, we introduce an extension to the GMLVQ learning process that uses the correction matrix to limit training to the space that does not contain the variance between the centers. We use this novel technique on our multi-center dataset in various settings and show that center effects can indeed be removed. Consequently, this produces cleaner prototypes and more expressive relevance profiles for further interpretation by medical experts. Furthermore, this method can be used on similar problems outside the neuroimaging domain, as long as an appropriate 'relevance space' can be found to construct the correction matrix.

Towards Informed Posterior Construction by Structural Identifiability Analysis

Janis Norden

Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, The Netherlands

Abstract

Theory-guided Machine Learning (see [1] for an overview) allows insight into the underlying nature and dynamics of the processes which produce the data, as opposed to conventional (i.e. purely data-driven) methods. This is particularly important in fields such as medicine and industrial engineering, where individual decisions potentially have grave consequences. In this presentation, we will give an overview of recent developments with respect to the problem of parameter identification, where it is assumed that the observed data has been generated by a parameterized dynamical system.

A first step in our approach is to study structural identifiability for linear systems of ODEs. Identifiability analysis following the Laplace transform approach (see e.g. Godfrey and DiStefano [2]) is demonstrated on a simple two-compartment linear system with four parameters. In this system, only one of the two compartments is observable, i.e. time-series data is obtained for only one of the compartments. The identifiability analysis yields a number of parameter relations, where any parameter configuration satisfying such a relation will yield identical observable output for the system of ODEs. We will discuss and demonstrate that such analysis provides insight into the nature of uncertainties arising in parameter identification which comprise 1) uncertainties due to noisy data and 2) uncertainties caused by structural unidentifiability inherent to the dynamical model.

References

[1] A. Karpatne, G. Atluri, J. H. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar. Theory-guided Data Science: A New Paradigm for Scientific Discovery from Data. *IEEE Transactions on knowledge and data engineering*, 29(10):2318–2331, 2017.

[2] K. Godfrey and J. DiStefano. Identifiability of Model Parameters. *IFAC Proceedings Volumes*, 18(5):89–114, 1985. 7th IFAC/IFORS Symposium on Identification and System Parameter Estimation, York, UK, 3-7 July.

Using Information Geometric Principles For Theory-Guided Machine Learning

Elisa Oostwal

Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, The Netherlands

Abstract

Theory-guided Machine Learning (see [4] for an overview) allows insight into the underlying nature and dynamics of the processes which produce the data, as opposed to conventional (i.e. purely data-driven) methods. This is particularly important in fields such as medicine and industrial engineering, where individual decisions potentially have grave consequences. In this presentation, we discuss recent developments with respect to the problem of parameter estimation, where it is assumed that the observed data has been generated by some parameterized dynamical system.

An important concept is Learning in the Model Space [2, 5]. In this approach, each individual series of measurements (time-series) is represented by a posterior distribution over model parameters, which are assumed to be given by a parameterized dynamical system. The posterior will model uncertainty both due to sampling noise and due to the unidentifiabilities which are inherent to the dynamical system. We will introduce information geometric principles, such as the natural gradient [1], and will provide examples to demonstrate their use in this setting. Additionally, we will present an overview of sampling methods which make use of these principles [3].

References

[1] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry.* Vol. 191.Jan. 2000.

[2] uanhuan Chen, Peter Tino, Ali Rodan, and Xin Yao. “Learning in the ModelSpace for Fault Diagnosis”. In: *IEEE Transactions on Neural Networks* (Mar. 2013). DOI: 10.1109/TNNLS.2013.2256797.

[3] Mark Girolami and Ben Calderhead. “Riemann manifold Langevin and Hamiltonian Monte Carlo methods”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.2 (2011), pp. 123-214. DOI: https://doi.org/10.1111/j.1467-9868.2010.00765.x.

[4] Anuj Karpatne, Gowtham Atluri, James H. Faghmous, Michael S. Steinbach,Arindam Banerjee, Auroop R. Ganguly, Shashi Shekhar, Nagiza F. Samatova,and Vipin Kumar. "Theory-guided Data Science: A New Paradigm for Scientific Discovery". In: *CoRRabs/1612.08544* (2016). arXiv: `1612.08544`. URL: `http://arxiv.org/abs/1612.08544`.

[5] Yuan Shen, Peter Tino, and Krasimira Tsaneva-Atanasova. "A Classification Framework for Partially Observed Dynamical Systems". In: *Phys. Rev. E* 95 (4 Apr. 2017), p. 043303. DOI: `10.1103/PhysRevE.95.043303`. URL: `https://link.aps.org/doi/10.1103/PhysRevE.95.043303`.

Geodesic ensembling of MLVQ models

Kerstin Bunte

Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, The Netherlands

Abstract

Ensembling is well known to often improve the generalization of machine learning. However, the naive combination of models, which are trained using some stochastic variation comes at the cost of computational effort, since all model parameters have to be saved. Additionally it complicates the interpretation even when intrinsically transparent and interpretable models are used. In this talk we will present a strategy to build an average ensemble model that uses the geodesic over the model parameters. We demonstrate results over biomedical data sets in comparison to the naive ensemble and statistics of individual model performances, as well as analysis of the interpretability.

On the transition from hyperspectral to multispectral data - Implications for functional data based ML modeling

Patrick Menz[1], Valerie Vaquet[2], Udo Seiffert[1] and Barbara Hammer[2]

1 - Cognitive Processes and Systems, Fraunhofer Institute for Factory Operation and Automation IFF Magdeburg, Germany
2 - Machine Learning Group, Bielefeld University, Bielefeld, Germany

Hyperspectral imaging is increasingly establishing itself as an important and powerful instrument through its non-invasive evaluation method in different fields of application, such as quality control, agriculture, forensic, and much more. However, it is still an expensive measurement system and not applicable for everyone. In many applications, it can be shown that only a few wavelengths are actually needed to perform the same task without much loss of performance. We want to show the transition from an expensive hyperspectral measurement system to cost-effective multispectral measurement hardware. Furthermore, we want to show that we have to deal with drifts, which are already well known for hyperspectral systems [1] due to various influences.

Since this is now semi-functional data, it will also be examined whether existing methods [2] and new methods [3] are still applicable to get rid of drifts.

References

[1] Patrick Menz, Andreas Backhaus, Udo Seiffert. Transferring machine learning models within a soft sensor system to achieve constant task performance under changing sensor hardware. In *Machine Learning Reports 2016.*

[2] Bouveresse, E. & Massart, D.L.. Improvement of the piecewise direct standardizationprocedure for the transfer of nir spectra for multivariate calibration. In *Chemom. Intell.Lab. Syst., 32:201-213, 1996.*

[3] Valerie Vaquet, Patrick Menz, Udo Seiffert & Barbara Hammer. Investigating Drift in Hyperspectral Imaging Data In *Machine Learning Reports 2021.*

Adaptive Gabor Filters for Color Texture Classification

Gerrit Luimstra

Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, University of Groningen, The Netherlands

Abstract

In this thesis we present a method of creating interpretable and adaptable feature extractors for image classification based on the Gabor function. The feature extractors are built into a variant of learning vector quantization, after which the algorithm is tweaked so that it uses an adaptable Gabor filter bank instead of a static one. The advantage of adaptive Gabor filters is that strong visual features can be obtained automatically in a domain agnostic way, by constricting the filter hypothesis space of the model to crisp edge detectors, which also leads to a big reduction in trainable parameters. By studying the derivatives of the Gabor function, we find that 4 of the 5 parameters that make up a Gabor filter lend itself well to adaptation, and careful parameter regularization is included to prevent invalid Gabor filters from being obtained. Additionally, we add modifications to the learning vector quantization algorithm used, and derive the learning rules for a real-valued and compact version. We evaluate the proposed techniques on an artificial dataset and real dataset consisting of color texture images using 4-fold stratified cross-validation and compare results with three similar models. The proposed techniques exhibit very good performance as well as stability, interpretability and the ability to generalize well on out-of-sample data. Hence adaptive Gabor filters provide promising results for accessible and efficient image classification.

AutoML technologies for configuration of the sparse and accurate models

Aleksei Liuliakov

Machine Learning Group, Bielefeld University, Germany

Abstract

Automated machine learning (AutoML) technologies established by a number of frameworks which perform efficient machine learning (ML) model configuration and its hyperparametrs adjustment for specific ML tasks and given suitable training data. Typically the objective for AutoML technologies relies on some performance metric with respect to a specific machine learning task. One common choice of such a metric is the resulting model accuracy. Complimentary objectives such as sparsity typically are not explicitly established in AutoML frameworks. Sparsity and accuracy are often conflicting goals, and optimum solutions form a Pareto front. Sparsity objective could be integrated by preprocessing pipeline operators or by penalty terms in the objective function which yields a single nondominated solution from the Pareto front and without guarantee that solutions at different positions of Pareto front share the same architectural choices. Thus it might yield sub-optimal solutions. The novel proposed method is based on the AutoML method TPOT and enables automated ML pipelines configuration with sparse input features along the whole Pareto front. It was shown that, indeed, architectures and model configurations vary at different points of the Pareto front for baseline examples from the domain of system security.

Reservoir Neural Computers

Benjamin Paaßen

Bielefeld University, The University of Sydney, Humboldt-University of Berlin, German Research Center for Artificial Intelligence

Abstract

At least since the ground-breaking work of Siegelmann [1], researchers have wondered how neural networks relate to more traditional, discrete models of computation like Turing machines. Interestingly, even simple discrete computations, like regular expressions, are hard for neural networks under conditions of noise [1, 2]. A key reason for this difficulty is that traditional neural network models can not keep memory without interference – every computational step influences all bits of memory [3]. To address this issue, memory-augmented neural networks extend a (recurrent) neural network model with an external memory block where information can remain without interference [3].

Unfortunately, training such models is hard due to the dependency between computation and memory: To optimally write to memory, I need to anticipate which computations I want to support with it in the future. Conversely, which computations I can perform depends on what is already in memory. While memory access behavior can be learned with backpropagation, it often takes hundreds of thousands of training sequences and long training times [4].

In light of these training difficulties, we propose to just not train any recurrent part of the model. More precisely, we propose to use a fixed reservoir as a recurrent neural net, to provide a small amount of supervision for the desired memory access behavior, and to then train all memory access modules as well as the output function via convex optimization – no deep learning required [5, 6]. This way, we can solve some neural computing tasks from only a few training sequences and with much less computation time compared to current neural computers.

References

[1] Siegelmann, H., and Sontag, E. (1995). On the Computational Power of Neural Nets. Journal of Computer and System Sciences, 50(1): 132–150. doi:10.1006/jcss.1995.1013

[2] Hammer, B., and Tiňo, P. (2003). Recurrent Neural Networks with Small Weights Implement Definite Memory Machines. Neural Computation, 15(8): 1897–1929. doi:10.1162/089976603606750 80

[3] Graves, A., Wayne, G., et al. (2016). Hybrid computing using a neural network with dynamic external memory. Nature, 538(7626): 471–476. doi:10.1038/nature20101

[4] Collier, M., and Beel, J. (2018). Implementing Neural Turing Machines. Proceedings of the ICANN 2018: 94–104. doi:10.1007/978-3-030-01424-7_10

[5] Paaßen, B., Schulz, A., Stewart, T., and Hammer, B. (2021). Reservoir Memory Machines as Neural Computers. IEEE Transactions on Neural Networks and Learning Systems. in press. https://arxiv.org/abs/2009.06342

[6] Paaßen, B., Schulz, A., and Hammer, B. (2021). Reservoir Stack Machines. Neurocomputing. accepted. https://arxiv.org/abs/2105.01616

Contrastive Learning for Classifying based on Class Possibility Assignment with Reject Option

Seyedfakhredin Musavishavazi

university of Applied Sciences Mittweida, SICIM, Germany

Abstract

In this contribution we try to introduce a classification model with rejection as an option, inspired by *RSLVQ* [1] and *GLVQ* [2]. As a matter of simplification, in the beginning we restrict the discussion to probability density functions. After establishing the outline we pivot to possibility functions for generalization. To start we define a *likelihood ratio* function and use it to predict a *label* for a given data point. With the help of the *likelihood ratio* function and motivated by Chow [3] an *error-reject trade-off* is proposed which is a function of a *threshold* value t. Later it will be proved that not only such a *threshold* exists but also it can be optimized, regardless of the choice of posterior as a probability or possibility function. The next step is about proposing a *contrastive-learning* function and a *class-wise-decision* rule based on it. The latter is used to define the total loss function. Finally, after modification of the model for integration of rejection we conclude the whole proposal with discussion on the choice of the possibility function as a *posterior*.

References

[1] S. Seo, K. Obermayer. Soft Learning Vector Quantization The Neural Computation Journal, 2003

[2] A. Sato, K. Yamada. Generalized Learning Vector Quantization, Advances in Neural Information Processing Systems, 1995

[3] C.K. Chow. On Optimum Recognition Error and Reject Trade-Off, IEEE Transactions on Information Theory Vol. IT6 NO.1, 1970

The Moment Problem with Applications in Machine Learning

Fabian Hinder and Barbara Hammer

Bielefeld University - Cognitive Interaction Technology (CITEC)
Bielefeld - Germany

Abstract

When it comes to continuous quantities, many approaches in statistics or machine learning focus mainly on estimating (conditional) expectations. However, from a modeling perspective, it is usually more appropriate to consider the properties of the entire distribution, e.g. statistical independence. Bridging this gap is often an important aspect in algorithm development and leads to the question of correctness or finding a suitable representation of the data that allows more direct approaches. In this talk we will present a non-parametric representation of distributions, given by the non-centralized moments, which can be used directly in various estimators and machine learning models to solve a wide range of problems: Two-sample and (conditional) independence tests, feature relevance analysis, bi-clustering/data segmentation and conditional density estimation.

Quantum Computing and LVQ

Alexander Engelsberger

University of Applied Sciences Mittweida, Germany

Abstract

In the past years there have been major achievements in the development of quantum computing hardware. These developments are accompanied by better documentation and simulation of quantum algorithms on classical computers. First supervised quantum machine learning algorithms [1] have been presented.

Algorithms in the context of quantum computing can be divided into three types

- quantum inspired
- quantum hybrid
- quantum native

In this contribution advances in quantum inspired and quantum hybrid algorithms for prototype learning are presented.

Our publications on quantum computing are presented. A quantum inspired Learning Vector Quantizer variant called Qu-GLVQ [2][3], which learns on a manifold inspired by quantum states. And a quantum hybrid variant [4] , which uses a quantum computer to estimate the inner product for the distance calculations.

Another quantum-hybrid approach is presented, that combines the known idea of learning on a spherical manifold with the structure of quantum states during calculations on a quantum processor.

References

[1] Schuld, M., Petruccione, F. (2018). Supervised Learning with Quantum Computers.

[2] Villmann, T., Ravichandran, J., Engelsberger, A., Villmann, A. and Kaden, M.(2020). Quantum-Inspired Learning Vector Quantization for Classification Learning. ESANN

[3] Villmann, T., Ravichandran, J., Engelsberger, A., Villmann, A. and Kaden, M.(2020). Quantum-inspired learning vector quantizers for prototype-based classification. Neural Computing and Applications

[4] Villmann, T. and Engelsberger, A.(2021). Quantum-hybrid Neural Vector Quantization - A Mathematical Approach. ICAISC

ASAP - A Sub-sampling Approach for Preserving Topological Structures

Abolfazl Taghribi

University of Groningen

Abstract

Topological data analysis tools enjoy increasing popularity in a wide range of applications, such as Computer graphics, Image analysis, Machine learning, and Astronomy for extracting information. However, due to computational complexity, processing large numbers of samples of higher dimensionality quickly becomes infeasible. This contribution is two-fold: We present an efficient novel sub-sampling strategy inspired by Coulomb's law to decrease the number of data points in d-dimensional point clouds while preserving its homology. The method is not only capable of reducing the memory and computation time needed for the construction of different types of simplicial complexes but also preserves the size of the voids in d-dimensions, which is crucial e.g. for astronomical applications. Furthermore, we propose a technique to construct a probabilistic description of the border of significant cycles and cavities inside the point cloud. We demonstrate and empirically compare the strategy in several synthetic scenarios and an astronomical particle simulation of a dwarf galaxy for the detection of superbubbles (supernova signatures). [1],[2]

References

[1] A. Taghribi, M. Mastropietro, S. Rijcke, K. Bunte and P. Tino, "ASAP-A Sub-sampling Approach for Preserving Topological Structures." Proceedings of the 28th European Symposium on Artificial Neural Networks (ESANN). Ciaco-i6doc. com, 2020.

[2] A. Taghribi, M. Canducci, M. Mastropietro, S. Rijcke, K. Bunte and P. Tino, "ASAP - A Sub-sampling Approach for Preserving Topological Structures Modeled with Geodesic Topographic Mapping", Neurocomputing, 2021. Available: 10.1016/j.neucom.2021.05.108.

Metric Learning in Federated Learning

Johannes Brinkrolf

Bielefeld University, CITEC, Germany

Abstract

In practice, a large amount of data is produced by distributed devices which is expedited by the internet of things (IoT). These data eventually result in big data that can be vital in uncovering hidden patterns. Many decisions are directly made on edge devices for quickness, economy, or security reason. In this context, federated learning plays a major role, i. e. learning schemes which enable an efficient, possibly privacy preserving integration of the information of local data produced by individual persons towards an integrated model.

Prototype-based methods such as learning vector quantization (LVQ) techniques combine discriminative and generative aspects by representing models in terms of representative locations in the data space which enable an intuitive nearest-neighbor based classification. This fact has already been used in the context of incremental learners for streaming data which might be subject to drift. In this contribution, we demonstrate that this intuitive representation enables a very simple strategy also for federated learning. Here, we will focus on LVQ as particularly robust training method, and its extensions to metric learning schemes in a global as well as in a local manner. We will rely on the fact that LVQ models naturally offer a sufficient statistics of the model, a fact which has already been used in the context of drift-resistant incremental LVQ learning schemes for streaming data, and in privacy preserving versions of LVQ. Based on this observation, we propose a novel method to fuse different LVQ models to enable distributed optimization also in the context of imbalanced classes.

In our experiments we compare federated LVQ with the original version on real-world benchmarks and two different scenarios including imbalanced classes.

GMLVQ based Transfer Learning - Nullspace Transfer Classification Learning

Daniel Staps[1,*],Jensun Ravichandran[1], Sascha Saralajew[2], Marika Kaden[1,*], Michael Biehl[3] and Thomas Villmann[1]

[1] - University of Applied Sciences Mittweida, SICIM, Germany
[3] -Bosch Center for Artifficial Intelligence, Renningen - Germany
[3] - Bernoulli Institute for Mathematics, Computer Science and Arti cial Intelligence, University of Groningen, Groningen - The Netherlands

Abstract

In real world we can find several scenarios where we obtain data from several sources but only a little amount of each source. The training of a valid classifier is at least difficult when you have only little data available. Accordingly, we introduce a siamese-like setting with the training two Generalized Matrix Learning Vector Quantization (GMLVQ) models. One model is learned to map the data for a good source distinction and the other model in parallel to solve the original classification task. Thereby the two linear mappings are connected. The respective null space projection provides a common data representation of the different source data for joint classification learning. We call this setting Transfer GMLVQ (T-GMLVQ). In the presentation, we give some more details of the models and show its potentials.

[*] Supported by ESF: Nachwuchsforschergruppe *Malekita*

Investigating Drift in Hyperspectral Imaging Data

Valerie Vaquet[1], Patrick Menz[2], Udo Seiffert[2] & Barbara Hammer[1]

[1]Machine Learning Group, Bielefeld University, Bielefeld, Germany
[2]Cognitive Processes and Systems, Fraunhofer Institute of Factory Operation and Automation (IFF), Magdeburg, Germany

Abstract

Drift, e.g. the change of the underlying data distribution, is a well known problem that occurs when the sensing device is changed or when a device is aging. For hyperspectral data, this drift is a combination of intensity and wavelength shifts. We propose a novel method to eliminate the shifts between devices and measurement times. We experimentally show that the proposed method performs on par or better in comparison to existing methods [1, 2] on both artificial data, containing only one shift type, and on real world measurements. Besides, assuming bounds on the smoothness of the data are given, we provide a theoretical motivation why our technology can deal with both shifts. Finally, we investigate the method further and find a possible extension to enhance the applicability in real world settings.

References

[1] Melchert, F., Seiffert, U., & Biehl, M. (2015). Polynomial Approximation of Spectral Data in LVQ and Relevance Learning. In Workshop on New Challenges in Neural Computation 2015 (blz. 25-32). (Machine Learning Reports; Vol. 03-2015).

[2] Bouveresse, E. & Massart, D.L. Improvement of the piecewise direct standardizationprocedure for the transfer of nir spectra for multivariate calibration. In Chemom. Intell.Lab. Syst., 32:201-213, 1996

Alpha, Beta and Gamma Balls

Maryam Alipour

University of Applied Sciences Mittweida, SICIM, Germany

Abstract

Surface reconstruction methods have become very important tool to get digital representations of real world existing objects,. In other words, we apply these methods to be able to define the shape that a set of sample points from in the plane. In this presentation, a few of such methods are summarized. After a short introduction to Alpha shapes [1], Gamma shapes [2], the ball-pivoting algorithm [3] and beta balls method, the motivation behind them, their progress and the challenge with limitation that are still on their way are discussed.

References

[1] H. Edelsbrunner, D. G. Kirkpatrick, and R. Seidel. On the Shape of a Set of Point in the Plane. Transaction on Information theory, Vol IT29, No 4., 1983

[2] Y. Sun. Surface Reconstruction using Gamma Shapes. The University of Albama, 2006

[3] F. Bernardini, J. Mittleman, H. Rushmeier, C. Silva, and G. Taubin. The Ball-Pivoting Algorithm for Surface Reconstruction. Transaction on Visualization and Computer Graphics, Vol. 5, No4, 1999

Comparison of Protein Sequence Embeddings to Classify Molecular Functions

Philipp Väth

University of Applied Sciences Würzburg-Schweinfurt,
Department of Computer Science and Business Information Systems, Germany

Abstract

As a result of advances in protein sequencing, for example the UniProt database is seeing the available protein sequences doubled approximately every two years. With more data and increasing computational power, many machine learning tasks have emerged in bioinformatics. One of those tasks is the challenge to find fixed-sized representations of variable length protein sequences, known as representation learning. As more algorithms are available to obtain such representations, in this work we pursue the question of how different algorithms perform in generating general protein sequence representations. Techniques range from traditional Smith-Waterman protein sequence alignment combined with dissimilarity representation of Duin and Pekalska 2011, to state-of-the-art transformer networks and self-supervision. We also take into consideration some models adapted from natural language processing, such as the ProtVec model based on the well known Word2Vec Skip-gram architecture. Our comparison also covers multiple types of neural network architectures such as fully connected-, CNN, LSTM, and transformer models. Finally, we attempt to compare the embeddings in several categories, e.g., semantic tasks, clustering of the embedding space, and computational complexity.

Machine Learning based modeling in industrial applications

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

1 - Cognitive Processes and Systems, Fraunhofer Institute for Factory Operation and Automation IFF Magdeburg, Germany
2 - Compolytics GmbH, Barleben, Germany

Abstract

Industrial applications of machine learning require not only the expected precision, but also in particular high robustness and interpretability in order to achieve a desired acceptance. In addition, unlike numerous academic experiments on carefully selected data sets, there are often limitations in the scope, structure, and availability of training data. Further challenges relate to missing or even inconsistent a priori background knowledge on the application at hand.

This paper illustrates exemplary applications from the domain of processing high-dimensional functional sensor data with machine learning methods in an industrial context. It covers the planning of the measurement campaign, the provision of reference / target data for mathematical modeling, and the interpretation of the results in the context of the underlying application.

Learning classification models from multiple, heterogeneous perspectives

Maximilian Münch[1,2] and Frank-Michael Schleif[1]

1 - University of Applied Sciences Würzburg-Schweinfurt,
Department of Computer Science and Business Information Systems,
D-97074 Würzburg, Germany
2 - University of Groningen, Bernoulli Institute for Mathematics,
Computer Science and Artificial Intelligence,
P.O. Box 407, NL-9700 AK Groningen, The Netherlands

Abstract

Considering a complex problem from different perspectives is often a better way to solve it - this includes both real-world problems and machine learning problems. In the fusion of data from different origins, for example, a particular situation is analysed from the different perspectives of the respective sources. However, these sources can be very dissimilar and in heterogeneous formats. For example, when analysing a complex information system, the data is often simultaneously available as image material, audio recordings, as text, as well as time series, sequences or histograms. This results in two main challenges: (1) Translating non-vectorial data (sequences, histograms, texts) into a vectorial representation to apply ML techniques. (2) The combination of this vectorial data such that all information is integrated in the learning process. If the data is non-vectorial, proximity-based embedding techniques provide a powerful way to transform the non-vectorial data into a vector space [2]. However, standard embeddings lead to the desired fixed-length vector encoding but are costly and have substantial limitations in preserving the original data's full information. As an information preserving alternative, we proposed a complex-valued vector embedding of proximity data [1]. Subsequently, the combination of the different vector spaces can easily be done by concatenating the complex-valued features. This allows suitable machine learning algorithms to use these fixed-length, complex-valued vectors for further processing. Based on our current research, we suggest the complex-valued GMLVQ for this purpose [3]. The cGMLVQ not only enables the use of complex-valued data but it also offers two other significant advantages: (1) After the training phase, we obtain an interpretable model that captures the most significant data points in terms of prototypes. (2) Relevance learning also provides information about which of the perspectives (and also data sources) were particularly relevant during the learning process and which may not need to be considered at all in the future. The proposed approach is evaluated on a variety of benchmarks and shows strong performance compared to traditional techniques in processing non-vectorial data with heterogeneous data format.

References

[1] Münch, M., Straat, M., Biehl, M., Schleif, F.: Complex-valued embeddings of generic proximity data. In: Torsello, A., Rossi, L., Pelillo, M., Biggio, B., Robles-Kelly, A. (eds.) Structural, Syntactic, and Statistical Pattern Recognition - Joint IAPR International Workshops, S+SSPR 2020, Padua, Italy, January 21-22, 2021, Proceedings. Lecture Notes in Computer Science, vol. 12644, pp. 14–23. Springer (2020)

[2] Pekalska, E., Duin, R.P.W.: The Dissimilarity Representation for Pattern Recognition - Foundations and Applications, Series in Machine Perception and Artificial Intelligence, vol. 64. WorldScientific (2005)

[3] Straat, M., Kaden, M., Gay, M., Villmann, T., Lampe, A., Seiffert, U., Biehl, M., Melchert, F.: Learning vector quantization and relevances in complex coefficient space. Neural Comput. Appl. **32**(24), 18085–18099 (2020)

Detecting Hate Speech In Multimodal Memes Using Vision-Language Models

Riza Velioglu

Machine Learning Group, Bielefeld University, Germany

Abstract

Memes on the Internet are often harmless and sometimes amusing. The apparently innocent meme, though, becomes a multimodal form of hate speech when certain kinds of pictures, text, or variations of both are used – a *hateful meme*. The Hateful Memes Challenge [1] is a one-of-a-kind competition that focuses on detecting hate speech in multimodal memes and proposes a new data collection with 10,000+ new examples of multimodal content. We use VisualBERT [2], which is also known as "BERT for vision and language" and Ensemble Learning to boost the performance. In the Hateful Memes Challenge [1], our solution received an AUROC of 0.811 and an accuracy of 0.765 on the challenge test set, placing us **third out of 3,173 participants** [2]. The code is available at GitHub [3]. [3, 4]

References

[1] Kiela, Douwe and Firooz, Hamed and Mohan, Aravind and Goswami, Vedanuj and Singh, Amanpreet and Ringshia, Pratik and Testuggine, Davide (2020).The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. arXiv preprint arXiv:2005.04790.

[2] Li, Liunian Harold and Yatskar, Mark and Yin, Da and Hsieh, Cho-Jui and Chang, Kai-Wei (2019). Visualbert: A simple and performant baseline for vision and language, arXiv preprint arXiv:1908.03557.

[3] Sharma, Piyush and Ding, Nan and Goodman, Sebastian and Soricut, Radu (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning.Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2556–2565.

[4] Singh, Amanpreet and Goswami, Vedanuj and Natarajan, Vivek and Jiang, Yu and Chen, Xinlei and Shah, Meet and Rohrbach, Marcus and Batra, Dhruv and Parikh, Devi. (2020). MMF: A multimodal framework for vision and language research

[1] `https://www.drivendata.org/competitions/70/hateful-memes-phase-2/`

[2] `https://hatefulmemeschallenge.com/#leaderboard`

[3] `https://github.com/rizavelioglu/hateful_memes-hate_detectron`

Intelligent Gait Analysis using Marker Based Motion Capturing System

Danny Möbius[1], Marika Kaden[2*], Daniel Staps[2*] and Thomas Villmann[2]

[1] Institut für Mechatronik, Chemnitz, Germany
[2] University of Applied Sciences Mittweida, SICIM, Germany

Abstract

Marker-based systems can digitally record human movements in detail. Using the digital bio-mechanical human model *Dynamicus*, which was developed by the Institut für Mechatronik (IfM), it is possible to model joint angles and their velocities such accurately that it can be used to improve motion analysis in competitive sports or for ergonomic evaluation of motion sequences. We will present a project between IfM and the Saxon Institute of Computational Intelligence and Machine Learning (SICIM) of the UAS Mittweida about the use of interpretable machine learning techniques to analyze the different phases of the gait. The motion data for training the model is labeled using force plates. We analyze how we could apply our machine learning models directly on new motion data recorded in a different scenario compared to the initial training, more precise on a treadmill. We use the properties of a interpretable model to detect drift, to transfer our model if necessary and give a confidence value for the prediction.

[*] Supported by ESF: Nachwuchsforschergruppe *Malekita*

Predicting elbow movement from electromyography data

Markus Vieth

Machine Learning Group, Bielefeld University, Germany

Abstract

Electromyography gives insight into muscle activity and promises to make prediction of movement with very little latency possible. Most of the existing research tries to predict which of several movements is intended (a classification problem). This project focuses on the prediction of continuous movement of the elbow from four EMG sensors on the upper arm. The goal is an interpretable model that allows understanding of the activity of individual muscles. We further intend to explore how the learned models differ between people and how a model can be transferred to another person.

AI-based Multi Sensor Fusion for Smart Decision Making: A Bi-Functional System for Single Sensor Evaluation in a Classification Task

Feryel Zoghlami

Automation, Maintenance and Factory Integration Infineon Technologies Dresden GmbH Co KG Dresden, Germany

Abstract

This work is part of my research, which is about developing AI based sensor fusion solution for making smart decisions. We focus in this part on developing a smart and interpretable bi-functional AI system, which has to discriminate the combined data regarding predefined classes. Furthermore, the system can evaluate the single source signals used in the classification task. The evaluation here covers each sensor contribution and robustness. More precisely, we train a smart and interpretable prototype-based neural network, which learns automatically to weight the influence of the sensors for the classification decision. Moreover, the prototype-based classifier is equipped with a reject option to measure classification certainty. To validate our approach's efficiency, we refer to different industrial sensor fusion applications.

Prototype selection based on set covering and large margins

Benjamin Paaßen

German Research Center for Artificial Intelligence (DFKI)

Thomas Villmann

University of Applied Sciences Mittweida

September 2021

Abstract

Classification via nearest prototypes is a fast, interpretable, and flexible scheme for classification [5]. The selection of prototypes ought to achieve two goals: Minimizing classification errors *and* representing the classes well. In this paper, we explore two cost functions from the literature which incorporate these goals, namely the large margin nearest neighbor cost function of Weinberger and Saul [7] as well as the prototype selection scheme of Bien and Tibshirani [1]. We highlight similarities and differences of both, thus sharpening our understanding of the prototype selection problem.

Classification via prototypes follows a very simple rule: To any data point, we assign the label of the closest prototype. In order to classify a point x_i correctly, the distance to the closest prototype from the same class d_i^+ must be smaller than the distance to any prototype from another class $d_{i,j}$ [5].

So, how does one select prototypes which achieve this criterion? A trivial method is to treat *any* training data point as a prototype, yielding a 1-nearest neighbor classifier [3]. While surprisingly accurate in many practical applications, this scheme has issues in terms of space and time complexity: To make predictions, we need to store the entire training data set and perform a similarity search across it, which takes at least $\mathcal{O}(\log(n))$ time [2]. By contrast, parametric models generally take only constant memory and time.

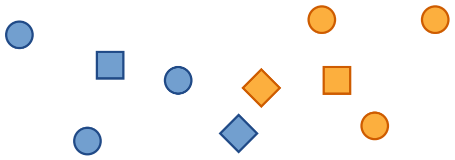


Figure 1: An illustration of prototype selection according to problem 1. Class labels are indicated by color. The optimal solution requires prototypes close to the classification boundaries (diamonds). The most representative prototypes would be the medoids (rectangles).

Accordingly, many methods take the route of reducing the prototype set to a much smaller size $K \ll n$, such that the classification decisions are as similar as possible to the full training set [1]. Formally speaking, let $x_1, \ldots, x_n$ be a set of data points with labels $y_1, \ldots, y_n$, let $d_{i,j}$ be the distance between point i and j, let $\alpha_i \in \{0, 1\}$ be a binary indicator whether point x_i is a prototype (i.e. $\alpha_i = 1$) or not (i.e. $\alpha_i = 0$), and let $d_i^+ = \min_{j:\alpha_j=1, y_i=y_j} d_{i,j}$. Then, we would like to find the smallest number of prototypes such that all classifications are still correct. As an optimization problem, we obtain:

$$\begin{aligned} \min_{\vec{\alpha} \in \{0,1\}^n} \quad & \sum_{i=1}^{n} \alpha_i && (1) \\ \text{s.t.} \quad & d_i^+ < d_{i,j} && \forall j \in N_i, \end{aligned}$$

where N_i is the set $N_i = \{j | \alpha_j = 1, y_i \neq y_j\}$.

This optimization problem has two drawbacks. For one, it is NP-hard [1]. More subtly, though, it promotes selecting prototypes which lie close to the class boundary and, thus, do not represent the data distribution in the class well. Consider the two-class problem in Figure 1. We can classify all points correctly by using only one prototype per class. However, to do so we need to select prototypes close to the class boundary (diamonds). These are relatively atypical examples for their class. The class medoids (rectangles) would be more representative [6] but would misclassify points near the boundary.

1 Two loss functions for prototype selection

Loss functions have since attempted to find a compromise between discrimination and representativeness. As an example, consider the large margin nearest neighbor loss function of Weinberger and Saul [7] which consists of a “push” term and a “pull” term:

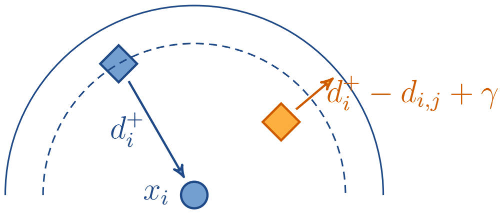


Figure 2: An illustration of the LMNN loss function 2. For each data point x_i, we pull the closest prototype from the same class closer and push prototypes from other classes out of the ball with radius $d_i^+ + \gamma$.

$$\min_{\vec{\alpha} \in \{0,1\}^n} \quad \overbrace{\sum_{i=1}^{n} d_i^+}^{\text{pull}} + \overbrace{\sum_{i=1}^{n} \sum_{j \in N_i} \left[d_i^+ - d_{i,j} + \gamma \right]_+}^{\text{push}}, \tag{2}$$

where $[x]_+ = \max\{0, x\}$ and $\gamma > 0$ is some hyper-parameter called *margin*.

While Weinberger and Saul [7] use this loss function for metric learning, it is quite natural to apply it to prototype selection as well: We are looking for prototypes that represent the data well by minimizing the data to all points in their Voronoi cell (pull term), and which ensure correct classification with a margin of safety γ (push term). The latter interpretation holds because $[d_i^+ - d_{i,j} + \gamma]_+$ is zero if and only if $d_i^+ + \gamma < d_{i,j}$ for all $j \in N_i$, which means that point i is classified correctly. Figure 2 displays a geometric intuition in line with the original paper [7]. Data point x_i is located in the center of a ball with radius $d_i^+ + \gamma$. The pull term of the loss encourages the closest prototype from the same class to move closer to x_i, such that this ball shrinks. The push term encourages prototypes from other classes to move out of the ball.

Bien and Tibshirani [1] propose an alternative loss function for prototype selection based on set covering. In particular, they argue that prototypes should be selected such that they cover as many data points as possible in balls of radius ϵ and such that as few data points of other classes intrude the balls. More formally, we obtain the following loss function.

$$\min_{\vec{\alpha} \in \{0,1\}^n} \quad \sum_{i=1}^{n} h_\epsilon(d_i^+) + \sum_{i=1}^{n} \sum_{j \in N_i} 1 - h_\epsilon(d_{i,j}) + \lambda \cdot \sum_{i=1}^{n} \alpha_i, \tag{3}$$

where $h_\epsilon(x)$ is the Heaviside function with input $x - \epsilon$, i.e. $h_\epsilon(x) = 1$ if $x > \epsilon$ and $h_\epsilon(x) = 0$ otherwise [1]. Note that ϵ and λ are hyper-parameters of this

[1] Note that this formulation is *not* strictly equivalent to Bien and Tibshirani [1]. Im-

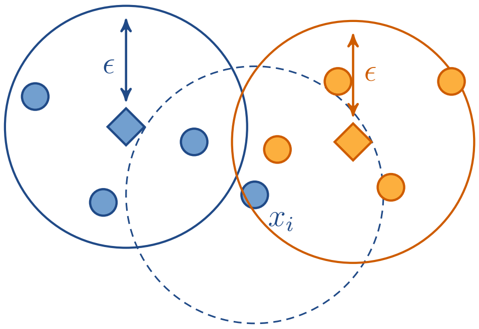


Figure 3: An illustration of the set cover loss function 3. We punish each data point that is not covered by an ϵ-ball around a prototype (diamonds) of the same class and each data point that intrudes in an ϵ-ball around a prototype from a different class.

method which regulate the size of the balls and the cost of adding a new prototype, respectively.

Figure 3 shows a graphical illustration of this cost function. The prototypes (diamonds) cover almost all points of the same class with ϵ-balls. The only exception is point x_i, which lies outside the ϵ-ball of its class prototype but intrudes the ϵ-ball of the other prototype. Accordingly, the value of the cost function would be $2+2\lambda$, punishing x_i with a value of 2 and the presence of two prototypes with 2λ.

2 Similarities and Differences

Our notation of the loss functions 2 and 3 is already indicating key similarities: both loss functions contain a “pull” and a “push” term. The former pulls prototypes close to points of the same class in their Voronoi cell, the latter pushes prototypes away from points with different labels. We also note that, in both cases, the terms are purely distance-based: the pull term minimizes some expression of d_i^+, the push term increases some expression of $d_{i,j}$. Still, there are also notable differences which influence the behavior of both loss functions. We now cover these differences one by one, discussing options to alleviate them.

First, the set cover loss 3 includes a third term which encourages sparsity in $\vec{\alpha}$. However, this term could easily be added to the LMNN loss 2 without

portantly, we only permit points to be covered by prototypes of their own class. Refer to Appendix A for more details.

changing the fundamental behavior.

Second, the LMNN loss 2 is intended to be read “from the perspective of the data”, i.e. draw a ball with radius $d_i^+ + \gamma$ around each point x_i and punish both the size of the ball as well as intrusions by prototypes with different labels. By contrast, the set cover loss 3 is designed “from the perspective of the prototypes”, i.e. draw a ball with radius ϵ around each prototype and try to cover all points with these ball and punish intrusions by points with different labels. However, this change in perspective is easily accounted for, as illustrated in Figure 3. We can draw a ball with radius ϵ around each point x_i and punish if no prototype of the same label is inside and or a prototype with a different label is inside. This change in perspective is also reflected in our notation of loss 3.

Third, loss 2 is continuous whereas loss 3 is discrete due to the Heaviside function. While this difference does not fundamentally change the behavior of the push term, the pull term behaves quite different. Consider an outlier data point x_i. For this point, d_i^+ will naturally be very large. Accordingly, the LMNN loss 2 ha‘s a large incentive to move the closest prototype toward the outlier. However, for the set cover loss 3 the size of the incentive is constant, no matter how far the outlier is away. This is a more robust behavior which may be preferable in practice. Still, there is room for a compromise between both losses: In particular, we can replace the Heaviside functions in loss 3 with sigmoids. The ’flatter’ our sigmoid, the more linear it will behave and the more similar the pull behavior will be to loss 2.

Finally, the push term in loss 2 punishes misclassifications whereas the push term in loss 3 only punishes intrusions in ϵ-balls. As long as $d_i^+ < \epsilon$, this does not matter because the behavior is essentially equivalent: the push term will ensure that $d_{i,j}$ becomes larger ϵ, which in turn is larger d_i^+, such that point x_i will be classified correctly if the push term is zero. However, if $d_i^+ \geq \epsilon$, the push term can become zero even though x_i is still misclassified. Consider Figure 4. In this figure, the push loss is zero, because no point intrudes the ϵ-ball of a prototype from another class, yet both points are misclassified.

3 Conclusion

The comparison of the LMNN loss 2 and the set cover loss 3 reveals advantages of both. The LMNN loss is more stringent in preventing misclassifications, whereas the set cover loss is more robust with respect to outliers. Accordingly, an interesting direction for future work would be to combine concepts from both losses, namely the pull term from the set cover loss and

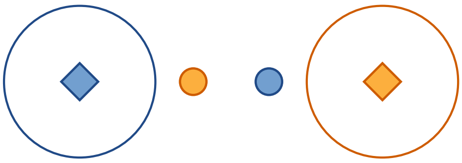


Figure 4: A model with two prototypes that would misclassify both points in the center but where the "push" term of loss 3 is nonetheless zero.

the push term from the LMNN loss. It will, however, be a challenge to achieve a solver for such a partially discrete and partially continuous loss. Perhaps, techniques from median generalized vector quantization [4] could be adapted to that end.

Besides the opportunity for novel machine learning methods, though, the consideration of both losses has granted us deeper insight into the design of losses for prototype selection which may support future studies into the theory of prototypes as well as the interpretation of prototype-based models.

References

[1] Jacob Bien and Robert Tibshirani. "Prototype selection for interpretable classification". In: *The Annals of Applied Statistics* 5.4 (2011), pp. 2403–2424. DOI: `10.1214/11-AOAS495`.

[2] Edgar Chávez et al. "Searching in Metric Spaces". In: *ACM Computing Surveys* 33.3 (2001), pp. 273–321. DOI: `10.1145/502807.502808`.

[3] T. Cover and P. Hart. "Nearest neighbor pattern classification". In: *IEEE Transactions on Information Theory* 13.1 (1967), pp. 21–27. DOI: `10.1109/TIT.1967.1053964`.

[4] David Nebel et al. "Median variants of learning vector quantization for learning of dissimilarity data". In: *Neurocomputing* 169 (2015), pp. 295–305. DOI: `10.1016/j.neucom.2014.12.096`.

[5] David Nova and Pablo A. Estévez. "A review of learning vector quantization classifiers". In: *Neural Computing and Applications* 25.3 (2014), pp. 511–524. DOI: `10.1007/s00521-013-1535-3`.

[6] Hae-Sang Park and Chi-Hyuck Jun. "A simple and fast algorithm for K-medoids clustering". In: *Expert Systems with Applications* 36.2, Part 2 (2009), pp. 3336–3341. DOI: `10.1016/j.eswa.2008.01.039`.

[7] Kilian Weinberger and Lawrence Saul. “Distance Metric Learning for Large Margin Nearest Neighbor Classification”. In: *Journal of Machine Learning Research* 10 (2009), pp. 207–244. URL: `http://www.jmlr.org/papers/v10/weinberger09a.html`.

A Re-writing the set-cover problem

As noted above, loss 3 is not immediately equivalent to the set cover problem of Bien and Tibshirani [1]. In this appendix we derive loss 3 from their formulation. We start with problem (3) from their work:

$$\begin{aligned} \min_{\alpha_i^l, \xi_i, \eta_i} \quad & \sum_{i=1}^n \xi_i + \sum_{i=1}^n \eta_i + \lambda \cdot \sum_{i=1}^n \sum_{l=1}^L \alpha_i^l & & (4) \\ \text{s.t.} \quad & \sum_{j:d_{i,j} \leq \epsilon} \alpha_j^{y_i} \geq 1 - \xi_i & \forall i \qquad & (5) \\ & \sum_{\substack{j:d_{i,j} \leq \epsilon \\ y_i \neq l}} \alpha_j^l \leq \eta_i & \forall i \qquad & (6) \\ & \alpha_i^l \in \{0,1\}, \xi_i \geq 0, \eta_i \geq 0 & \forall i, l, \qquad & (7) \end{aligned}$$

where l iterates over all classes. Importantly, this problem permits that any point x_i can be a prototype for *any* class. Wile this appears counter-intuitive, it can be beneficial if a point x_i is surrounded by points of a single other class $l \neq y_i$. Then, setting $\alpha_i^l = 1$ incurs a loss because x_i itself intrudes in its own ϵ-ball, but all surrounding points are covered as well. We will simplify this structure later on. First, however, we turn our attention to side constraint 5. Since our objective function aims to minimize ξ_i, it is helpful to minimize this constraint as a lower bound for ϵ_i.

$$\xi_i \geq 1 - \sum_{j:d_{i,j} \leq \epsilon} \alpha_j^{y_i}.$$

The other lower bound is $\xi_i \geq 0$. Since all α_j^l are binary variables, the optimal ϵ_i can only obtain two values: One, if the sum $\sum_{j:d_{i,j} \leq \epsilon} \alpha_j^{y_i}$ is zero, or zero, if it is 1 or larger. We can simplify both cases by introducing the auxiliary variable $\tilde{d}_i^+ := \min_{j:\alpha_j^{y_i}=1} d_{i,j}$. Then, the sum is zero if $\tilde{d}_i^+ > \epsilon$. Otherwise, the sum is at least 1. Accordingly, we can re-write $\xi_i = h_\epsilon(\tilde{d}_i^+)$, where h_ϵ is the Heaviside function with offset ϵ, i.e. $h_\epsilon(x) = 1$ if $x > \epsilon$ and $h_\epsilon(x) = 0$ otherwise.

Next, we inspect side constraint 6. As the objective function encourages us to minimize η_i, an optimal value η_i will always exactly obtain the value $\sum_{\substack{j:d_{i,j}\leq\epsilon \\ y_i\neq l}} \alpha_j^l$. Let us now define $\tilde{N}_i = \{j | \exists l \neq y_i : \alpha_j^l = 1\}$, i.e. the set of all points x_j which are prototypes of other classes. With this definition, we obtain:

$$\eta_i = \sum_{\substack{j:d_{i,j}\leq\epsilon \\ y_i\neq l}} \alpha_j^l = \sum_{j\in\tilde{N}_i} 1 - h_\epsilon(d_{i,j}).$$

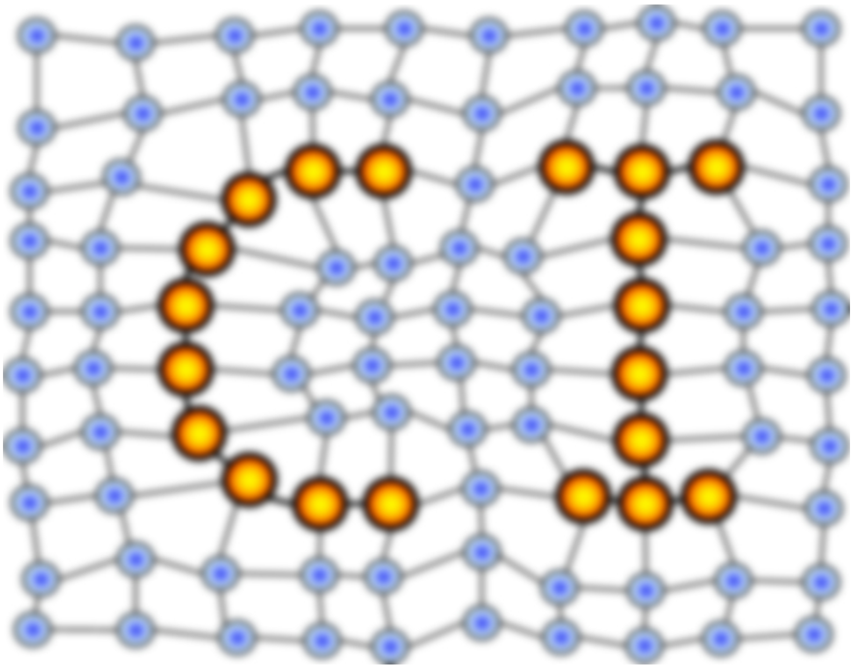
Accordingly, our overall optimization problem becomes

$$\min_{\alpha_i^l} \quad \sum_{i=1}^n h_\epsilon(\tilde{d}_i^+) + \sum_{i=1}^n \sum_{j\in\tilde{N}_i} 1 - h_\epsilon(d_{i,j}) + \lambda \cdot \sum_{i=1}^n \sum_{l=1}^L \alpha_i^l.$$

Note that this form is already strikingly similar to loss 3. To eliminate the remaining difference, we need to impose that $\alpha_i^l = 0$ if $l \neq y_i$, i.e. data point x_i is only allowed to be a prototype for its own class. We believe that this is only a minor restriction in practice and facilitates interpretability. We now define $\alpha_i := \alpha_i^{y_i}$. Accordingly, $\tilde{d}_i^+$ becomes equivalent to d_i^+, $\tilde{N}_i$ becomes equivalent to N_i, and $\sum_{i=1}^n \sum_{l=1}^L \alpha_i^l = \sum_{i=1}^n \alpha_i^{y_i} = \sum_{i=1}^n \alpha_i$. Overall, the objective function collapses exactly to 3, as claimed.

MACHINE LEARNING REPORTS

Report 03/2021



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 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.