A Data-Science Pipeline to Enable the Interpretability of Many-Objective Feature Selection

Uchechukwu F. Njoku[1,2,*], Alberto Abelló[1], Besim Bilalli[1] and Gianluca Bontempi[2]

[1]Universitat Politècnica de Catalunya, Jordi Girona 31, 08034 Barcelona, Spain
[2]Université Libre de Bruxelles, Franklin Roosevelt 50, 1050 Brussels, Belgium

Abstract

Many-Objective Feature Selection (MOFS) approaches use four or more objectives to determine the relevance of a subset of features in a supervised learning task. As a consequence, MOFS typically returns a large set of non-dominated solutions, which have to be assessed by the data scientist in order to proceed with the final choice. Given the multi-variate nature of the assessment, which may include objectives (e.g., fairness) unrelated to predictive accuracy, this step is often not straightforward and suffers from the lack of existing tools. For instance, it is common to make use of a tabular presentation of the solutions, which provides little information about the trade-offs and the relationships between objectives over the set of solutions.

Adopting a GA-based MOFS with six objectives (number of selected features, balanced accuracy, F1-Score, variance inflation factor, statistical parity, and equalised odds) for two feature selection tasks, this paper illustrates the complex challenge of assessing MOFS results and the need for a methodology to aid and justify the final choice of a solution.

Keywords

Feature selection, Many-objective optimisation, Genetic algorithm, Interpretability, Fairness

1. Introduction

Data scientists are increasingly confronted with datasets of huge size and dimensionality. Effective Feature Selection (FS) is then more and more important in order to extract *valuable information* by identifying relevant features and discarding irrelevant or redundant ones.

Given a supervised task with m input features, the number of possible subsets is exponential (exactly $2^m - 2$, excluding the empty and full set). FS denotes the techniques to find one or more subsets of features (in the following, also called solutions) containing the most relevant features based on certain objectives.

The set of objectives is highly dependent on the task and goal of the data analysis. For example, if we aim at predicting hospital readmission of patients with diabetes [1], the objectives could include generalisation-related measures (e.g., size of the subset, predictive performance, redundancy) as well as fairness. The inclusion of fairness objectives is more and more required in the application of artificial intelligence (AI) to domains involving confidential information and having a potential impact on citizens' rights and welfare [2] (e.g., healthcare).

Moving from one to several objectives requires adapting the FS strategy to account for the trade-offs among the considered objectives. In particular, many-objective FS methods return a set of non-dominated solutions, among which the data scientist is expected to select the most convenient one. In the case of two objectives, the set of solutions can be illustrated with 2D line plots [3], scatter plots [4], bar charts [5], or simply tables [6]. With three objectives, it becomes more complex to present the solutions on a single graph, requiring 3-D scatter plots [7], multiple charts [8] or tables [9]. When we move to four or more objectives (MOFS), it is even harder for the data analyst to extract some useful insights.

The few recent works on MOFS solutions are either based on a large number of descriptive tables [10] or rely on a single measure [11, 10, 12] hiding the multi-variate nature of the objective space to present the solutions. Thus, after the massive computational effort spent generating interesting solutions, the final selection does not take advantage of any interpretable and user-friendly exploration mechanism [13].

In this work, we highlight the need to extend the classical data science pipeline with a MOFS interpretation module (Figure 1) to enable effective exploration of the MOFS set of solutions by the data scientist. We do this by considering two classification tasks with six objectives (i.e., the subset size, balanced accuracy, F1-score, Variance Inflation Factor – V.I.F, statistical parity, and equalised odds). We use a Genetic Algorithm (GA), namely NSGA-III [14], as a search strategy to generate the set of solutions with Naive Bayes (NB) and Logistic Regression (LR) as classifiers.

DOLAP 2024: 26th International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data, co-located with EDBT/ICDT 2024, March 25, 2024, Paestum, Italy
*Corresponding author.
uchechukwu.fortune.njoku@upc.edu (U. F. Njoku); alberto.abello@upc.edu (A. Abelló); besim.bilalli@upc.edu (B. Bilalli); gianluca.bontempi@ulb.be (G. Bontempi)
0000-0002-2599-9645 (U. F. Njoku); 0000-0002-3223-2186 (A. Abelló); 0000-0002-0575-2389 (B. Bilalli); 0000-0001-8621-316X (G. Bontempi)

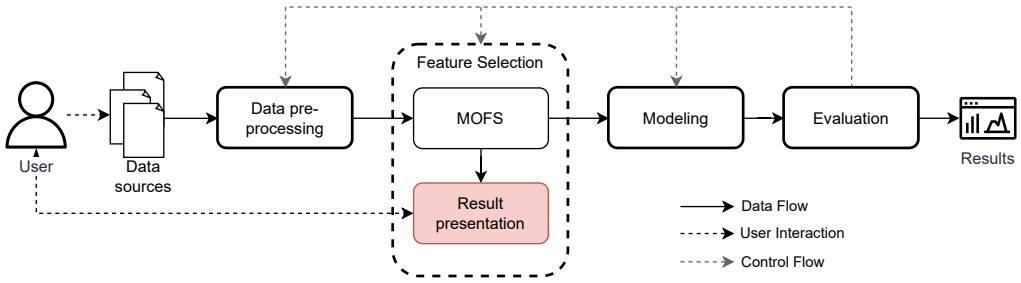

Figure 1: An end-to-end data science process.

2. Related Work

Many-objective optimisations are problems that require the simultaneous optimisation of four or more objectives. The goal is to find diverse solutions that lie close to the true Pareto font [15]. On the other hand, FS aims to find a subset of features from a dataset that is most relevant to the task at hand; a key part of this is defining their relevance, which cannot be fully described by a single objective [16]. Depending on the task, it requires considering up to four or more objectives, termed MOFS.

MOFS has been used to optimize multiple objectives in various applications, such as network anomaly detection [10], motor imagery prediction [17], multi-label classification [18], and large-scale feature selection [19]. However, there is still a need for research on how to choose a final solution from the large set of non-dominated solutions.

The mechanisms used to present and choose a final solution in previous works include box plots [10], bar charts [17], scatter plots [18], Parallel Coordinate Plots (PCP) [19], and tables [10, 17, 18]. Additionally, measures such as the hypervolume indicator [17] and inverse generation distance — which is the distance from the ideal solution vector [10], are also used. However, all these ignore one or more viewpoints (perspectives), especially the feature viewpoint, unique to the FS problem. MOFS requires collective (solutions) and individual (features) evaluation of the resulting set of solutions to find the most appropriate one.

In many-objective optimisation, providing support for the data scientist remains challenging because the commonly used mechanisms for presenting the solutions become inadequate with the growing number of dimensions. To remedy this, the dimensions can be reduced, or multiple presentations can be used for completeness [20]. Yet, the former comes with a loss of information, and the latter requires clear guidelines.

It is, therefore, imperative to design a methodology that facilitates the interpretability of MOFS results holistically, taking into account all three viewpoints that make up the process: objectives, solutions, and features.

3. METHODOLOGY

We apply the MOFS method to two well-known fairness datasets. The MOFS simultaneously optimises six objectives: subset size, balanced accuracy, F1-score, VIF, statistical parity, and equalised odds. We use NSGA-III as the search method to find a set of non-dominated solutions. This section presents NSGA-III, the MOFS implementation, datasets used in this work, and the execution details for reproducibility.

3.1. NSGA-III

NSGA-III is a pareto and reference-based elitist GA for many-objective problems [14]. It uses well-spread-out reference points to maintain population (i.e., a set of candidate solutions) diversity. Also, it is elitist because it is designed to preserve the set of best individuals at each iteration. Thus, NSGA-III begins by initialising a population and optimises towards better solutions until a termination criterion is satisfied.

3.2. MOFS implementation

We describe the starting point (i.e., initialisation of the first population), search strategy, feature subset evaluation (i.e., the objectives), and the termination criterion for the MOFS used in this work.

3.2.1. Starting Point

We begin with a population of one-sized candidate solutions, where each feature is selected at least once. The population size p is set as an even number greater than the overall number of features m. So, $p = m + 1$ if m is odd. Otherwise, $p = m + 2$.

3.2.2. Search Strategy and Subsets Evaluation

We use NSGA-III with several parameters as the search strategy for MOFS. In particular, we set the mutation probability to $1/m$ and the crossover probability to 1.

Knowing that one measure of relevance for a feature subset is insufficient, we evaluate the relevance of a subset of features by up to six objectives:

1. The **subset size** is a fundamental objective of FS, and the goal is to minimise it (↓).
2. **Balanced accuracy** measures accurate predictions. It is a commonly used predictive performance metric in classification problems, and we aim to maximise it (↑).
3. **F1-Score**, which is the harmonic mean of precision and recall [16], is another predictive performance measure which we maximise (↑).
4. **Variance Inflation Factor (VIF)** measures multicollinearity between features [21]. A high VIF implies redundancy, and we want to minimise it (↓).
5. **Statistical parity** checks if classifier predictions are void of group membership sentiments [2], and we aim to maximise it (↑).
6. **Equalised odds** measures if the classifier performs equally well across sensitive groups [2], and we aim to maximise it (↑).

Therefore, the goal is to find minimal, uncorrelated feature subsets that provide accurate and fair predictions.

3.2.3. Termination Condition

In this work, we set the termination criterion to a given maximum number of evaluations, which is set up to $2p^2$. When this condition is satisfied, the non-dominated [14] solutions in the current population are returned.

3.3. Datasets

Table 1 shows the datasets' details. Diabetes[1] for predicting hospital readmission of diabetes patients, and German credit[2] for predicting credit risk of clients.

Table 1
Dataset properties.

Name	#Features	#Instances	#Classes	Sensitive	Pop. Size	Max. Evals.
Diabetes	57	101,766	2	Gender	58	6728
German credit	21	1000	2	Sex	22	968

3.4. Execution

We implemented the experiments using Python 3 and particularly, the *jMetalPy*[3] library implementation of NSGA-III [22]. The source code for our experiments is available on GitHub.[4]

[1]https://doi.org/10.24432/C5230J
[2]https://doi.org/10.24432/C5QG88
[3]https://github.com/jMetal/jMetalPy
[4]https://github.com/F-U-Njoku/many-objective-fs-nsgaiii

4. Results and Discussion

For both datasets used in this work, we scrutinise and present their solutions below.

4.1. Diabetes

For this dataset, we try to predict hospital readmission of patients with Diabetes using the NB classifier, and the sensitive feature on which we strive for fairness is *gender*. The MOFS produced 52 non-dominated solutions, and the ranges of values for the objectives are as follows: subset size $[7, 19]$, balanced accuracy $[0.5049, 0.5422]$, F1-score $[0.0456, 0.1790]$, VIF $[0.0093, \infty]$, statistical parity $[0.8432, 0.9246]$, and equalised odds $[0.7130, 0.8403]$.

With 52 solutions for the Diabetes dataset, this becomes overwhelming to depict with a table. Also, with greater subset sizes, it becomes impossible to fully display all features that have been selected for each solution. Using a table, in this case, gives little support to the data scientist in choosing a final solution.

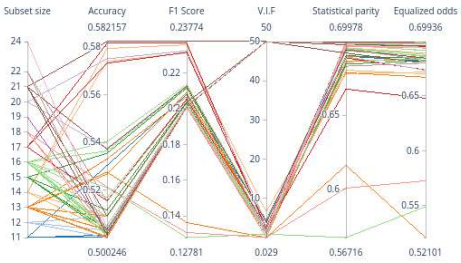

Figure 2: PCP presentation for Diabetes dataset results.

A more sophisticated way to present MOFS results is the Parallel Coordinate Plot (PCP) shown in Figure 2. For each solution, PCP shows the values for each objective, and it is typically interactive (i.e., a range of values can be chosen for each objective, and the solutions that satisfy the configuration will be filtered). With this, the data scientist can find solutions that satisfy the given criteria. However, it is unlikely that only one solution satisfies this configuration. Whether this returns many or a few solutions, there is still a need to compare alternatives and with more alternatives, the complexity increases. Also, the PCP does not show the content of the solutions (i.e., the features that have been chosen). Finally, as shown in Figure 2, with more solutions, the PCP becomes overcluttered.

	Solution	Size	Acc	F1	VIF	SP	EO
1.	laufkont	1	0.6040	0.7808	0.0000	0.8058	0.2122
2.	beszeit	1	0.5000	0.8235	0.0000	1.0000	0.8000
3.	moral	1	0.5310	0.8253	0.0000	0.9096	0.6592
4.	hoehe, beszeit	2	0.5202	0.8181	2.0244	0.9199	0.6301
5.	moral, beszeit	2	0.5338	0.8246	4.6076	0.8982	0.5976
6.	laufkont, bishkred	2	0.5888	0.7803	3.4792	0.7862	0.2686
7.	laufkont, laufzeit	2	0.6129	0.8192	2.3920	0.7671	0.3375
8.	moral, sparkont, wohnzeit	3	0.5548	0.8305	3.8364	0.8793	0.6035
9.	laufkont, wohnzeit, pers	3	0.5769	0.7776	6.3914	0.7804	0.3100
10.	moral, beszeit, rate, gastarb	4	0.5357	0.8206	9.8773	0.8961	0.6367
11.	hoehe, beszeit, alter, bishkred	4	0.5214	0.8204	6.1980	0.9129	0.6531
12.	laufkont, laufzeit, moral, verw	4	0.6255	0.8149	3.4898	0.7636	0.3011
13.	beszeit, beruf, telef, gastarb	4	0.5002	0.8226	16.6683	0.9817	0.7741
14.	moral, beszeit, alter, bishkred	4	0.5417	0.8254	8.1784	0.9017	0.6249
15.	laufkont, laufzeit, hoehe, beszeit, alter	5	0.6062	0.8171	6.1966	0.7537	0.3417
16.	moral, buerge, verm, alter, weitkred, bishkred	6	0.5593	0.8268	7.8315	0.8840	0.5892
17.	laufkont, laufzeit, moral, sparkont, beszeit, rate, wohnzeit, weitkred	8	0.6474	0.8254	6.7702	0.7758	0.3541
18.	laufkont, laufzeit, moral, sparkont, beszeit, rate, bishkred, gastarb	8	0.6607	0.8297	8.7166	0.7637	0.3491
19.	laufkont, laufzeit, moral, verw, hoehe, sparkont, famges, verm, telef, ...	10	0.6474	0.8290	8.8106	0.7488	0.4174

Figure 3: Tabular presentation for German credit dataset results.

4.2. German credit

The task for this dataset is to predict the credit risk of bank account holders using an LR classifier, having *famges (Sex)* as a sensitive feature. MOFS produced 19 non-dominated solutions where the range of values for each objective is: subset size $[1, 10]$, balanced accuracy $[0.5000, 0.6607]$, F1-score $[0.7776, 0.8305]$, VIF $[0, 16.668]$, statistical parity $[0.7488, 1]$, and equalised odds $[0.2122, 0.8000]$. Most commonly, MOFS results are presented in descriptive tables, such as the one shown in Figure 3, which shows all 19 MOFS results for the German Credit dataset. In most cases, we can see all the features in each solution except for the nineteenth solution with 10 features. Looking at the objectives, one after the other (since that is all the support the table provides), solutions one to three offer the smallest subset size (1), solution 18 has the best balanced accuracy (0.6607), solution eight has the highest F1-score (0.8305), solutions one to three offer the smallest VIF (0), and finally, solution two has the best statistical parity (1) and equalised odd (0.8) scores. Solution two (*beszeit*) appears to have a better score for four out of six objectives. However, building a model with one feature is not a good idea. Therefore, we must continue comparing alternatives until a satisfactory solution is found.

Indeed, a better, more informative and systematic methodology will accelerate this process and make it easier for the data scientist.

5. Conclusion

MOFS produces a potentially large set of solutions with completely different values for the considered objectives. Thus, we have shown the need for a methodology to systematically analyse the results obtained by MOFS through two concrete fairness benchmark datasets. The results of our research framed under the studies performed in the project DEDS (MSCA-ITN G.A. No. 955895) allows us to conclude that MOFS is required in practical use cases like those requiring fairness; however, it still lacks methodologies and tools for interpretability.

Future work includes developing methodologies for MOFS interpretability and incorporating them into FS tools for accessibility.

6. Acknowledgements

The project leading to this publication has received funding from the European Union's Horizon 2020 research and innovation programme (G.A. No. 955895). A. Abelló and B. Bilalli are funded by the Spanish Ministerio de Ciencia e Innovación under project PID2020-117191RB-I00, funding scheme AEI/10.13039/501100011033. G. Bontempi is supported by the Service Public de Wallonie Recherche under G.A. No. 2010235–ARIAC by Digital-Wallonia4.ai. Computational resources have been provided by the Consortium des Équipements de Calcul Intensif, funded by the Fonds de la Recherche Scientifique de Belgique (G.A. No. 2.5020.11) and Walloon Region.

References

[1] J. Clore, K. Cios, J. DeShazo, B. Strack, Diabetes 130-US hospitals for years 1999-2008, UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5230J.

[2] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, E. Ntoutsi, A survey on datasets for fairness-aware machine learning, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 12 (2022) e1452.

[3] E. Hancer, B. Xue, M. Zhang, D. Karaboga, B. Akay, Pareto front feature selection based on artificial bee colony optimization, Information Sciences 422 (2018) 462–479.

[4] Y. Zhang, S. Cheng, Y. Shi, D.-w. Gong, X. Zhao, Cost-sensitive feature selection using two-archive multi-objective artificial bee colony algorithm, Expert Systems with Applications 137 (2019) 46–58.

[5] J. Grzyb, M. Topolski, M. Woźniak, Application of multi-objective optimization to feature selection for a difficult data classification task, in: International Conference on Computational Science, Springer, 2021, pp. 81–94.

[6] Z. Liu, B. Chang, F. Cheng, An interactive filter-wrapper multi-objective evolutionary algorithm for feature selection, Swarm and Evolutionary Computation 65 (2021) 100925.

[7] Y. Xue, Y. Tang, X. Xu, J. Liang, F. Neri, Multi-objective feature selection with missing data in classification, IEEE Transactions on Emerging Topics in Computational Intelligence 6 (2021) 355–364.

[8] P. Barbiero, E. Lutton, G. Squillero, A. Tonda, A novel outlook on feature selection as a multi-objective problem, in: International conference on artificial evolution (evolution artificielle), Springer, 2019, pp. 68–81.

[9] Y. Zhou, J. Kang, S. Kwong, X. Wang, Q. Zhang, An evolutionary multi-objective optimization framework of discretization-based feature selection for classification, Swarm and Evolutionary Computation 60 (2021) 100770.

[10] Z. Zhang, J. Wen, J. Zhang, X. Cai, L. Xie, A many objective-based feature selection model for anomaly detection in cloud environment, IEEE Access 8 (2020) 60218–60231.

[11] K. Jha, S. Saha, Incorporation of multimodal multiobjective optimization in designing a filter based feature selection technique, Applied Soft Computing 98 (2021) 106823.

[12] D. Rodrigues, V. H. C. de Albuquerque, J. P. Papa, A multi-objective artificial butterfly optimization approach for feature selection, Applied Soft Computing 94 (2020) 106442.

[13] J. Zacharias, M. von Zahn, J. Chen, O. Hinz, Designing a feature selection method based on explainable artificial intelligence, Electronic Markets 32 (2022) 2159–2184.

[14] K. Deb, H. Jain, An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part i: solving problems with box constraints, IEEE transactions on evolutionary computation 18 (2013) 577–601.

[15] S. Chand, M. Wagner, Evolutionary many-objective optimization: A quick-start guide, Surveys in Operations Research and Management Science 20 (2015) 35–42.

[16] S. A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M. A. Riegler, P. Halvorsen, S. Parasa, On evaluation metrics for medical applications of artificial intelligence, Scientific reports 12 (2022) 5979.

[17] M. Pal, S. Bandyopadhyay, Many-objective feature selection for motor imagery eeg signals using differential evolution and support vector machine, in: 2016 International Conference on Microelectronics, Computing and Communications (MicroCom), IEEE, 2016, pp. 1–6.

[18] H. Dong, J. Sun, X. Sun, R. Ding, A many-objective feature selection for multi-label classification, Knowledge-Based Systems 208 (2020) 106456.

[19] H. Li, F. He, Y. Liang, Q. Quan, A dividing-based many-objective evolutionary algorithm for large-scale feature selection, Soft computing 24 (2020) 6851–6870.

[20] P. Korhonen, J. Wallenius, Visualization in the multiple objective decision-making framework, in: Multiobjective optimization: interactive and evolutionary approaches, Springer, 2008, pp. 195–212.

[21] J. Cheng, J. Sun, K. Yao, M. Xu, Y. Cao, A variable selection method based on mutual information and variance inflation factor, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 268 (2022) 120652.

[22] A. Benítez-Hidalgo, A. J. Nebro, J. García-Nieto, I. Oregi, J. Del Ser, jmetalpy: A python framework for multi-objective optimization with metaheuristics, Swarm and Evolutionary Computation 51 (2019) 100598.