TITRE DE LA THESE: Discrete **G**raphical **M**odel **L**earning and **S**olving using **D**ata **M**ining (GMLaSDM)

Direction de thèse : Samir LOUDNI

Co-encadrant·es : Simon De Givry (INRAE Toulouse)

Laboratoire(s) : LS2N

Équipe(s) de recherche : TASC

Département(s) IMT Atlantique : DAPI

S'agit-il d'une thèse en cotutelle internationale ? Non

Le sujet proposé présente-il un caractère interdisciplinaire ? Oui
This thesis has an interdisciplinary character: on the one hand, the fusion of two distinct solving paradigms, constraint programming (AI) and data science (ML/DM) and on the other hand, the techniques developed during this thesis will be validated on the field of protein design which has many applications in biotechnology, health, and green chemistry (biofuels,...).

La source du co-financement est-elle identifiée ? Oui/Non
We are looking for regional co-financing or with our partnership INRAE Toulouse.

Autres informations :
A good candidate has been identified for this thesis who is currently doing her Master's internship on the same subject in the TASC team under the supervision of Samir Loudni.

Information candidate:

- Nom : CHAOUI Prénom : Sokayna
- MASTER 2 EME ANNÉE en cours à l'Université Paris 8 – M2 CYBERSÉCURITÉ ET SCIENCES DES DONNÉES
- Très bons résultats obtenus au premier semestre du M2, avec des notes entre 16 et 18/20
- Actuellement en stage M2R à IMT Atlantique (du 04/03/2024 au 02/09/2024) sur un sujet intitulé « Hybridations des méthodes d'optimisation par des méthodes de fouille de données ».

Scientific Context

In recent years, the fields of machine learning (ML) and data mining (DM) have seen tremendous progress, becoming a ubiquitous technology across a wide range of applications. One area that can significantly benefit from the use of ML is **constraint reasoning and optimization**. Three components of constraint problem solving, i.e., modeling, search and optimization, makes ML especially relevant. The hybridization of Data Mining, Machine, and Deep Learning with state-of-the-art (SOTA) discrete reasoning and optimization is one of the major challenges in Artificial Intelligence. Such hybridization has opened a research avenue with two directions [1]. The first direction focuses on using ML to improve the capacity of SOTA discrete solvers in tackling specific families of problems. The second direction extends the reasoning capabilities of DL on complex problems by having them collaborate with SOTA discrete solvers. **Our proposal explores the challenge of data-driven decision-making through a combined machine learning (ML), discrete reasoning/optimization framework, and associated solver**: Graphical Models (GMs) and the **toulbar2**[1] solver, winner of several competitions: UAI 2022 solver competition (winner on all MPE and MMAP task categories), XCSP3 2023 Competition (first place on Mini COP track and second place on Parallel COP track). Discrete GMs, especially additive GMs such as Cost Function Networks (CFNs) and Markov Random Fields [2] are attractive because they can represent logical and probabilistic information seamlessly. Then, toulbar2 is a SOTA solver for reasoning on such discrete models. It includes both exact solvers (e.g., Hybrid Best First Search [3]) and sophisticated meta-heuristics (e.g., Variable Neighborhood Search: **toulbar2-VNS** [4–6]).

When approximate methods, such as meta-heuristics (MHs) are required, ML and DM techniques can also cooperate for solving CO problems [7,8]. During exploration, MHs generate a considerable volume of data including good or bad solutions, evolution trajectories of different solutions, local optima, etc. These data potentially carry useful knowledge such as the properties of good and bad solutions, the performance of different operators in different stages of the search process, etc. ML/DM techniques can serve MHs by extracting useful knowledge from the generated data throughout the search process. Several successful studies have demonstrated the usefulness of ML/DM to enhance meta-heuristics, specifically for guiding neighborhood generation [9–15].

Goal and Scientific Issues of the thesis

ML/DM techniques can be used to leverage the generation of good neighbors by extracting common characteristics and patterns often present in high-quality solutions during or before the search. Hence, **The objective of this thesis is to design a general learning-based module** to extract these characteristics, and exploit the learned knowledge to guide the **toulbar2-VNS** meta-heuristic search towards compact and promising solution regions. More precisely, we aim at exploiting the learned knowledge to select relevant neighborhood structures (i.e. subset of variables allowed to be changed in the current solution) that leads to a promising neighborhood. However, the main technical difficulty lies in the ability to design such a module with reasonable complexity in terms of trade-off between the accuracy of extracted patterns and the computational overhead of the knowledge extraction. This requires finding *good representations of the search history* to fit pattern mining algorithms. We will design new representations built from the data collected from the current solutions and the neighborhood where they have been found. We will also explore different paradigms like sampling approaches [16] to achieve seamless integration without introducing heavy computational costs. Other research questions related to our module design is the *moments at which the data mining process should be performed* and the *frequency at which the knowledge should be updated and injected* to create new neighborhoods. While in most of the studies, learned knowledge takes the form of a set of rules or frequent patterns requiring thresholds constraints to extract them, this thesis will consider the *integration of more elaborated DM techniques* within **toulbar2-VNS** to search for high quality patterns (i.e., Pareto patterns) according to a set of interestingness measures without any thresholds [17]. We will also consider *discriminating patterns* highlighting the contrast

1 https://github.com/toulbar2/toulbar2

between the variables whose value modification results in an improvement and those leading to a non-improvement. We plan to experiment with our ML/CP hybrid techniques on UAI 2014/2022, XCSP 2022/2023, Weighted CP, and MIPLIB 2017 benchmarks, in addition to an extended set of difficult Computational Protein Design (CPD) instances developed by MIAT INRAE Toulouse.

Expected results: This thesis will endow the **toulbar2-VNS** meta-heuristic and its variants with learning abilities to better guide the generation of good neighborhoods from the generated data throughout the search process. As a criteria for success will be the potential resolution of difficult large-scale instances of the CPD problem in terms of run time and quality. This thesis will have a strong impact on the AI community as improvements in CFN solving and learning will directly impact the field of protein design.

Originality w.r.t. the state of the art

Existing hybridization of ML/DM techniques with meta-heuristics are specifically designed to solve a specific problem with a specific algorithm, which lacks generalization. More importantly, they are limited to either frequent patterns or association rules and require defining threshold constraints. This thesis targets a more general objective that goes beyond previous works in the sense that our hybridization intends to be problem independent and generally applicable to different optimization problems. **This thesis would be one of the first to consider** (1) Pareto patterns requiring no thresholds [17] and other forms of patterns, (2) strategies to ensure a trade-off between the accuracy of extracted knowledge and the computational overhead of the knowledge extraction process.

Compétences attendues du ou de la candidat·e :

Lister les principales compétences nécessaires pour ce sujet de thèse.

- Constraint Programming, Decision aid , machine learning
- Strong facility in software engineering and implementation (C++, Python)
- Strong mathematical and formal foundations
- A good command of written and oral English

Bibliography

1. Yoshua Bengio, Andrea Lodi, Antoine Prouvost: Machine learning for combinatorial optimization: A methodological tour d'horizon. Eur. J. Oper. Res. 290(2): 405-421, 2021.

2. Cooper, M., de Givry, S., Schiex, T. : Graphical models: queries, complexity, algorithms. In Proc. Of STACS-2020, 154, 4-1, 2020.

3. Allouche D, de Givry S, Katsirelos G, Schiex T, Zytnicki M. Anytime Hybrid Best-First Search with Tree Decomposition for Weighted CSP. Principles and Practice of Constraint Programming. Springer International Publishing; 2015. pp. 12–29.

4. Ouali A, Allouche D, de Givry S, **Loudni S**, Lebbah Y, Loukil L. Iterative decomposition guided Variable Neighborhood Search for graphical model energy minimization. Uncertain Artif Intell. 2017.

5. 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.

6. Boizumault P, de Givry S, **Loudni S**, Ouali A. Variable Neighborhood Search for Cost Function Networks. In: Kulkarni AJ, Gandomi AH, editors. Handbook of Formal Optimization. Singapore: Springer Nature Singapore; 2023. pp. 1–29.

7. Martins D, Vianna GM, Rosseti I, Martins SL, Plastino A. Making a state-of-the-art heuristic faster with data mining. Ann Oper Res. 2018;263: 141–162.

8. de Lima Martins S, Rosseti I, Plastino A. Data Mining in Stochastic Local Search. In: Martí R, Pardalos PM, Resende MGC, editors. Handbook of Heuristics. Cham: Springer International Publishing; 2018. pp. 39–87.

9. Guerine M, Rosseti I, Plastino A. Extending the hybridization of metaheuristics with data mining: Dealing with sequences. Intell Data Anal. 2016;20: 1133–1156.

10. Zhou Y, Hao J-K, Duval B. Frequent Pattern-Based Search: A Case Study on the Quadratic Assignment Problem. IEEE Trans Syst Man Cybern. 2022;52: 1503–1515.

11. Plastino A, Barbalho H, Santos LFM, Fuchshuber R, Martins SL. Adaptive and multi-mining versions of the DM-GRASP hybrid metaheuristic. J Heuristics. 2014;20: 39–74.

12. Santos HG, Ochi LS, Marinho EH, Drummond LMA. Combining an evolutionary algorithm with data mining to solve a single-vehicle routing problem. Neurocomputing. 2006;70: 70–77.

13. Raschip M, Croitoru C, Stoffel K. Guiding Evolutionary Search with Association Rules for Solving Weighted CSPs. Annual Conference on Genetic and Evolutionary Computation, GECCO 2015. New York, NY, USA: Association for Computing Machinery; 2015. pp. 481–488.

14. Zhou Y, Hao J-K, Duval B. Reinforcement learning based local search for grouping problems: A case study on graph coloring. Expert Syst Appl. 2016;64: 412–422.

15. Queiroz dos Santos JP, de Melo JD, Duarte Neto AD, Aloise D. Reactive Search strategies using Reinforcement Learning, local search algorithms and Variable Neighborhood Search. Expert Syst Appl. 2014;41: 4939–4949.

16. 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.

17. 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