ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Sparse Nyström Approximation for Non-Vectorial Data Using Class-informed Landmark Selection

Maximilian Münch[1,2], Katrin Sophie Bohnsack[2,3], Alexander Engelsberger[2,3], Frank-Michael Schleif[1] and Thomas Villmann[3] *

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

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

Abstract. We introduce an efficient approach for supervised landmark selection in sparse Nyström approximation of kernel matrices. Our method converts structured non-vectorial input data such as graphs or text into a vectorial dissimilarity representation, enabling class-informed landmark identification through prototype-based learning. Experimental results show competitive approximation quality compared to existing strategies and demonstrate the positive effect of integrating class information into the selection process of Nyström landmarks making our approach an efficient and versatile solution for large-scale kernel learning.

1 Introduction

Kernel methods have gained widespread popularity in the field of machine learning due to their ability to handle a variety of structured data types, such as graphs, time series, and text documents [3, 5, 8]. Essential for these methods is the computation of kernel matrices, which contain the pairwise similarities between all N data points in an implicit high-dimensional feature space. However, the computational complexity of calculating all pairwise comparisons is $\mathcal{O}(N^2)$, making a calculation of the complete kernel matrix computationally prohibitive.

A widely adopted technique for this bottleneck is the Nyström approximation [18], drastically reducing the computational costs by requiring only a small subset of so-called landmark points to create a kernel matrix. A crucial step in this technique is the selection of those landmarks that reflect the data information appropriately. Therefore, various strategies have been proposed in the past [2, 10, 12, 18]. Although these methods identify high-quality landmarks, accurately representing the data space, they completely ignore the overall task, namely the classification or regression problem in general. We overcome this limitation by a landmark selection strategy, incorporating label information during the landmark determination process.

*MM is supported by the Bavarian HighTech agenda and the Würzburg Center for Artificial Intelligence and Robotics (CAIRO), KSB is supported by the German Federal Ministry for Economic Affairs and Climate Action (BMWK), AE is supported by the German Federal Ministry of Education and Research (BMBF). Additionally, we thank Prof. Michael Biehl for creating an invaluable comfortable atmosphere for scientific exchange at the WISCI 2023 where the idea for this paper was born.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

We start with some concepts of Nyström matrix approximation, data embeddings, and supervised prototype learning, before detailing our proposed method. Subsequently, we evaluate our method's performance on several datasets compared to other landmark selection strategies and discuss the experimental results.

2 Basic Notation and Related Work

Kernel approximation via Nyström: Let $\Xi = \{\xi_1, \xi_2, \ldots, \xi_N\}$ be a set of N objects, and let $\mathbf{K}$ be the $N \times N$ kernel matrix with entries $K_{i,j}$ given by the kernel function $k(\xi_i, \xi_j)$. Yet, a major challenge in the field of kernel methods is the overall computational complexity of $\mathcal{O}(N^2)$ when the entire kernel matrix has to be calculated. Over the last decades, the Nyström method [18] has proven to be a highly efficient technique to approximate the kernel matrix $\mathbf{K}$ using only a subset of L landmark points, with $L \ll N$. Here, an approximated matrix $\tilde{\mathbf{K}}$ is derived by $\mathbf{K} \approx \mathbf{C}\mathbf{W}^{-1}\mathbf{C}^T$, where $\mathbf{C} \in \mathbb{R}^{N \times L}$ is the matrix containing the kernel values between all data and the landmark points, and the matrix $\mathbf{W} \in \mathbb{R}^{L \times L}$ contains the kernel function's results between the landmark points. Consequently, the Nyström method offers a substantial reduction in computational load requiring only $N \times L$ kernel function evaluations to calculate the entire kernel matrix. In general, a key factor for the method's approximation quality is a careful selection of the landmarks, as their ability to accurately represent the data space is of great importance. For this purpose, several landmark selection strategies have been suggested like random sampling [18], which is computationally inexpensive but may not be appropriate for sparse data regions, or non-uniform selection techniques such as leverage score-based sampling [12], clustering strategies [19] and various others [2, 4]. While all these techniques exhibit potential, their applicability is limited either due to quadratic computational costs or the input's structured form. One of the few methods achieving an efficient selection of proper landmarks is based on k-means++ in a n-dimensional dissimilarity representation space (see below). The overall complexity is the sum of $\mathcal{O}(N \cdot n)$ for the mapping, $\mathcal{O}(N \cdot L \cdot n)$ for k-means++, and $\mathcal{O}(N \cdot L^2 + L^3)$ for the final Nyström method [10]. However, since k-means++ conducts entirely unsupervised clustering like all aforementioned strategies as well, no further class or label information is incorporated into the landmark selection process.

Data representation via dissimilarity space: The dissimilarity representation (DR) introduced by [11] is an intuitive approach for making standard vector-based machine learning methods applicable to structured data. Given a set of such (non-vectorial) data objects $\Xi = \{\xi_1, \ldots, \xi_N\}$ with dissimilarity measure d, e.g. the distance $d_k(\xi_i, \xi_j) = \sqrt{k(\xi_i, \xi_i) - 2\mathfrak{Re}\left(k(\xi_i, \xi_j)\right) + k(\xi_j, \xi_j)}$ obtained from a task- and data-type-specific kernel k. Then, a data object can be embedded into an n-dimensional vector space by cumulating its dissimilarities to a small subset of so-called references, $R = \{ r_1, \ldots, r_n \} \subset \Xi$ with $n \ll N$ such that the object ξ_i is represented by $\boldsymbol{x}_i = \big(d_k(\xi_i, r_1), \ldots, d_k(\xi_i, r_n)\big)^T$. Selection strategies for the $r_k \in R$ should minimize the effort in potentially complex proximity calculations, between the structured data [1], with random selection as a reasonable starting point [17].

Classification via learning vector quantization: Generalized learning

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

vector quantization (GLVQ) [13] is a cost-function-based variant of LVQ constituting an interpretable classification model, optimizing the hypothesis margin for class decision. Given data vectors $\boldsymbol{x} \in X = \{\boldsymbol{x}_i\}_{i=1}^{N} \subseteq \mathbb{R}^n$ together with their class labels $c(\boldsymbol{x}) \in \mathcal{C}$, GLVQ aims to find an optimal placement of class-dependent prototype vectors $\boldsymbol{w} \in \{\boldsymbol{w}_j\}_{j=1}^{L} \subset \mathbb{R}^n$ with $c(\boldsymbol{w}) \in \mathcal{C}$ such that misclassification of data by means of the nearest prototype principle, $c(\boldsymbol{w}_{s(\boldsymbol{x})})$ with $s(\boldsymbol{x}) = \arg\min_j \big(d(\boldsymbol{x}, \boldsymbol{w}_j)\big)$, is minimized. The prototype update realizes an attraction-repelling scheme (ARS), i.e. a vector shift towards or away from training data depending on their class label agreement. However, due to the integral repelling term, care should be taken with the interpretation of the prototypes [9]: these may be "pushed" far outside of the data distribution, making them highly data-untypical (non-representative), but rather class-discriminative. Nonetheless, LVQ variants enjoy popularity due to their simplicity and low complexity ($\mathcal{O}(N \cdot L \cdot n)$ for GLVQ).

3 Class-informed Landmark Selection for Nyström

In this section, we introduce an alternative landmark selection strategy, enabling the integration of class information given by an LVQ classifier. This might be particularly beneficial when considering only a small amount of landmarks, i.e. $1 - 10\%$ of the original input data: selection schemes solely based on data density may yield non-optimal landmarks w.r.t. the later classification, particularly in case of severely overlapping class distributions [7]. In contrast, the scheme proposed below results in an approximated kernel matrix that is adjusted, i.e. customized to the particular classification task to be solved. Hence, the approximation is not necessarily optimal for a general class-independent data representation but instead constitutes a kind of metric learning [6]. Yet, learning-based class-informed landmark selection in the non-vectorial data space is generally too costly because respective models require the full data dissimilarity matrix, which contradicts the objective of the Nyström approximation. A promising alternative is an Euclidean data embedding such that optimal landmarks can be obtained from a prototype-based classification model in this embedding space by an appropriate reverse mapping.

For the realisation of this overall strategy, first, we adopt the dissimilarity space representation (see Sec. 2) for embedding as suggested by [10, 11]: The non-vectorial data are represented by $X = \{\boldsymbol{x}_i\}_{i=1}^{N} \subseteq \mathbb{R}^n$ based on an appropriate kernel. We consider L class dependent prototypes $\boldsymbol{w}_j \in \mathbb{R}^n$ such that each class is covered and optimize them via GLVQ (see Sec. 2). Subsequently, landmark determination is done by reverse mapping realized as $s(\boldsymbol{w}_j) = \arg\min_{i=1,\ldots,N} \big(d_E(\boldsymbol{x}_i, \boldsymbol{w}_j)\big) \in \{1, \ldots, N\}$ to obtain the class-informed landmarks $l_j = \xi_{s(\boldsymbol{w}_j)} \in \Xi$, which finally are used in the Nyström approximation.[1] Alg. 1 summarizes this procedure.

[1] A note of caution: The concepts of references, prototypes and landmarks should be clearly distinguished: references determine a new representation of structured data in a vector space, prototypes define the decision boundary in classification settings in this representation space and landmarks correspond to structured objects facilitating the kernel matrix approximation.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Algorithm 1 Supervised landmark selection for kernel matrix approximation

Input: Data $\Xi = \{\xi_1, \ldots, \xi_N\}$ with class labels, number of references n, number of landmarks L

Output: Approximated kernel matrix $\tilde{\mathbf{K}}$ with $k_{ij} = k(\xi_i, \xi_j)$

1: Randomly select $n \ll N$ data $\xi_j \in \Xi$ serving as references $r_1, \ldots, r_n$
2: Map all data $\xi_i \in \Xi$ to the dissimilarity space by means of $\boldsymbol{x}_i = \big(d_k(\xi_i, r_1), \ldots, d_k(\xi_i, r_n)\big)^T \in X \subset \mathbb{R}^n$ ▷ DR
3: Train GLVQ on X with L prototypes $\boldsymbol{w}_j$ ▷ GLVQ
4: For all learned prototypes $\boldsymbol{w}_j$ identify their nearest data point with index $s(\boldsymbol{w}_j) = \arg\min_{i=1,\ldots,N}\big(d_E(\boldsymbol{x}_i, \boldsymbol{w}_j)\big) \in \{1, \ldots, N\}$
5: Consider the $l_j = \xi_{s(\boldsymbol{w}_j)} \in \Xi$ as final class-sensitive landmarks $l_1, \ldots, l_L$
6: Approximate the kernel matrix via Nyström $\tilde{\mathbf{K}} \approx \mathbf{C}\mathbf{W}^{-1}\mathbf{C}^T$ with $\mathbf{C} \in \mathbb{R}^{N \times L}$ and $\mathbf{W} \in \mathbb{R}^{L \times L}$ ▷ Nyström

Finally, the class-adjusted kernel approximation serves as an input for a (linear) support vector machine (SVM) with fast convergence because of convexity and proved classification robustness due to maximum class-separation margin [15]. We remark that this class-informed landmark determination has the same complexity as the approach proposed by [10] due to equivalent complexities of GLVQ and k-means++.

4 Experiments and Results

In order to show the superiority of our class-informed landmark selection, we evaluate our approach against other fast sampling strategies, namely random sampling and kernel k-means sampling. Our proposed approach proves to be especially beneficial when employing a minimal proportion of landmarks ($1 - 10\%$ of the original data size), as demonstrated by the evaluation of various well-established structured data sets. We evaluated our approach on three data sets AIDS, MSRC_21C (MSRC), and NCI1 using the Weisfeiler-Lehman kernel from the GraKeL library [16], as a state-of-the-art kernel in graph representations. Additionally, our experimental setup includes indefinite kernel data sets [14]: Flowcyto (normalized histograms), Music (earth movers distance), and Sonatas (normalized compression distance). To quantify the quality of matrix approximations, we use *classification accuracy* and *relative error* as suggested in [2].

Classification performance: As in this paper an optimal approximation for classification problems is prioritized, we first evaluate approximation quality through classification performance. Therefore, we evaluate the performance of an SVM trained on the approximated kernel matrix of the aforementioned data sets in a ten-fold cross-validation with nested cross-validation for parameter optimization. Due to numerical issues, we employed an eigenvalue shift for indefinite kernel matrices [2] when reconstructing the approximated matrix. Table 1 compares the SVM accuracy (mean accuracy over ten-fold cross-validation) of our approach with random sampling and kernel k-means.

In general, our approach shows an improved landmark sampling compared

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Dataset	Random	k-means	GLVQ
AIDS	**1.0**	**1.0**	0.99
MSRC	0.58	0.6	**0.81**
NCI1	**0.59**	0.53	0.57
Flowcyto	0.67	0.68	**0.69**
Music	0.56	0.56	**0.62**
Sonatas	0.79	0.79	**0.84**

Table 1: Mean classification accuracy using only 1% of the data for Nyström landmark selection.

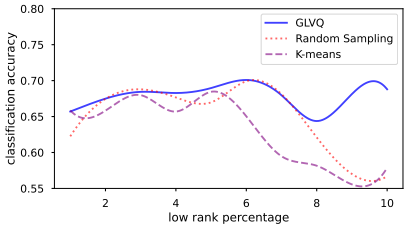

Fig. 1: Classification accuracy as a function of the number of landmarks (high to low) for Flowcyto dataset.

to random sampling and kernel k-means, particularly for MSRC, Music, and Sonatas datasets. For AIDS and NCI1, the performance of our strategy was slightly below, for Flowcyto it was slightly above the competitive methods showing nearly equal performance. In Fig. 1, we show exemplary for the *Flowcyto* data set the stable accuracy of our approach compared to the other sampling strategies over a varying amount of used landmarks for matrix reconstruction.

Relative error: Additionally, we evaluate the approximation's quality by means of the relative error between original and reconstructed kernel matrix in Table 2. The relative error is defined as $||\tilde{\mathbf{K}}-\mathbf{K}||_F/||\mathbf{K}||_F$ measuring the relative difference of $\mathbf{K}$ and $\tilde{\mathbf{K}}$ using the Frobenius norm.

Dataset	Random	k-means	GLVQ
AIDS	0.055	**0.041**	0.141
MSRC	0.541	0.585	**0.503**
NCI1	0.019	**0.016**	0.024
Flowcyto	0.405	**0.351**	0.377
Music	0.847	0.968	0.858
Sonatas	0.795	0.817	**0.760**

Table 2: Relative error between original and approximated matrix (by means of Frobenius norm).

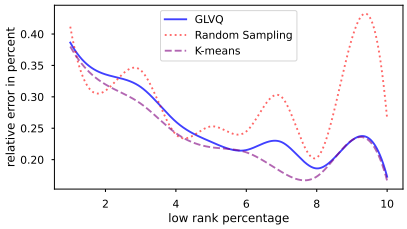

Fig. 2: Relative matrix approximation error as a function of the number of landmarks.

As expected, the relative error experiments show that our approach does not necessarily provide the best approximation in terms of a general data representation since it is tailored towards classification tasks. However, for MSRC and Sonatas datasets, our approach still manages to produce a superior or comparable approximation, as indicated by the lower relative error values. This is an expected behaviour since GLVQ landmark selection is class specific and does not focus on an optimal representation of the entire data space. In summary, our approach offers superior classification performance and competitive approximation error compared to other methods over various data sets, yielding an efficient supervised solution for large-scale, structured data challenges.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

5 Conclusions

In this paper, we introduced a novel landmark selection strategy for Nyström approximation that integrates label information into the selection process by means of prototype-based learning models. The experiments demonstrated the effectiveness of our method in approximating kernel matrices and its applicability to various structured data types with psd or non-psd kernel functions. In future work, we will apply our approach to a larger variety of data sets and analyze elaborated supervised prototype-based models to improve the overall performance and reduce the required landmarks in matrix approximation.

References

[1] K. S. Bohnsack, A. Engelsberger, M. Kaden, and T. Villmann. Efficient representation of biochemical structures for supervised and unsupervised machine learning models using multi-sensoric embeddings. In *Proc. of the 16th Int. Joint Conf. on Biom. Eng. Sys. and Techn. - Vol. 3: BIOINFORMATICS*, pages 59–69, 2023.

[2] D. Cai, J. Nagy, and Y. Xi. Fast deterministic approximation of symmetric indefinite kernel matrices with high dimensional datasets. *SIAM SIMAX*, 43(2):1003–1028, 2022.

[3] H. Chen, F. Tang, P. Tiňo, and X. Yao. Model-based kernel for efficient time series analysis. In *KDD*, pages 392–400. ACM, 2013.

[4] M. Fanuel, J. Schreurs, and J. A. K. Suykens. Diversity sampling is an implicit regularization for kernel methods. *SIAM J. Math. Data Sci.*, 3(1):280–297, 2021.

[5] M. Farhan, J. Tariq, A. Zaman, M. Shabbir, and I. Khan. Efficient approximation algorithms for strings kernel based sequence classification. In I. Guyon, editor, *Proc. of the 30th NIPS 2017*, pages 6935–6945, 2017.

[6] A. Globerson and S. Roweis. Metric learning by collapsing classes. In *Adv. in Neural Inf. Proc. Sys.*, volume 18. MIT Press, 2005.

[7] T. Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1997.

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

[9] K. L. Oehler and R. M. Gray. Combining image classification and image compression using vector quantization. *IEEE TPAMI*, 17(5):461–473, 1995.

[10] D. Oglic and T. Gärtner. Nyström method with kernel k-means++ samples as landmarks. In Doina Precup and Yee Whye Teh, editors, *Proc. of the 34th Int. Conf. on Mach. Learn., ICML 2017*, volume 70 of *Proc. of Mach. Learn. Res.*, pages 2652–2660. PMLR, 2017.

[11] E. Pekalska and R. P. W. Duin. *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications*. WORLD SCIENTIFIC, November 2005.

[12] A. Rudi, D. Calandriello, L. Carratino, and L. Rosasco. On fast leverage score sampling and optimal learning. In *Adv. in Neural Inf. Proc. Sys. 31*, pages 5677–5687, 2018.

[13] A. Sato and K. Yamada. Generalized learning vector quantization. In Hasselmo ME Touretzky DS, Mozer MC, editor, *Adv. in Neural Inf. Proc. Sys.*, volume 8, pages 423–429. MIT Press, Cambridge, 1996.

[14] F.-M. Schleif and P. Tiňo. Indefinite proximity learning: A review. *Neural Computation*, 27(10):2039–2096, 2015.

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

[16] G. Siglidis, G. Nikolentzos, S. Limnios, C. Giatsidis, K. Skianis, and M. Vazirgiannis. Grakel: A graph kernel library in python. *arXiv preprint arXiv:1806.02193*, 2018.

[17] L. Wang, M. Sugiyama, C. Yang, K. Hatano, and J. Feng. Theory and algorithm for learning with dissimilarity functions. *Neural Comp.*, 21(5):1459–1484, 2009.

[18] C. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *Adv. in NIPS*, pages 682–688. MIT Press, 2001.

[19] K. Zhang and J. T. Kwok. Clustered Nyström method for large scale manifold learning and dimension reduction. *IEEE TNN*, 21(10):1576–1587, 2010.