Sujet de thèse de doctorat – PhD position offer

Learning-based Module for Neighborhood Design to Solve Discrete Graphical Models.

Supervision: Pr. Jean-Luc Lamotte, Pr. Samir Loudni and Dr. Abdelkader Ouali.

Location: CODAG team, GREYC laboratory UMR CNRS 6072, University of Caen Normandy. 6 Boulevard du Maréchal Juin Bâtiment Sciences 3. CS 14032, 14032 CAEN cedex 5.

Funding: 36 months PhD scholarship by ANR-GMLAS project.

Salary: Approximately 2300 euros gross per month.

Starting date: by October 1, 2025.

Desired Profile: Master's level (or equivalent) in computer science (or applied mathematics) with an interest in reinforcement learning, data mining, metaheuristics, constraint programming. Programming skills mainly in C++ and Python as well as a good understanding of algorithms and constraint solving will be appreciated. The working language is French or English.

Application: email to abdelkader.ouali@unicaen.fr and samir.loudni@imt-atlantique.fr your CV, motivation letter, academic results (notes/grades/ranking) of the last two years, recommendation letters, one or two project reports (internship or big course projects). Please mention in the email subject "Application PhD ANR GMLAS". The application deadline is June 30.

Context

This research is part of the GMLaS project, funded by the ANR 2024 (Agence Nationale de la Recherche), which focuses on *Discrete Graphical Model Learning and Solving* using data mining and deep learning techniques. The main objective of the GMLaS project is to automatically learn promising search areas from search history, with the aim of rapidly finding high-quality solutions. In collaboration with the SaAB team from the Mathematics and Applied Computer Science unit in Toulouse (MIAT), a key application is being explored to improve the resolution of large instances in protein design.

Subject

As part of this PhD thesis, we study problems modeled using Cost Function Networks (CFNs), a generalization of constraint networks in which relationships between variables are associated with costs. This framework allows the formalization of optimization problems aiming to minimize the total cost across the network, in order to find an optimal solution while accounting for possible trade-offs between different relationships [6]. Cost Function Networks are applied in various fields, notably for solving very large-scale problems in bioinformatics (genetics, molecular biology, etc.), as well as in the optimization of large stochastic graphical models, such as Bayesian networks and Markov random fields [7].

This thesis aims at proposing and developing data-driven methods and algorithms for mining and learning search neighborhoods by leveraging pattern mining techniques [1,2,3,4,5] to improve the efficiency of the UDGVNS method [8]. The main challenge here is to achieve a trade-off between the accuracy of extracted patterns and the computational overhead of the mining and learning process. This thesis is organized in three main tasks.

Pattern mining module for neighborhood search:

We begin by focusing on the integration of classical algorithms for mining frequent closed patterns [1,4] from the search history. Frequent closed patterns are particularly useful as they reduce redundancy among discovered patterns, leading to a more compact and efficient set of patterns for neighborhood construction. To ensure low-overhead integration within the search process, we will explore various representations of the search history that are compatible with standard pattern mining algorithms. These representations will be based on data collected from the current solutions and their associated neighborhoods during the search. Additionally, the integration of the pattern mining step will be made dynamic and adaptive.

Exploring high-order neighborhood search structures using complex patterns:

Finding appropriate threshold values for classical pattern mining algorithms without prior knowledge is particularly challenging. Poorly chosen thresholds can lead to the extraction of less informative patterns, which in turn results in less effective neighborhood constructions. To address this limitation, we propose integrating more advanced data mining techniques within toulbar2-VNS to directly search for high-quality patterns, such as Pareto-optimal patterns, based on multiple interestingness measures without requiring predefined thresholds [2,3]. Additionally, we will explore techniques involving higher-order relationships, such as pattern set mining [9], to further enrich the information extracted from the search history. Beyond pattern mining, we will design a learning procedure that learns guiding rules for escaping local optima, allowing the search process to identify promising directions for improvement.

Client-server architecture for UDGVNS

The parallel version of toulbar2-VNS for discrete GMs [8] could benefit from a new cooperation model to efficiently integrate the mining process during the search. We plan to delegate the mining tasks to workers and provide direct communication schemes between workers based on the results of their mining process. The concrete result of this line of research will be a module laying out rules for the general integration of pattern mining within the toulbar2-VNS metaheuristic.

References:

[1] Uno T, Asai T, Uchida Y, Arimura H. An Efficient Algorithm for Enumerating Closed Patterns in Transaction Databases. Discovery Science. Springer Berlin Heidelberg; 2004. pp. 16–31.

[2] Ugarte W, Boizumault P, Crémilleux B, Lepailleur A, Loudni S, Plantevit M, et al. Skypattern mining: From pattern condensed representations to dynamic constraint satisfaction problems. Artif Intell. 2017;244: 48–69.

[3] Vernerey C, Loudni S, Aribi N, Lebbah Y. Threshold-free pattern mining meets Multi-objective Optimization: Application to association rules. Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. 2022. doi:10.24963/ijcai.2022/261

[4] Djawad Bekkoucha, Abdelkader Ouali, Patrice Boizumault, and Bruno Crémilleux. Efficiently mining closed interval patterns with constraint programming. In Bistra Dilkina, editor, *CPAIOR 2024, Uppsala, Sweden, May 28-31, 2024, Proceedings, Part I*, volume 14742 of *Lecture Notes in Computer Science*, pages 51--67. Springer, 2024.

[5] Riondato M, Upfal E. Efficient Discovery of Association Rules and Frequent Itemsets through Sampling with Tight Performance Guarantees. ACM Trans Knowl Discov Data. 2014;8: 1–32.

[6] Schiex T, Fargier H, Verfaillie G, Others. Valued constraint satisfaction problems: Hard and easy problems. IJCAI . 1995;95: 631–639.

[7] Cooper MC, de Givry S, Schiex T. Graphical Models: Queries, Complexity, Algorithms (Tutorial). 37th International Symposium on Theoretical Aspects of Computer Science (STACS 2020). Schloss Dagstuhl – Leibniz-Zentrum für Informatik; 2020. p. 4:22.

[8] Ouali A, Allouche D, de Givry S, Loudni S, Lebbah Y, Loukil L, et al. Variable neighborhood search for graphical model energy minimization. Artif Intell. 2020;278: 103194.

[9] Ouali A, Loudni S, Lebbah Y, Boizumault P, Zimmermann A, Loukil L. Efficiently finding conceptual clustering models with integer linear programming. Int Jt Conf Artif Intell. 2016; 647–654.