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/.

Efficient classification learning of biochemical structured data by means of relevance weighting for sensoric response features

Katrin Sophie Bohnsack, Marika Kaden, Julius Voigt and Thomas Villmann *

University of Applied Sciences Mittweida,
Saxon Institute for Computational Intelligence and Machine Learning
Mittweida - Germany

Abstract. We present an approach for generating vectorial representations of graphs for machine learning applications based on a sensoric response principle and multiple graph kernels. The sensor perspective reduces the graph kernel computations significantly. Thus, multiple kernel (relevance) learning can be realized using the interpretable generalized matrix learning vector quantization (GMLVQ) classifier. Results obtained in small molecule classification serve as proof of concept.

1 Introduction

Data comparison forms the backbone of machine learning and its applications in chem- and bioinformatics. Both fields require the handling (and thus the comparison) of structured data in the form of graphs, e.g. structural formulas or proteins. Frequently, graph-derived feature vectors, so-called topological descriptors [1] are compared. Instead, a direct processing of structures can be achieved by use of graph kernels [2], where a respective feature map is only considered implicitly.

The particular choice for one or another topological descriptor or graph kernel comes with a feature bias, i.e., narrowing the model's view to certain graph properties while disregarding others. This necessitates either great computational effort during model selection or strategies like multiple kernel learning [3] for combining different kernel matrices in a single learning procedure. As powerful as kernels can be, they generally limit the choice of the classifier to Support Vector Machines (SVMs) [4] or models designed to handle proximity data such as median and relational variants of Learning Vector Quantization (LVQ) [5].

Here, we present an approach inspired by [6, 7] that maps graph data to a proximity space by a *sensoric response principle* (SRP) [8, 9] based on different graph comparison strategies allowing for *relevance learning* of these by Generalized Matrix LVQ (GMLVQ) as a prominent interpretable classifier model [10, 11]. This SRP, significantly reduces the number of kernel computations and, hence, makes this method practicable also for huge data sets.

2 Background

The following two subsections provide primers on kernels for structured data and instances of interpretable machine learning models for relevance learning.

*K.S.B and M.K. are supported by a grant of the European Social Fund (ESF).

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/.

Yet, instead of graph kernels, any other topological descriptors can be used in the SRP-approach explained later.

2.1 Graph comparison by kernels

Let $\mathcal{G}$ be a non-empty set of data points and $\kappa : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ be a function. Then κ is a kernel on $\mathcal{G}$ if there is a Hilbert space $\mathcal{H}_\kappa$ and a feature map $\phi : \mathcal{G} \to \mathcal{H}_k$ such that $\kappa(x, y) = \langle \phi(x), \phi(y) \rangle$ for $x, y \in \mathcal{G}$, where $\langle \cdot, \cdot \rangle$ denotes the inner product of $\mathcal{H}_k$. Such a feature map exists iff the function κ is psd and symmetric.

Kernels on structured data such as graphs are usually instances of so-called convolution kernels [12]. The concept bases on substructure decomposition, such that a graph kernel results from evaluating combinations of base kernel functions defined on *parts*. The validity of this approach follows directly from the closure properties of positive definite functions. Obviously, kernels may be designed by choosing $\mathcal{H}$ and ϕ, and simply evaluating $\langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$. This, however, requires operations in $\mathcal{H}$, which might be computationally demanding such that efficient calculations of $\kappa(x, y)$ are aspired instead (kernel trick). Every real kernel determines a distance between structures x and y by $d_\kappa(x, y) = \sqrt{\kappa(x, x) - 2\kappa(x, y) + \kappa(y, y)}$.

2.2 Classification by Learning Vector Quantization

Generalized Learning Vector Quantization (GLVQ) as introduced by [13] supposes data vectors $\boldsymbol{x} \in \mathcal{X} = \{ \boldsymbol{x}_i \}_{i=1}^{|\mathcal{X}|} \subseteq \mathbb{R}^n$ with class labels $c(\boldsymbol{x}) \in \mathcal{C} = \{ 1, \ldots, C \}$ for training. Further, trainable prototype vectors $\boldsymbol{w} \in \mathcal{W} = \{ \boldsymbol{w}_j \}_{j=1}^{|\mathcal{W}|} \subset \mathbb{R}^n$ with class labels $c(\boldsymbol{w}_j) \in \mathcal{C}$ are required such that each class of $\mathcal{C}$ is represented by at least one prototype. GLVQ aims at distributing the prototype vectors such that the class label of any new input $\boldsymbol{x}$ can be inferred by means of the nearest prototype principle given by $c(\boldsymbol{w}_{s(\boldsymbol{x})})$ where $s(\boldsymbol{x}) = \operatorname{argmin}_j d(\boldsymbol{x}, \boldsymbol{w}_j)$. This is realized by minimization of the cost function

$$\mathcal{E} = \sum_{\boldsymbol{x} \in \mathcal{X}} f\left(\frac{d(\boldsymbol{x}, \boldsymbol{w}^+) - d(\boldsymbol{x}, \boldsymbol{w}^-)}{d(\boldsymbol{x}, \boldsymbol{w}^+) + d(\boldsymbol{x}, \boldsymbol{w}^-)} \right) \tag{1}$$

by stochastic gradient descent learning with respect to $\mathcal{W}$, where f is a monotonically increasing function, d is a dissimilarity measure in $\mathbb{R}^n$ and $\boldsymbol{w}^+$ and $\boldsymbol{w}^-$ denote the closest prototypes to $\boldsymbol{x}$ with a matching label, i.e. $c(\boldsymbol{x}) = c(\boldsymbol{w}^+)$, and non-matching label, i.e. $c(\boldsymbol{x}) \neq c(\boldsymbol{w}^-)$, respectively.

When d is set to be the (squared) Euclidean distance, as frequently done, all input dimensions are weighted equally. To overcome this drawback, in [14] the distance measure is adapted along with the prototypes, yielding a relevance profile of the respective vector space dimensions. The Generalized Matrix LVQ (GMLVQ) by [10] generalizes the relevance profile to a relevance matrix by considering an adaptive distance measure of the form $d_{\boldsymbol{\Omega}}(\boldsymbol{x}, \boldsymbol{w}) = (\boldsymbol{\Omega}(\boldsymbol{x} - \boldsymbol{w}))^2$ with $\boldsymbol{\Omega} \in \mathbb{R}^{m \times n}$ being a mapping matrix $m \leq n$ subject to adaptation during learning. The resulting matrix $\boldsymbol{\Lambda} = \boldsymbol{\Omega}^\intercal \boldsymbol{\Omega}$ is denoted as classification correlation matrix (CCM) with entries Λ_{ij} reflecting the correlations between the i^{th} and

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/.

j^{th} data features, which contribute to a class discrimination. The classification influence profile (CIP), defined as $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)^\intercal$ with $\lambda_i = \sum_j |\Lambda_{ij}|$ provides the importance of the i^{th} data feature for classification.

3 Sensoric representations of graphs for efficient learning

Consider a reference graph $r \in \mathcal{G}$ that may be understood as sensor in the space $\mathcal{G}$ of graphs. Thus, each graph $x \in \mathcal{G} \setminus r$ can then be represented relatively to r, i.e. in terms of multiple measurement schemes relating x to r. We denote this approach as *sensoric response principle* (SRP). A way of comprehending those responses is in terms of object proximities, e.g. in our context graph kernels. Consequently, the sensoric representation $\boldsymbol{x} \in \mathcal{X}$ of a graph $x \in \mathcal{G}$ is given by the sensoric response vector $\boldsymbol{x} = \boldsymbol{d}(x, r)$, with

$$\boldsymbol{d}(x, r) = (d_{\kappa_1}(x, r), d_{\kappa_2}(x, r), \ldots, d_{\kappa_n}(x, r))^\intercal \tag{2}$$

where d_{κ_i} denotes the kernel distance corresponding to the graph kernel κ_i. Hence, $\boldsymbol{r} = \boldsymbol{d}(r, r)$ serves as a reference in the feature vector (proximity) space. Fig. 1 visualizes this SRP. At the simplest, the reference r is chosen randomly from $\mathcal{G}$. To ensure no outlier was selected, a refinement step may be applied that determines a new reference as the graph whose sensor representation w.r.t. r is closest to the mean representation $\bar{\boldsymbol{x}}$ of all graphs, i.e. $r_{\text{new}} = x_s$ with $s = \operatorname{argmin}_j d(\bar{\boldsymbol{x}}, \boldsymbol{x}_j)$.

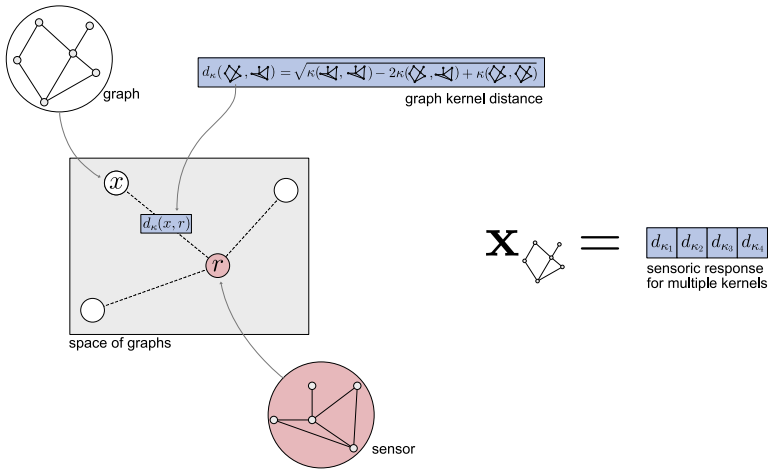

Fig. 1: Sensoric response features of graphs.

It is remarked that multiple references $r \in \mathcal{R} = \{r_i\}_{i=1}^{|\mathcal{R}|} \subset \mathcal{G}$, $|\mathcal{R}| \ll |\mathcal{G}|$ may be chosen as also suggested in [7], yielding $\boldsymbol{x} = \big(\boldsymbol{d}(x, r_1), \ldots, \boldsymbol{d}(x, r_{|\mathcal{R}|})\big)$ as a graph's sensor representation. A set of optimal representatives $\mathcal{R}$ may be learned by median Neural Gas [15], which, however, requires full distance

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/.

matrices, whose calculation was to be avoided in the first place. The obtained sensor representations of graphs (feature vectors) serve as input to GMLVQ described in Sec. 2.2.

Why the sensor perspective? Generally, this procedure enables machine learning methods that intrinsically rely on vectorial data, now to be applied to structured data. By mapping the graphs to proximity space we avoid the restriction to median and relational variants of classifiers. Further, the design of the sensor principle results in savings in time complexity: Given $|\mathcal{G}| = N$ data points, $|\mathcal{R}|$ references and N_κ kernel methods to be evaluated. Then the sensor principle only requires calculation of each kernel distance from every graph to the reference and thus $\mathcal{O}(2N \cdot N_\kappa \cdot |\mathcal{R}|)$ with $N_\kappa \ll N$. In comparison, evaluating N_κ kernel values between each pair of data points yields $\mathcal{O}(\frac{N^2}{2} \cdot N_\kappa)$.

Why relevance learning? Graph kernels used in practice are *incomplete*, i.e. there are non-isomorphic graphs x and $y \in \mathcal{G}$ with $\phi(x) = \phi(y)$ that cannot be distinguished by the kernel [2]. This incomplete perspective on the graph demands a careful kernel selection to reduce bias. However, problem-adequate approaches are usually unknown in advance, making methods to learn their optimal combination promising. Besides multiple kernel learning for SVM [3], a kernel-related approach based on matrix learning by neighborhood components analysis (NCA) [16] was presented in [6] for graphs. However, NCA is based on the k-Nearest Neighbour algorithm, whose disadvantages are well-known.

4 Experiments

We evaluated our approach by training a GMLVQ on sensoric representations of chem- and bioinformatic graphs from the publicly available TUDataset [17]. Graph representations were obtained by computing graph kernel distances using the GraKeL library [18].

Data set, graph kernel and classification settings In particular, we considered the small molecule data sets MUTAG and AIDS and the bioinformatic data set PROTEINS. All graphs are undirected and node-labeled. We constructed graph feature vectors using the SRP based on a randomly selected reference from the data set and distance computation based on eight graph kernels: the vertex histogram (VH), shortest path (SP), Weisfeiler-Lehman-subtree (WL-VH), Core shortest path (CORE-SP), ordered directed acyclic graph decomposition - subtree h (Odd Sth), pyramid match (PM), propagation (PK) and neighborhood hash (NH) kernel. We evaluated the GMLVQ-based relevance learning of sensoric graph representations using 3 prototypes per class and 10-fold cross-validation. Further, two baseline models were considered: 1) a simple GLVQ for comparison with no feature, i.e. distance weighting and 2) a SVM from the benchmark [19] for comparison with a non-SRP based pairwise proximity computation using single graph kernels. Results from the latter are directly comparable as it bases on the same data sets and kernel implementations.

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/.

Results and discussion Table 1 gives an overview of the achieved classification accuracies for the previously described models and datasets. Fig. 2 provides insights into the influence of individual kernels on the classification.

	GMLVQ	GLVQ	SVM[1]
MUTAG	89.7 (±8.5)	86.2 (±4.6)	88.3 (±6.3)
AIDS	99.1 (±1.2)	98.3 (±0.5)	99.7 (±0.3)
PROTEINS	75.1 (±3.5)	72.0 (±4.0)	76.5 (±3.9)

[1] results from [19]: best performing kernels NH, PM and CORE-SP

Table 1: Mean classification accuracies with the standard deviation [%].

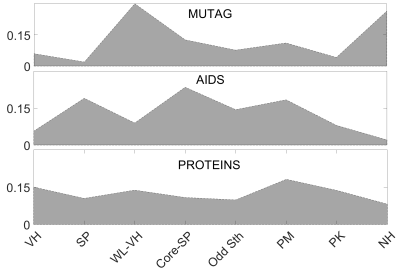

Fig. 2: CIPs obtained by GMLVQ.

GMLVQ outperforms GLVQ on all data sets, which suggests that a relevance learning, i.e. a weighted kernel combination is beneficial for the class discrimination. Furthermore, GMLVQ yields accuracies comparable to SVM with a single graph kernel although its computation load is reduced significantly due to the SRP. In [19] it is further found that state-of-the-art graph neural networks (GNNs) do not outperform the simpler graph kernel-SVM combination for classifying (discrete) node-labeled graphs. Nevertheless, GNNs enjoy great popularity, especially since graph kernels scale poorly for large data sets (with hundreds of thousands of graphs). With the SRP, we now propose an alternative that relies on sparse and interpretable models still applicable to large data sets.

5 Conclusion

In this contribution, we propose the use of a sensoric response principle for converting proximity data obtained by graph kernels into vectorial features for machine learning. Considering multiple kernels avoids the so-called feature bias or an elaborate model selection, while using the sensor perspective reduces the computation time of the kernels significantly. In combination with matrix learning LVQ, the approach allows problem-specific weighting of individual features. In the experiments, we empirically showed the potential of the SRP in the context of small molecule classification.

As the next step, we intend to extend the response principle introduced for graph kernels to that of topological descriptors and graph edit distances [20] in order to capture graph commonalities and differences even more comprehensively. Also other strategies for prototype determination like in [21] should be investigated in this context. We see future applications for graphs beyond the scope of chem- and bioinformatics, as well as for data structures that might be transformed into graphs [22].

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/.

References

[1] Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63, 2015.

[2] Nils M. Kriege, Fredrik D. Johansson, and Christopher Morris. A survey on graph kernels. *Applied Network Science*, 5(1):6, 2020.

[3] Michele Donini, Nicolò Navarin, Ivano Lauriola, Fabio Aiolli, and Fabrizio Costa. Fast hyperparameter selection for graph kernels via subsampling and multiple kernel learning. In *ESANN 2017 Proceedings*, pages 287–292, Bruges, Belgium, 2017.

[4] Bernhard Schölkopf and Alexander Smola. *Learning with Kernels*. MIT Press, Cambridge, 2002.

[5] David Nebel, Barbara Hammer, Kathleen Frohberg, and Thomas Villmann. Median variants of learning vector quantization for learning of dissimilarity data. *Neurocomputing*, 169:295–305, 2015.

[6] Adam Woźnica, Alexandros Kalousis, and Melanie Hilario. Adaptive Matching Based Kernels for Labelled Graphs. In *Advances in Knowledge Discovery and Data Mining*, volume 6119, pages 374–385, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[7] Robert P.W. Duin and Elżbieta Pękalska. The dissimilarity space: Bridging structural and statistical pattern recognition. *Pattern Recognition Letters*, 33(7):826–832, 2012.

[8] Marika Kaden, Ronny Schubert, Mehrdad Mohannazadeh Bakhtiari, Lucas Schwarz, and Thomas Villmann. The LVQ-based Counter Propagation Network – an Interpretable Information Bottleneck Approach. In *ESANN 2021 Proceedings*, pages 581–586, Online event (Bruges, Belgium), 2021.

[9] Feryel Zoghlami, Marika Kaden, Thomas Villmann, Germar Schneider, and Harald Heinrich. AI-Based Multi Sensor Fusion for Smart Decision Making: A Bi-Functional System for Single Sensor Evaluation in a Classification Task. *Sensors*, 21(13):4405, 2021.

[10] Petra Schneider, Michael Biehl, and Barbara Hammer. Adaptive Relevance Matrices in Learning Vector Quantization. *Neural Computation*, 21(12):3532–3561, 2009.

[11] Paulo Lisboa, Sascha Saralajew, Alfredo Vellido, and Thomas Villmann. The coming of age of interpretable and explainable machine learning models. In M. Verleysen, editor, *ESANN 2021 Proceedings*, pages 547–556, Bruges, Belgium, 2021.

[12] David Haussler. Convolution kernels on discrete structures. *Technical Report*, 1999.

[13] Atsushi Sato and Keiji Yamada. Generalized Learning Vector Quantization. In *Advances in Neural Information Processing Systems*, pages 423–429, Cambridge, MA, USA, 1996. MIT Press.

[14] Barbara Hammer and Thomas Villmann. Generalized relevance learning vector quantization. *Neural Networks*, 15(8):1059–1068, 2002.

[15] Marie Cottrell, Barbara Hammer, Alexander Hasenfuß, and Thomas Villmann. Batch and median neural gas. *Neural Networks*, 19(6-7):762–771, 2006.

[16] Jacob Goldberger, Sam Roweis, Geoff Hinton, and Ruslan Salakhutdinov. Neighbourhood Components Analysis. In *Advances in Neural Information Processing Systems*, pages 513–520, Vancouver, British Columbia, Canada, 2005.

[17] Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. TUDataset: A collection of benchmark datasets for learning with graphs. In *ICML 2020 Workshop on Graph Representation Learning and Beyond*, 2020.

[18] Giannis Siglidis, Giannis Nikolentzos, Stratis Limnios, Christos Giatsidis, Konstantinos Skianis, and Michalis Vazirgiannis. GraKeL: A graph kernel library in python. *Journal of Machine Learning Research*, 21:54:1–54:5, 2020.

[19] Giannis Nikolentzos, Giannis Siglidis, and Michalis Vazirgiannis. Graph Kernels: A Survey. *Journal of Artificial Intelligence Research*, 72:943–1027, 2021.

[20] Alberto Sanfeliu and King-Sun Fu. A distance measure between attributed relational graphs for pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(3):353–362, 1983.

[21] Benjamin Paaßen and Thomas Villmann. Prototype selection based on set covering and large margins. *Machine Learning Reports*, 14(MLR-03-2021):35–42, 2021.

[22] Lucas Lacasa, Bartolo Luque, Fernando Ballesteros, Jordi Luque, and Juan Carlos Nuño. From time series to complex networks: The visibility graph. *Proceedings of the National Academy of Sciences*, 105(13):4972–4975, 2008.