ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

Adaptive multi-modal positive semi-definite and indefinite kernel fusion for binary classification

Maximilian Münch[1,2], Christoph Raab[1], Simon Heilig[1], Manuel Röder[1] and Frank-Michael Schleif[1] *

1- Center for Artificial Intelligence and Robotics (CAIRO), University of Applied Sciences Würzburg-Schweinfurt, Würzburg, Germany

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

Abstract. Data and information are nowadays frequently available in multiple modalities like different sensor signals, textual descriptions, graph structures, and other formats. The maximum information from these heterogeneous representations can be obtained by fusing the various modalities by specific embeddings or proximity measures. Current approaches are widely limited in the fusion model and the applied measures, especially when the given data is non-vectorial. We propose a model to learn the spectral properties of the different inner product representations in a joined optimization problem. The approach is evaluated on various multi-modal data and compared to modern multiple-kernel learning and baseline techniques.

1 Introduction

Modern data analysis has become increasingly challenging and the expectations on machine learning models are higher than ever: These days, data is no longer given in just a single format, but rather simultaneously in multiple different formats that are not always in vectorial form [2]. For both, vectorial and non-vectorial input data, kernel methods have proven to be highly efficient and robust [16, 3]. Due to its great results in multi-modal data analysis, so-called *Multiple Kernel Learning* (MKL) has become very popular [1]. Despite their impressive results, MKL methods are still widely limited by mathematical constraints of the models, such as the kernel function's positive definiteness (pd).

Here, we present a technique that exploits the spectral properties of multiple kernels to learn a new representation of the data as a single information-rich kernel over multiple modalities. We recap main concepts of Multiple Kernel Learning and the particularities of non-positive semi-definite (non-psd) kernel functions. Subsequently, we outline our novel approach of kernel fusion and evaluate our approach on a variety of benchmark data sets from the MKL domain. We conclude with a detailed discussion of the results and an outlook on further research.

*MM and MR are supported by the Bavarian HighTech agenda and the Würzburg Center for Artificial Intelligence and Robotics (CAIRO). Additionally, we thank Dr. Benjamin Paaßen for the invaluable discussions about this research topic during a fantastic boating trip.

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

2 Learning from multiple indefinite kernel functions

In machine learning, information is now often spread across different heterogeneous formats and classical techniques are insufficient [17, 12]. Frequently, deep learning and embedding techniques can be employed to generate vectorial representations, but they require huge amounts of training data, dedicated deep learning models and have extensive computational costs [17]. Multiple representations can also be addressed by MKL models, where each kernel was derived by a different similarity measure or from different input sources (text, video, audio data, or other) [1]. MKL aims to use various base kernel functions and capture their most important information in order to obtain one single information-rich kernel. The most prominent MKL techniques are *SimpleMKL* [14] and *EasyMKL* [1] which learn the weights of a convex combination of kernels. The underlying kernel functions are often expected to be Mercer kernels and need well-chosen meta-parameters to ensure convexity and convergence guarantees of the kernel consuming machine learning methods [9]. In general, non-psd kernels arise much more frequently than typically assumed and can already occur due to normalization procedures or careless parameter settings [5]. Several correction and adjustment procedures were proposed (eigenspectrum correction, proxy matrix learning or dedicated models for indefinite kernels) [15]. We consider a finite collection of objects $\mathbf{X} = \{\mathbf{x}_i\}$, $i = 1, \ldots, N$ in some (implicit) input space $\mathcal{X}$ and one or multiple similarity functions $k_m(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$ with $m = 1, \ldots, M$ to compare the input data objects. The $\langle \cdot, \cdot \rangle$ can be any symmetric similarity function. In case of Mercer kernels this could be the Euclidean inner product or other types of kernels, but also domain-specific non-psd similarities as alignment functions for sequential data and alike [15].

A kernel matrix $K_m \in \mathbb{R}^{N \times N}$ is obtained for each similarity function by evaluating all pairwise similarities on $\mathbf{X}$. In order to modify K_m to become psd (we denote $\tilde{K_m}$ as the modified psd-version of K_m), the eigenspectrum of K_m can be adapted. This is achieved by an eigendecomposition $K_m = Q_m \Lambda_m Q_m^T$, with Λ_m containing the eigenvalues and Q_m the corresponding eigenvectors of K_m and modifications of Λ_m (like clip, flip, shift, square) to ensure that $\tilde{\lambda}_m = \tilde{\Lambda}_{m[ii]} \geqslant 0$ for $i = 1, \ldots, N$, which eventually results in a positive semi-definite $\tilde{K}_m = Q_m \tilde{\Lambda}_m Q_m^T$ [15]. Instead of modifying the eigenspectrum of K_m directly, the authors in [10] proposed to learn a psd proxy matrix with maximum alignment to K_m. In general this leads to a clip strategy, where all negative eigenvalues are removed (see also [15, 11]). By now, only very few MKL approaches addressed indefinite kernels. The authors in [6] suggested a primal formulation following a *SimpleMKL* style, but without requiring psd constraints. The model provides a binary classifier for multiple input kernels and is optimized by gradient descent. However, the approach does not exploit the available information in the multiple modalities since it sticks with an averaging strategy. The approach in [18] can only be applied to vectorial input data, but not for general similarity measures. In contrast to other previous work in this area, our proposed method modifies the spectral properties of the multi-modal data in a joined adaptive MKL approach.

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

3 Adaptive spectral properties for kernel fusion

Multiple Kernel Learning with non-psd matrices is challenging. We propose a new strategy to learn one strong kernel matrix from a variety of weak - potentially indefinite - kernels over a support vector classifier, in the following referred to as **A***daptive* **S***ubspace* **K***ernel* **F***usion - Support Vector Machine* (ASKF-SVM). The dual of the support vector machine (e.g. [16]) is given as:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\alpha^T Y K Y \alpha - \mathbf{1}^T\alpha \\ \text{subject to} \quad & 0 \leqslant \alpha_i \leqslant c \quad \text{and} \quad \sum_i y_i\alpha_i = 0, \qquad i = 1,\dots,N; \quad C \in \mathbb{R}, \end{aligned} \tag{1}$$

with $y_i = \{-1, 1\}, Y = [y_1, \dots, y_N]^\top$ as the labels and the label vector. The vector α contains the weights for the support vectors, C is a regularization parameter and K a psd kernel matrix. Our optimization approach is derived from Eq. (1), with the additional objective to learn a psd kernel $\tilde{K}$ over multiple modalities or kernel functions. Again, we assume a set of M proximity functions for pairwise comparisons of the input data yielding a collection of kernel matrices $\mathbf{K} = \{K_1, \dots, K_M\}$, all of them of size $N \times N$. To determine the relevant spectral properties, we apply an eigen-decomposition $K_m = Q_m \Lambda_m Q_m^T$ to obtain the eigenvalues Λ_m and eigenvectors Q_m for each K_m. Let $\mathbf{\Lambda} = \{\Lambda_1, \dots, \Lambda_M\}$ be the collection of all eigenvalues Λ_m for all K_m and $\mathbf{Q} = \{Q_1, \dots, Q_M\}$ the collection of their respective eigenvectors Q_m. Now we select those N eigenvectors from $\mathbf{Q}$ over all Q_m whose corresponding eigenvalues have the greatest importance over all eigenspectra, i.e. the N eigenvalues from $\mathbf{\Lambda}$ with the maximum absolute value. We refer to this set as $\hat{Q}$ and optimize the spectral properties to obtain a discriminative and psd weighting of the eigenvectors. Therefore, we adopt the optimization problem from Eq.(1) and extend it by some additional stress factors and constraints:

$$\begin{aligned} \min_{\alpha,\tilde{\lambda}} \quad & \frac{1}{2}\alpha^T Y \underbrace{\hat{Q}\tilde{\Lambda}\hat{Q}^T}_{\tilde{K}} Y\alpha - \mathbf{1}^T\alpha - \underbrace{\beta \cdot \mathbf{1}^T\tilde{\Lambda}\mathbf{1}}_{R_1} + \underbrace{\gamma \cdot ||\hat{Q}\Lambda\hat{Q}^T - \hat{Q}\tilde{\Lambda}\hat{Q}^T||_F}_{R_2} \\ \text{subject to} \quad & 0 \leqslant \alpha_i \leqslant C \quad \text{and} \quad \sum_i y_i\alpha_i = 0, \qquad i = 1,\dots,n; \quad c \in \mathbb{R} \\ & \underbrace{\tilde{\lambda}_i \geqslant 0}_{C_1} \quad \text{and} \quad \underbrace{\sum_{i=1} |\tilde{\lambda}_i| \leqslant \delta \cdot \sum_{i=1} |\lambda_i|}_{C_2} \qquad \forall \lambda_i \in [\Lambda]_{ii}, \end{aligned}$$

with α and $\tilde{\lambda}$ as optimization variables, $\mathbf{1}$ as vectors of ones, and β, γ and δ as hyperparameters. Compared to Eq. (1), we replace K by a new kernel matrix $\tilde{K} = \hat{Q}\tilde{\Lambda}\hat{Q}^T$ containing the eigenvalues we want to learn. Furthermore, we introduce two additional regularization factors R_1 and R_2 in to the objective as well as two more constraints C_1 and C_2. R_1 is the sum of the new eigenvalues and prevents the optimization solver from automatically setting all eigenvalues to 0. R_2 keeps all values of the new kernel matrix from deviating too far from

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

dataset	M	N	ι	λ_{min}	λ_{max}
FlowCyto	4	612	0.09	-19.73	152.49
PD	96	83	0.2	-169.46	173.94
NR-AR	10	8164	0.12	-4317.47	5963.99
NR-AhR	10	9357	0.12	-4883.85	6766.25
NR-ER	10	7693	0.12	-4106.23	5662.21
SR-ATAD5	10	9086	0.12	-4804.69	6627.91
SR-MMP	10	7316	0.12	-3949.7	5411.08

Table 1: Properties of the benchmark data sets - details in the text.

the values of the old kernel matrix. The impact of the two terms R_1 and R_2 are controlled by the scalars β and γ. The additional constraints C_1 and C_2 guarantee that the new eigenvalues λ_i are neither too small nor too large. C_1 is the guarantee for learning only new eigenvalues with $\lambda_i \geqslant 0$, implying that $\tilde{K}$ must be psd in any case. C_2 defines the upper bound for λ_i as the sum of all new eigenvalues that cannot be greater than a multiple of the original eigenvalues.

The optimization yields optimized vectors α and λ to reassemble the kernel matrix $\tilde{K} = \hat{Q}\tilde{\Lambda}\hat{Q}^T$ and to derive the decision function.

4 Experiments

In this section, we evaluate the performance of our approach against established methods on a variety of benchmark data sets for Multiple Kernel Learning tasks.

4.1 Datasets

Benchmark data and their spectral properties are detailed in Table 1. All data sets used in our experimental setup include M similarity matrices of size $N \times N$ according to the applied similarity or kernel functions.[1] The degree of indefiniteness is quantified by ι, where $\iota = 0$ indicates a psd matrix and $\iota = 1$ a negative semi-definite matrix. The **FlowCyto** data set[2] is based on 612 FL3-A DNA flow cytometer histograms from breast cancer tissues in 256 resolution, divided into two classes for our binary classification setup.[3] The **Presence Detection** (PD) data set[2] analyzes the occupancy of lecture halls based on multiple wireless Bluetooth Low Energy signals [12]. The **Tox21** challenge[4] was the computational analysis about toxic effects of substances on body regions such as stress response (SR) or effects on nuclear receptors (NR). We used **NR-AhR**, **NR-AR**, **NR-ER**, **SR-MMP** out of the original 12 assays of the challenge.

[1] We normalized each of those matrices before the evaluation by $\hat{k}(x,z) = \frac{k(x,z)}{\sqrt{k(x,x)\cdot k(z,z)}}$.

[2] The proximity matrices are given as dissimilarities and converted to similarities by double centering [13]: $\mathbf{S} = -\mathbf{JDJ}/2$ with $\mathbf{J} = (\mathbf{I} - \mathbf{1}\mathbf{1}^\top/N)$, identity matrix $\mathbf{I}$ and vector of ones $\mathbf{1}$.

[3] `http://rduin.nl/prtools.html`

[4] `https://tox21.gov/overview/`

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

Data	NN	EasyMKL	AverageMKL	ASKF-SVM
FlowCyto	0.60 ± 0.07	0.73 ± 0.05	0.72 ± 0.05	**0.80 ± 0.03**
PD	0.80 ± 0.15	0.86 ± 0.12	0.82 ± 0.19	**1.00 ± 0.00**
NR-AR	0.95 ± 0.01	0.96 ± 0.02	0.95 ± 0.02	**0.97 ± 0.01**
NR-AhR	0.88 ± 0.01	0.89 ± 0.04	0.88 ± 0.02	**0.90 ± 0.02**
NR-ER	0.87 ± 0.03	0.86 ± 0.02	0.86 ± 0.02	**0.87 ± 0.01**
SR-ATAD5	0.96 ± 0.01	**0.97 ± 0.01**	**0.97 ± 0.01**	**0.97 ± 0.00**
SR-MMP	0.82 ± 0.03	0.83 ± 0.03	0.82 ± 0.01	**0.84 ± 0.01**

Table 2: Classification results on multi-modal data sets comparing ASKF-SVM to a NN baseline classifier and other models from the MKL domain.

4.2 Evaluation & Results

In the experimental setup, we evaluate our ASKF-SVM against other baseline classifiers and methods from the MKL domain.

Nearest neighbor (NN) is used as a baseline model. The similarity matrices are evaluated individually and the respective predictions are averaged. This approach is computationally expensive as all matrix values have to be stored as total in RAM or recalculated for new points. Besides the baseline nearest neighbor, there exist various advanced variants as in [4].

EasyMKL & AverageMKL are two methods from the MKL domain and implemented in MKLpy [8]. Both models are directly combined with an SVM classifier within the MKLpy-framework.

Adaptive Subspace Kernel Fusion-SVM (ASKF-SVM: We created our solver based on the optimization problem of Eq. (3) using the GENO-project framework [7] to extend the classical SVM formulation.

Experiments are done in a five-fold cross-validation with hold-out test set. We tested and trained all classification models on the same training and test splits and optimized the model-specific parameters via grid search. Mean accuracy and standard deviation of classification models are shown in results table 2.

The NN classifier performed in general slightly worse than the other classification models in the **Tox21** data. Only for **FlowCyto** and **Presence-Detection**, the performance of EasyMKL and AverageMKL as well as ASKF-SVM was considerably superior. In general, EasyMKL and AverageMKL performed reasonably well and sometimes slightly better than the baseline NN. Note that EasyMKL and AverageMKL are valid only for psd kernels. The best results in the benchmark are obtained by the proposed kernel fusion technique ASKF-SVM. The ASKF-SVM formulation permits to adapt the eigenspectrum to become psd, while not being limited to a clip approach. It also allows a mixture of common correction methods to get adapted such that the classification task can be taken into account. This flexible weighting enables the method to create a new positive semi-definite kernel with considerable alignment to the old multi-modal kernels and good performance. However, a current limitation of the proposal is the computational complexity.

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

5 Conclusions

In this paper, we presented *Adaptive Subspace Kernel Fusion - SVM*, allowing the fusion of data from different sources or measures for binary classification. It learns how to combine the spectral properties of several kernels in a joined optimization problem. The proposed method performed competitive and in some cases significantly better compared to other MKL methods. In future work, we will extend the approach to multi-class problems and address complexity issues, by low-rank approximation techniques [5], which are promising in first tests.

References

[1] Fabio Aiolli and Michele Donini. EasyMKL: a scalable multiple kernel learning algorithm. *Neurocomputing*, 169:215–224, 2015.

[2] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE TPAMI*, 41(2):423–443, 2019.

[3] Yihua Chen, Eric K. Garcia, Maya R. Gupta, Ali Rahimi, and Luca Cazzanti. Similarity-based classification: Concepts and algorithms. *J. Mach. Learn. Res.*, 10:747–776, 2009.

[4] Jan Philip Göpfert, Heiko Wersing, and Barbara Hammer. Interpretable locally adaptive nearest neighbors. *Neurocomputing*, 470:344–351, 2022.

[5] Simon Heilig, Maximilian Muench, and Frank-Michael Schleif. Memory efficient kernel approximation for non-stationary and indefinite kernels. In *IJCNN*. IEEE, 2022.

[6] M. Kowalski, M. Szafranski, and L. Ralaivola. Multiple indefinite kernel learning with mixed norm regularization. In *ICML*, volume 382, pages 545–552. ACM, 2009.

[7] Sören Laue, Matthias Mitterreiter, and Joachim Giesen. GENO – GENeric Optimization for classical machine learning. In *Advances in NIPS (NeurIPS)*. 2019.

[8] I. Lauriola and F. Aiolli. Mklpy: a python-based framework for multiple kernel learning. *CoRR*, abs/2007.09982, 2020.

[9] G. Loosli, S. Canu, and C. S. Ong. Learning SVM in Krein Spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(6):1204–1216, June 2016.

[10] Ronny Luss and Alexandre d'Aspremont. Support vector machine classification with indefinite kernels. *Math. Program. Comput.*, 1(2-3):97–118, 2009.

[11] S Mehrkanoon, X. Huang, and J. A. K. Suykens. Indefinite kernel spectral learning. *Pattern Recognit.*, 78:144–153, 2018.

[12] M. Münch, K. Huffstadt, and F.-M. Schleif. Towards a device-free passive presence detection system with bluetooth low energy beacons. In *27th Europ. Symp. on Artif. Neur. Netw., ESANN 2019, Bruges, Belgium, 2019*, 2019.

[13] E. Pekalska and R. P. W. Duin. *The Dissimilarity Representation for Pattern Recognition - Foundations and Applications*. WorldScientific, 2005.

[14] Alain Rakotomamonjy, Francis R. Bach, Stéphane Canu, and Yves Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9(83):2491–2521, 2008.

[15] Frank-Michael Schleif and Peter Tiňo. Indefinite proximity learning: A review. *Neural Comput.*, 27(10):2039–2096, 2015.

[16] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[17] Philipp Väth, Maximilian Münch, Christoph Raab, and F.-M. Schleif. Proval: A framework for comparison of protein sequence embeddings. *Journal of Computational Mathematics and Data Science*, page 100044, 2022.

[18] Hui Xue, Yu Song, and Haiming Xu. Multiple indefinite kernel learning for feature selection. *Knowl. Based Syst.*, 191:105272, 2020.