Publications of Vanja Doskoč

This document lists all peer-reviewed publications of Vanja Doskoč, Chair for Algorithm Engineering, Hasso Plattner Institute, Potsdam, Germany. This listing was automatically generated on December 1, 2021. An up-to-date version is available online at hpi.de/friedrich/docs/publist/doskoc.pdf.

Conference papers

[1] Quinzan, F., Doskoč, V., Göbel, A., Friedrich, T., Adaptive Sampling for Fast Constrained Maximization of Submodular Functions. In: *Artificial Intelligence and Statistics (AISTATS)*, pp. 964–972, 2021.

Several large-scale machine learning tasks, such as data summarization, can be approached by maximizing functions that satisfy submodularity. These optimization problems often involve complex side constraints, imposed by the underlying application. In this paper, we develop an algorithm with poly-logarithmic adaptivity for non-monotone submodular maximization under general side constraints. The adaptive complexity of a problem is the minimal number of sequential rounds required to achieve the objective. Our algorithm is suitable to maximize a non-monotone submodular function under a p-system side constraint, and it achieves a $(p+O(\sqrt{p}))$-approximation for this problem, after only poly-logarithmic adaptive rounds and polynomial queries to the valuation oracle function. Furthermore, our algorithm achieves a $p + O(1))$-approximation when the given side constraint is a p-extendible system. This algorithm yields an exponential speed-up, with respect to the adaptivity, over any other known constant-factor approximation algorithm for this problem. It also competes with previous known results in terms of the query complexity. We perform various experiments on various real-world applications. We find that, in comparison with commonly used heuristics, our algorithm performs better on these instances.

[2] Doskoč, V., Kötzing, T., Normal Forms for Semantically Witness-Based Learners in Inductive Inference. In: *Computability in Europe (CiE)*, 2021.

In inductive inference, we study learners (computable devices) inferring formal languages. In particular, we consider semantically witness-based learners, that is, learners which are required to justify each of their semantic mind changes. This natural requirement deserves special attention as it is a specialization of various important learning paradigms. As such, it has already proven to be fruitful for gaining knowledge about other types of restrictions. In this paper, we provide a thorough analysis of semantically converging, semantically witness-based learners, obtaining normal forms for them. Most notably, we show that set-driven globally semantically witness-based learners are equally powerful as their Gold-style semantically conservative counterpart. Such results are key to understanding the, yet undiscovered, mutual relation between various important learning paradigms of semantically converging learners.

[3] Doskoč, V., Kötzing, T., Mapping Monotonic Restrictions in Inductive Inference. In: *Computability in Europe (CiE)*, 2021.

In inductive inference we investigate computable devices (learners) learning formal languages. In this work, we focus on monotonic learners which, despite their natural motivation, exhibit peculiar behaviour. A recent study analysed the learning capabilities of strongly monotone learners in various settings. The therein unveiled differences between explanatory (syntactically converging) and behaviourally correct (semantically converging) such learners motivate our studies of monotone learners in the same settings. While the structure of the pairwise relations for monotone explanatory learning is similar to the strongly monotone case (and for similar reasons), for behaviourally correct learning a very different picture emerges. In the latter setup, we provide a self-learning class of languages showing that monotone learners, as opposed to their strongly monotone counterpart, do heavily rely on the order in which the information is given, an unusual result for behaviourally correct learners.

[4] Berger, J., Böther, M., Doskoč, V., Gadea Harder, J., Klodt, N., Kötzing, T., Lötzsch, W., Peters, J., Schiller, L., Seifert, L., Wells, A., Wietheger, S., Learning Languages with Decidable Hypotheses. In: *Computability in Europe (CiE)*, 2021.

In language learning in the limit, the most common type of hypothesis is to give an enumerator for a language, a W-index. These hypotheses have the drawback that even the membership problem is undecidable. In this paper, we use a different system which allows for naming arbitrary decidable languages, namely programs for characteristic functions (called C-indices). These indices have the drawback that it is now not decidable whether a given hypothesis is even a legal C-index. In this first analysis of learning with C-indices, we give a structured account of the learning power of various restrictions employing C-indices, also when compared with W-indices. We establish a hierarchy of learning power depending on whether C-indices are required (a) on all outputs; (b) only on outputs relevant for the class to be learned or (c) only in the limit as final, correct hypotheses. We analyze all these questions also in relation to the mode of data presentation. Finally, we also ask about the relation of semantic versus syntactic convergence and derive the map of pairwise relations for these two kinds of convergence coupled with various forms of data presentation.

[5] Berger, J. Fine-Grained Localization, Classification and Segmentation of Lungs with Various Diseases. In: *CVPR Workshop on Fine-Grained Visual Categorization (FGVC@CVPR)*, 2021.

The fine-grained localization and classification of various lung abnormalities is a challenging yet important task for combating diseases and, also, pandemics. In this paper, we present one way to detect and classify abnormalities within chest X-ray scans. In particular, we investigate the use of binary image classification (to distinguish between healthy and infected chests) and the weighted box fusion (which constructs a detection box using the proposed boxes within range). We observe that both methods increase the performance of a base model significantly. Furthermore, we improve state of the art on lung segmentation, even in the presence of abnormalities. We do so using transfer learning to fine-tune a UNet model on the Montgomery and Shenzhen datasets. In our experiments, we compare standard augmentations (like crop, pad, rotate, warp, zoom, brightness, and contrast variations) to more complex ones (for example,

block masking and diffused noise augmentations). This way, we obtain a state-of-the-art model with a dice score of 97.9%. In particular, we show that simple augmentations outperform complex ones in our setting.

[6] Kißig, O., Taraz, M., Cohen, S., Doskoč, V., Friedrich, T., Drug Repurposing for Multiple COVID Strains using Collaborative Filtering. In: *ICLR Workshop on Machine Learning for Preventing and Combating Pandemics (MLPCP@ICLR)*, 2021.

The ongoing COVID-19 pandemic demands for a swift discovery of suitable treatments. The development of completely new compounds for such a novel disease is a challenging, time intensive process. This amplifies the relevance of drug repurposing, a technique where existing drugs are used to treat other diseases. A common bioinformatical approach to this is based on knowledge graphs, which compile relationships between drugs, diseases, genes and other biomedical entities. Then, graph neural networks (GNNs) are used for the drug repurposing task as they provide a good link prediction performance on such knowledge graphs. Building on state-of-the-art GNN research, Doshi & Chepuri (2020) construct the remarkable model DR-COVID. We re-implement their model and extend the approach to perform significantly better. We propose and evaluate several strategies for the aggregation of link predictions into drug recommendation rankings. With the help of clustering of similar target diseases we improve the model by a substantial margin, compiling a top-100 ranking of candidates including 32 currently being in COVID-19-related clinical trials. Regarding the re-implementation, we offer more flexibility in the selection of the graph neighborhood sizes fed into the model and reduce the training time significantly by making use of data parallelism.

[7] Doskoč, V., Kötzing, T., Cautious Limit Learning. In: *Algorithmic Learning Theory (ALT)*, pp. 251–276, 2020.

We investigate language learning in the limit from text with various *cautious* learning restrictions. Learning is *cautious* if no hypothesis is a proper subset of a previous guess. While dealing with a seemingly natural learning behaviour, cautious learning does severely restrict explanatory (syntactic) learning power. To further understand why exactly this loss of learning power arises, Kötzing and Palenta (2016) introduced weakened versions of cautious learning and gave first partial results on their relation. In this paper, we aim to understand the restriction of cautious learning more fully. To this end we compare the known variants in a number of different settings, namely full-information and (partially) set-driven learning, paired either with the syntactic convergence restriction (explanatory learning) or the semantic convergence restriction (behaviourally correct learning). To do so, we make use of normal forms presented in Kötzing et al. (2017), most notably strongly locking and consistent learning. While strongly locking learners have been exploited when dealing with a variety of syntactic learning restrictions, we show how they can be beneficial in the semantic case as well. Furthermore, we expand the normal forms to a broader range of learning restrictions, including an answer to the open question of whether cautious learners can be assumed to be consistent, as stated in Kötzing et al. (2017).

[8] Doskoč, V., Friedrich, T., Göbel, A., Neumann, A., Neumann, F., Quinzan, F., Non-Monotone Submodular Maximization with Multiple Knapsacks in Static and Dynamic Settings. In: *European Conference on Artificial Intelligence (ECAI)*, pp. 435–442, 2020.

We study the problem of maximizing a non-monotone submodular function under multiple knapsack constraints. We propose a simple discrete greedy algorithm to approach this problem, and prove that it yields strong approximation guarantees for functions with bounded curvature. In contrast to other heuristics, this does not require problem relaxation to continuous domains and it maintains a constant-factor approximation guarantee in the problem size. In the case of a single knapsack, our analysis suggests that the standard greedy can be used in non-monotone settings. Additionally, we study this problem in a dynamic setting, in which knapsacks change during the optimization process. We modify our greedy algorithm to avoid a complete restart at each constraint update. This modification retains the approximation guarantees of the static case. We evaluate our results experimentally on a video summarization and sensor placement task. We show that our proposed algorithm competes with the state-of-the-art in static settings. Furthermore, we show that in dynamic settings with tight computational time budget, our modified greedy yields significant improvements over starting the greedy from scratch, in terms of the solution quality achieved.

[9] Doskoč, V. Confident Iterative Learning in Computational Learning Theory. In: *Current Trends in Theory and Practice of Computer Science (SOFSEM)*, pp. 30–42, 2018.

In inductive inference various types of learning have emerged. The main aim of this paper is to investigate a new type of learning, the confident iterative learning. Given a class to be learnt, the idea here is to merge the following two concepts. For confidence, we require the learner to converge on any set, however, it only needs to be correct on the sets in the class. To be iterative, we restrict the learner's memory on previ- ous inputs and calculations to its last hypothesis. Investigating the new learner, we will provide negative and positive examples, as well as some properties the confident iterative learner possesses. This will peak at a classification theorem for certain types of classes. Next, we will introduce and compare different types of confidence, focus- ing on the learner's behaviour on sets outside of the class. Lastly, we will focus on the possible hypotheses. Introducing learning with respect to hypothesis spaces, we will provide examples witnessing that exact, class preserving and class comprising learning are different.