São José dos Campos - SP
LV SBPO
LV Simpósio Brasileiro de Pesquisa Operacional
São José dos Campos, SP - 6 a 9 de novembro de 2023

Novel Data Mining Methods and Applications in Combinatorial Optimization and Bioinformatics

Marcelo Rodrigues de Holanda Maia[1,2], Alexandre Plastino[1], Uéverton dos Santos Souza[1]

[1]Instituto de Computação – Universidade Federal Fluminense
Av. General Milton Tavares de Souza s/n – Niterói, RJ, Brazil
{mmaia,plastino,ueverton}@ic.uff.br

[2]Escola Nacional de Ciências Estatísticas – Instituto Brasileiro de Geografia e Estatística
R. General Canabarro, 706 – Rio de Janeiro, RJ, Brazil
marcelo.h.maia@ibge.gov.br

ABSTRACT

This text summarizes the identically titled doctoral thesis, which compiles a series of research contributions on the proposal and application of data mining methods in the combinatorial optimization and bioinformatics fields. It explores traditional methods and, mainly, a novel approach, named MineReduce, for incorporating data mining into metaheuristics, applied to solve various problems. The derived methods achieved relevant results, outperforming state-of-the-art algorithms and finding new best solutions for several benchmark instances of the addressed problems. Additionally, two proposed methods were awarded first place in the MESS 2020+1 metaheuristics competition and second place in the capacitated vehicle routing problem track of the 12th DIMACS Implementation Challenge. The thesis also reports methodological contributions related to classification and their applications in bioinformatics. It introduces novel approaches to build classifier ensembles for uncertain data, applied for classifying ageing-related genes and predicting drug side effects. The results evidence that the proposed approaches improve the predictive performance of ensembles on uncertain data. Additionally, novel approaches for interpreting Naive Bayes ensembles were introduced and applied to identify relevant features to classify genes, producing consistent results and new insights in this field.

KEYWORDS. Data mining. Metaheuristics. Problem size reduction.

Paper topics. Combinatorial Optimization. Computational Intelligence.

http://live-proceedings.pantheonsite.io/p/174820

1. Introduction

Research on the synergies between metaheuristics and data science – which relates to fields like data mining (DM) and machine learning (ML) – has become increasingly popular in the combinatorial optimization literature [Talbi, 2021]. One particularly successful approach in this context involves incorporating data mining techniques into metaheuristics, leading to solution quality improvement and convergence speedup [Martins et al., 2018]. This thesis [Maia, 2022] introduces and explores a novel approach, named MineReduce, for incorporating data mining into metaheuristics, presenting its well-succeeded application to solve variants of the vehicle routing, vertex cover, facility location and travelling salesperson problems. The heuristics developed based on the MineReduce approach achieved relevant results, outperforming previous state-of-the-art algorithms and finding new best solutions for several instances of the addressed problems. Additionally, this thesis presents the application of a similar hybrid data mining approach – the Multi Data Mining (MDM) approach [Plastino et al., 2014] – to solve a vehicle routing problem variant, which also achieved state-of-the-art results. The MineReduce and MDM approaches are based on frequent itemset mining, a descriptive data mining task.

This thesis also reports methodological contributions related to the classification task and their applications in bioinformatics. Specifically, it introduces novel approaches for coping with uncertain data and for model interpretation. Previous work has shown that proper handling of uncertainty can improve predictive performance, but this is still an under-explored research topic, particularly for categorical features. This thesis introduces novel approaches for building ensembles that cope with uncertain categorical features. They were applied for classifying ageing-related genes and predicting drug side effects using real data. The results showed that the proposed approaches improved the predictive performance of ensembles. Interpretable modelling is another emerging topic of high importance in the ML field since understanding how and why computational models make predictions may be critical for some applications. Besides, it can help domain experts learn more about the considered problem. In particular, interpretable models have been applied in genetics and genomics to derive novel biological insights [Azodi et al., 2020]. This thesis also introduces novel approaches for interpreting Naive Bayes ensembles, which were applied to identify relevant protein interactions to classify ageing-related genes, producing results consistent with current biological knowledge and, most important, new insights in this field.

The following sections will present a summary of the studies in this thesis. Section 2 discusses this thesis' research contributions to the combinatorial optimization field. The proposal of novel classifier ensemble approaches and their application in bioinformatics domains are reported in Section 3. Section 4 lists products derived from the thesis and the awards received in recognition of this work's achievements.

2. Advances in the Hybridization of Metaheuristics with Data Mining

In this context, a novel general approach named MineReduce was proposed. Its application to various combinatorial optimization problems demonstrated its effectiveness in providing metaheuristic methods with enhanced performance, achieving state-of-the-art results regarding solution quality and convergence speed. Thus, the MineReduce approach constitutes a major methodological contribution to the combinatorial optimization field, whereas its applications constitute practical contributions. An additional practical contribution was obtained by applying the MDM approach to a vehicle routing problem variant, also achieving state-of-the-art results. Section 2.1 introduces the MineReduce approach (thesis Chapter 2), whereas Section 2.2 briefly reports the applications (thesis Chapters 3–7).

2.1. The MineReduce Approach

Any approach for solving a combinatorial optimization problem (COP) relies on some form of search on its solution space, which is the domain of the function to be optimized. Solving a COP is usually challenging because its solution space grows exponentially with its size. Therefore, problem size reduction (PSR) is beneficial in this context, especially in the case of large-scale COPs.

If a COP is analyzed with respect to an integer programming formulation, its size is defined considering the number of decision variables, the cardinality of the decision variables' domains, and the number of constraints. PSR techniques aim to reduce the number of decision variables of the problem or the range of values in their domains.

The general process of PSR techniques consists of the following steps:

1. Transforming a problem instance P into a problem instance P' such that the solution space of P' is smaller than that of P.
2. Solving P'.
3. Transforming the solution to P' into a solution to P.

The first step (the reduction itself) is naturally the most crucial procedure since better decisions in this step will produce solutions of better quality in the subsequent ones. Strategies to make these decisions are highly dependent on the structure and features of the problem, so they vary significantly among distinct classes of COPs.

One general form of accomplishing size reduction is fixing values of decision variables, which can be regarded as fixing elements either in or out of the solution.

2.1.1. Approach Description

Previous approaches that incorporate data mining into multi-start metaheuristics eliminate the independence of their iterations by introducing a memory mechanism [Martins et al., 2018]. At some point, iterations benefit from knowledge accumulated in previous iterations. This knowledge – patterns mined from an elite set of solutions – is used to generate better initial solutions.

The data mining procedure in these approaches mines maximal frequent itemsets [Grahne e Zhu, 2003]. Hence, mined patterns are composed of items that frequently appear together in the sub-optimal solutions of the elite set. Intuitively, these items should likely be part of the best solutions to the problem. Thus, they are included in initial solutions.

The MineReduce approach builds upon these ideas. Since the mined patterns are assumed to likely be part of the best solutions to a problem instance, they are well-suited for reducing the size of that instance by fixing or merging pattern items in the solution.

Fig. 1 presents a conceptual framework of the MineReduce approach, showing the steps that compose its PSR process. The proposed approach requires as input an *elite set* of solutions to the problem instance in consideration. A data mining method extracts a set of patterns from the elite set. Then, a pattern p is selected for driving the PSR process. The *Reduce* step uses pattern p to transform problem instance I into a reduced-size instance I'. In the *Optimize* step, a solution to I' is obtained. Finally, in the *Expand* step, the solution to I' is transformed into a solution to I.

The three final steps compose a complete PSR process conceived to work as a method for generating new solutions. For example, in a multi-start metaheuristic, at each iteration occurring after a set of patterns is mined, one of the patterns would be selected for carrying out the PSR process, producing a new initial solution.

The *Optimize* step may be accomplished by applying any method that produces a feasible (presumably good) solution for the reduced instance. For example, the original metaheuristic's

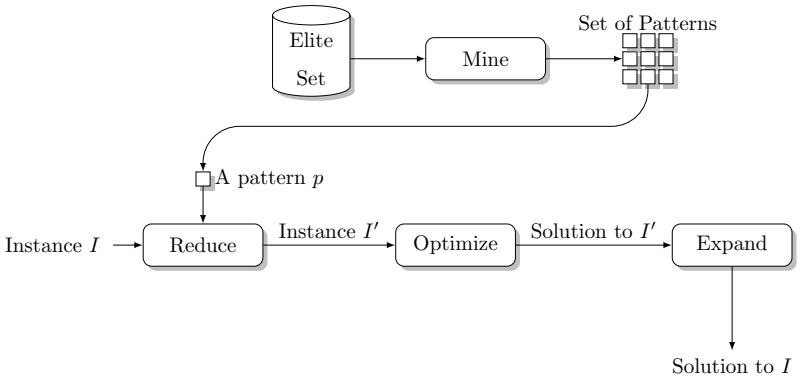

Figure 1: Conceptual framework of the MineReduce approach.

optimization procedures (e.g., solution initialization and local search) may be used, or an exact method may be applied to solve the reduced instance optimally if it is small enough.

A taxonomy for metaheuristics that incorporate ML in their design was presented in [Talbi, 2021]. According to that taxonomy, MineReduce-based methods are primarily classified as *problem-level* ML-supported metaheuristics since this approach uses data mining for hierarchical problem decomposition (defining and solving smaller subproblems). They can also be classified as *low-level* ML-supported metaheuristics since data mining is used in a process that drives the initialization of solutions. Finally, regarding the *learning time* criteria, they are classified as *online* ML-supported metaheuristics since they gather knowledge during the search while solving the problem.

2.2. Applications

In this thesis, the MineReduce approach was applied in the proposal of metaheuristics for the heterogeneous fleet vehicle routing problem (HFVRP), minimum weighted vertex cover problem (MWVCP), multi-source facility location problem with customer incompatibilities (MS-CFLP-CI), and minimum latency problem (MLP).

The metaheuristics proposed for the HFVRP (thesis Chapter 3), MWVCP (thesis Chapter 4) and MLP (thesis Chapter 6) were designed by incorporating the MineReduce approach into previous state-of-the-art methods from the literature. They achieved superior solution quality and convergence speed, outperforming the original algorithms and other state-of-the-art methods.

A methodology for assessing PSR effectiveness by interpreting it as a binary classification problem is also proposed (thesis Chapter 4). This allows using standard classification metrics like precision, recall and f-measure to assess essential factors of PSR: optimality (whether a reduced instance contains an optimal solution to the original one) and degree of reduction (how smaller a reduced instance is). This methodology was applied to compare the reduction provided by MineReduce and that provided by a kernelization [Downey e Fellows, 1999] algorithm. The results showed that MineReduce achieved an average precision of over 99% while fixing significantly more elements in the solution (higher recall).

The MS-CFLP-CI is a new variant of the facility location problem proposed in the MESS 2020+1 Competition[1]. A MineReduce-based metaheuristic designed from scratch was proposed for this problem (thesis Chapter 5), ranking first in the competition.

Additionally, a metaheuristic was proposed for the capacitated vehicle routing problem (CVRP), designed by incorporating the MDM approach into a previous state-of-the-art method from

[1] https://www.ants-lab.it/mess2020/#competition

the literature (thesis Chapter 7). The proposed method outperformed the original one regarding solution quality and convergence speed. Furthermore, it was a finalist in the CVRP track of the 12th DIMACS Implementation Challenge[2]. Besides being a close second in the overall ranking, the proposed method was the best for some classes of problem instances.

3. Advances in Classification Methods and their Application in Bioinformatics

In this context, novel approaches were proposed for coping with uncertain data and model interpretation. Their application to real data from bioinformatics domains demonstrated their effectiveness in providing ensembles of classifiers with enhanced predictive performance and interpretability. Thus, the proposed approaches constitute methodological contributions to the computational intelligence field, whereas the results obtained constitute contributions to the bioinformatics field. Section 3.1 discusses three approaches proposed for building classifier ensembles for uncertain categorical data (thesis Chapter 8), whereas Section 3.2 discusses two approaches proposed for adding interpretability to ensembles of NB classifiers (thesis Chapter 9).

3.1. Ensembles of Classifiers for Uncertain Data

Feature-value uncertainty occurs when some feature values in an instance are not precisely known. Previous work has shown that incorporating information on uncertainty into classification algorithms can improve predictive performance. However, this is still an under-explored research topic, particularly for categorical features. This work focuses on ensemble methods that learn many base classifiers independently on random subsets of the original training set and then aggregate the base classifiers' predictions. In particular, Bagging methods randomly sample subsets of the instances in the dataset with replacement, which is called bootstrap sampling, and Random Subspaces methods randomly sample subsets of the features in the dataset.

3.1.1. Proposed Approaches

This work proposes new approaches for building ensembles that incorporate uncertainty about the value of categorical features into the model. The intuition motivating these proposals is that the higher the uncertainty for a given feature (or instance), the less it might contribute to the predictive performance as it provides less reliable information. A bias value is computed for each feature f_j based on its degree of uncertainty and fraction of missing values in the dataset, given by:

$$b_{*j} = \left(1 - \frac{1}{|I \setminus M_{*j}|} \sum_{i \in I \setminus M_{*j}} E_{ij}\right) \times \frac{|I \setminus M_{*j}|}{|I|}$$

where $I = \{1, 2, \ldots, n\}$ is the set of indices of all instances in the dataset, M_{*j} is the set of indices of instances with a missing value for feature f_j, and E_{ij} is the normalized entropy of the probability distribution associated with feature f_j for instance i.

The computed bias is a value in the $[0, 1]$ range, with higher values indicating lower uncertainty degrees, i.e., more reliable features. The feature bias values are normalized over all features, defining a probability distribution $B = (\beta_1, \beta_2, \ldots, \beta_m)$, where a probability β_j associated with a feature f_j is given by $\beta_j = b_{*j}/(\sum_{l=1}^{m} b_{*l})$. The first proposed approach, named *Biased Random Subspaces* (BRS), uses the probability distribution B to sample the features to be considered by each base classifier in the ensemble instead of the default uniform distribution.

Random Forests usually do not sample features before generating each tree. Nonetheless, they sample a subset of features to be considered candidate features when splitting each tree node.

[2] http://dimacs.rutgers.edu/programs/challenge/vrp/cvrp

Hence, another approach, named *Biased Splitting* (BS), is proposed for building Random Forests. As the BRS approach, it uses the probability distribution B to sample the candidate features.

Finally, a third approach, named *Biased Bootstrap* (BB), is proposed for instance sampling analogously to the proposal of the BRS and BS approaches for feature sampling.

3.1.2. Applications

The proposed approaches have been evaluated on real uncertain data from two domains. The first domain is the classification of ageing-related genes regarding their effect on the lifespan of an organism, which may be positive (pro-longevity) or negative (anti-longevity). Four new datasets have been generated for this domain by integrating data from the GenAge and the STRING databases. The second domain involves the prediction of drugs' side effects. Six new datasets have been generated for this domain by integrating data from the SIDER and the STITCH databases.

In the experiments, three baseline ensemble methods were considered: two kinds of ensembles of Naive Bayes (NB) classifiers and one Random Forest (RF). The baseline ensembles use conventional strategies for sampling instances and features. Three versions have been built for each by incorporating different combinations of the proposed approaches: BB, BRS and BB+BRS for NB ensembles; BB, BS and BB+BS for RFs. The predictive performance of the algorithms has been assessed using two metrics: the Area Under the Receiver Operating Characteristic curve (AUROC) and the geometric mean of sensitivity and specificity (G-mean).

The results support the hypothesis that the BB, BRS and BS approaches improve the predictive performance of ensembles on uncertain data. The BB and BS approaches obtained the best overall results when combined with the RF baseline.

3.2. Interpretability Approaches for Naive Bayes Ensembles

3.2.1. Proposed Approaches

Given a feature vector $X_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ associated with an unlabelled instance identified by index i, an NB classifier predicts the class $y \in C$ that maximizes the value given by $P(y|X_i) \propto P(y) \prod_{j=1}^{m} P(x_{ij}|y)$. In the first approach proposed in this work, the importance of a feature value x_{ij} in a given binary NB classifier e in the ensemble is estimated by the following difference of conditional probabilities:

$$Diff(x_{ij}, c_1, c_2, e) = |P(x_{ij}|c_1) - P(x_{ij}|c_2)|$$

For datasets with more than two class labels ($|C| = q > 2$), this idea can be generalized by summing the differences between all pairs of class labels in C:

$$Importance(x_{ij}, C, e) = \sum_{r=1}^{q-1} \sum_{s=r+1}^{q} Diff(x_{ij}, c_r, c_s, e)$$

The importance of a feature value x_{ij} for an ensemble of NB classifiers is computed by averaging its importance across all classifiers in the ensemble. Finally, once the importance value has been computed for all feature values x_{ij}, all feature values are ranked in decreasing order of importance. Then a user (domain expert) can focus on interpreting the top-ranked feature values, i.e., the most important ones for predicting the class variable in the ensemble.

The approach based on conditional probability differences measures the importance of each feature value separately, ignoring its importance in the context of all other feature values. However, whether or not a conditional probability will make a difference in the choice of the predicted class depends on the entire set of conditional probabilities and the prior class probability.

Hence, a second approach that considers sets of feature values is proposed. The aim is to find, for each instance, a minimal set of features sufficient to preserve the class prediction, such that changes to the other feature values of the instance would not change the class predicted by the model. An algorithm for identifying a minimal sufficient set of features for the class prediction for a given instance is proposed in this thesis. Based on the sufficiency criterion, a measure of the importance of a feature f_j for a classifier e given the set of instances X, denoted $SImportance(f_j, e, X)$, is defined as the proportion of instances in X for which f_j is in the minimal sufficient set.

3.2.2. Applications

The proposed approaches were applied to the ageing-related datasets mentioned in Section 3.1 to identify the top-ranked features for classification. The top-ranked features were analyzed regarding their relevance to the biology of ageing by Prof. João Pedro de Magalhães, a microbiologist who leads the Genomics of Ageing and Rejuvenation Lab at the University of Birmingham, whose research broadly focuses on understanding the genetic, cellular, and molecular mechanisms of ageing. According to his analysis, the top-ranked features fit well with current knowledge on longevity/ageing and previous similar analyses. Furthermore, new insights into the genetics of ageing can be drawn from the rankings, indicating interesting targets for future biological studies.

4. Work Products and Awards

Seven research papers derived from this thesis have been published so far, and other two are under review. Additionally, software and datasets produced have been publicly released[3].

4.1. Awards

Theses' Contests

1st place – PhD Theses Contest – 8th IEEE Latin American Conf. on Computational Intelligence.

3rd place – 13th National Contest of PhD Theses on Artificial and Computational Intelligence.

Implementation Contests

1st place – Metaheuristics Competition – International Metaheuristics Summer School (MESS 2020+1).

2nd place – 12th DIMACS Implementation Challenge: Vehicle Routing Problems – CVRP Track.

4.2. Publications

Journal articles:

1. Maia et al., 2020. "MineReduce: An approach based on data mining for problem size reduction", *Computers & Operations Research*
https://doi.org/10.1016/j.cor.2020.104995

2. Maia et al., 2023. "Metaheuristic techniques for the capacitated facility location problem with customer incompatibilities", *Soft Computing*
https://doi.org/10.1007/s00500-022-07600-z

3. Maia et al., 2023. "Interpretable Ensembles of Classifiers for Uncertain Data with Bioinformatics Applications", *IEEE/ACM Transact. on Computational Biology and Bioinformatics*
https://doi.org/10.1109/TCBB.2022.3218588

[3] https://github.com/marcelorhmaia, https://github.com/MESS-2020-1

Conference papers:

4. Maia et al., 2023. "MineReduce-Based Metaheuristic for the Minimum Latency Problem", in *14th Metaheuristics International Conference (MIC'2022)*, *LNCS* https://doi.org/10.1007/978-3-031-26504-4_7

5. Maia et al., 2021. "An Ensemble of Naive Bayes Classifiers for Uncertain Categorical Data", in *2021 IEEE International Conference on Data Mining (ICDM)* https://doi.org/10.1109/ICDM51629.2021.00148

6. Maia et al., 2020. "MineReduce for the minimum weight vertex cover problem", in *Proceedings of the International Conference on Optimization and Learning (OLA'2020)*

Challenge workshop paper:

7. Maia et al., 2022. "An Improved Hybrid Genetic Search with Data Mining for the CVRP", in *12th DIMACS Implementation Challenge: Vehicle Routing Problems* http://dimacs.rutgers.edu/events/details?eID=2073

Acknowledgment

The author was supported by: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) [finance code 001]; and Instituto Brasileiro de Geografia e Estatística (IBGE). This work was enabled in part by computational resources and support provided by Calcul Québec (calculquebec.ca), the Digital Research Alliance of Canada (alliancecan.ca) and UFF's Laboratório de Inteligência Computacional (LabIC). This support is gratefully acknowledged.

References

Azodi, C. B., Tang, J., e Shiu, S.-H. (2020). Opening the black box: Interpretable machine learning for geneticists. *Trends Genet.*, 36(6):442–455. ISSN 0168-9525.

Downey, R. G. e Fellows, M. R. (1999). *Some Ad Hoc Methods: The Methods of Bounded Search Tree and Problem Kernel*, p. 29–48. Springer, New York. ISBN 978-1-4612-0515-9.

Grahne, G. e Zhu, J. (2003). Efficiently using prefix-trees in mining frequent itemsets. In Goethals, B. e Zaki, M. J., editors, *Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations*.

Maia, M. R. H. (2022). *Novel Data Mining Methods and Applications in Combinatorial Optimization and Bioinformatics*. D.Sc. thesis, Universidade Federal Fluminense, Niterói, RJ, Brazil. URL http://www.ic.uff.br/PosGraduacao/frontend-tesesdissertacoes/download.php?id=1053.pdf&tipo=trabalho.

Martins, S. L., Rosseti, I., e Plastino, A. (2018). *Data Mining in Stochastic Local Search*, p. 39–87. Springer, Cham. ISBN 978-3-319-07124-4.

Plastino, A., Barbalho, H., Santos, L. F. M., Fuchshuber, R., e Martins, S. L. (2014). Adaptive and multi-mining versions of the DM-GRASP hybrid metaheuristic. *J. Heuristics*, 20:39–74. ISSN 1572-9397.

Talbi, E.-G. (2021). Machine learning into metaheuristics: A survey and taxonomy. *ACM Comput. Surv.*, 54(6). ISSN 0360-0300.