Identifying maximal independent sets in many-objective optimization problems

Jorge Luiz Franco
Alexandre Cláudio Botazzo Delbem
Francisco José Mônaco
Kuruvilla Joseph Abraham
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - Brasil
jorge.luiz@usp.br acbd@icmc.usp.br monaco@icmc.usp.br abraham@fmrp.usp.br

Objectives

When dealing with Many-Objective Optimization problems, several obstacles are encountered. As described in [1], since the implementation of the first Multi-Objective Evolutionary Algorithm (MOEA), many others have been proposed. However, almost all of them have been used for problems with two or three objectives, which represent only a small portion of real-world problems.

Many of the designed MOEAs are based on Pareto optimality, and the main obstacle when working with many objectives ($n > 4$) is the curse of dimensionality. This is because the number of points required to faithfully represent the Pareto frontier grows exponentially with the number of objectives, with complexity $\mathcal{O}(nr^{n-1})$, which makes it challenging to solve Many-Objective Problems (MOPs), primarily due to the number of solutions that need to be evaluated. Additionally, visualization is of utmost importance in decision-making, which becomes unfeasible beyond the third dimension.

Therefore, the objective of this work is to compare existing techniques that can reduce the number of objectives without losing relevant information for the problem, i.e., finding a subset K of objectives that appropriately represent the set N of objectives, where $K \subset N$. To achieve this, algorithms proposed in [2] and [3] will be evaluated. Finally, we will propose an initial algorithm idea based on DAMICORE [3].

Methodology

The Duro's method described in [2] is:

The matrix $R = XX^T$, where X is the data and R is the linear covariance matrix. The kernel matrix K can be learned by solving the MVU-based semi-definite programming (SDP) problem:

$$\text{Max} \quad \text{trace}(K) = \sum_{ij} \frac{(K_{ii} - 2K_{ij} + K_{jj})}{2M}$$

$$(a) \quad \sum_{ij} K_{ij} = 0, \quad \forall \eta_{ij} = 1$$

$$(b) \quad K_{ii} - 2K_{ij} + K_{jj} = R_{ii} - 2R_{ij} + R_{jj}$$

$$(c) \quad \text{K is positive semi-definite}$$

Here, R_{ij} is the (i, j)-th element of the correlation matrix R, and η_{ij} is a binary indicator variable.

Eigenvalue Analysis: This step aims to eliminate non-conflicting objectives along the significant principal components: 1. Determine the number of significant principal components N_v as follows: $\sum_{j=1}^{N_v} \lambda_j \geq \theta$, where $\theta \in [0, 1]$ is a threshold parameter. 2. Interpret each significant principal component: a. Identify the objective f_i with the highest contribution to V_j (maximum absolute value). b. If all objectives have the same-sign contributions to V_j, select the top two contributions by magnitude.

Reduced Correlation Matrix (RCM) Analysis: Identify subsets of identically correlated objectives: 1. For each $f_i \in F_e$, identify potentially identically correlated subsets S_i based on correlation analysis. 2. If the correlation of f_j with f_i is stronger than a threshold T_{cor}, include f_j in S_i.

Selection Scheme for Objective Elimination: Compute a selection score for objectives within identically correlated subsets and retain the most significant objective: - Compute the selection score for each f_i in S using: score$_i$ = $\sum_{j=1}^{N_v} \lambda_j |f_{ij}|$ - Retain f_i with the highest score and eliminate the others, resulting in F_s (essential objective set).

The method proposed in [3] is based on FS-OPA, a clustering technique based on phylogram analysis. It uses DAMICORE to compute the distances from objectives and create a graph based on this distance matrix, then apply community detection for clustering. For each step, it uses (1) Normalized Compression Distance, (2) Neighbor Joining and (3) Fast Newman.

$$NCD(A,B) = \frac{C(AB) - \min[C(A),C(B)]}{\max[C(A),C(B)]} \qquad (1)$$

$$Q(i,j) = (n-2) \cdot d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k) \quad (2)$$

$$FN = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \qquad (3)$$

Finally, our new algorithm idea is based on the NCD generated distance matrix, compute a simplified graph with less edges that will represent the strongest correlated objectives, then perform a maximal independet set algorithm in the graph to find the independent set of objectives.

Results

A real problem in the Pantanal agroindustry, as described in [4], was used as a practical case study to apply algorithms techiniques. Since this and a lot of real world problems involve datasets that include non-numeric data types, the method proposed in [2] may not be well-suited. Real-world datasets often comprise a mix of categorical, textual, and numeric features. The transformations required to convert such data into numeric format can be problematic, potentially leading to loss of valuable information. Furthermore, Duro's Method using Kernel PCA, intuitively, try to find a higher dimension where the relations are linear, but, doing this it goes back to the curse of dimensionality. Additionally, when solving the SDP problem, it involves addressing the Pre-Image problem. The resolution of this problem depends on methods like fixed-point iteration, and their convergence relies on the Lipschitz constant, which can lead to convergence issues and getting stuck in local optima.

On the other hand, the methodology proposed in [3] suited really well in the Pantanal data, but needed a specialist to analyze the clusters in Figure 1 to choose the new set of objectives.

Conclusions

The comparative analysis of two methods, [2] and [3], has illuminated essential distinctions in their applicability to real-world datasets. While [3] exhibits remarkable versatility by accommodating a wide range of data types, including images, text, and other forms, it does necessitate the involvement of a domain specialist for cluster analysis.

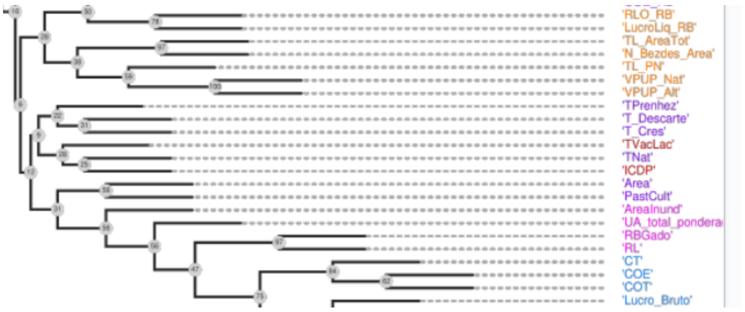

Figure 1: Consensus Tree from objectives

In contrast, our approach offers complete automation. By inputting the dataset of objectives, our algorithm efficiently identifies the maximal set of independent objectives, reducing their numbers. Leveraging Normalized Compression Distance (NCD), our method can compute distances across diverse data types and simplify complex graphs by optimizing the threshold.

Furthermore, our methodology excels in tackling the challenging problem of identifying the Maximal Independent Set in a graph. Although NP-Complete, efficient approximation algorithms ensure practicality across various real-world applications.

References

[1] JAIMES, A. Many-objective problems: Challenges and methods. Springer Berlin Heidelberg, 2015.

[2] DURO, J. A. et al. Machine learning based decision support for many-objective optimization problems. *Neurocomputing*, v. 146, p. 30–47, 2014. ISSN 0925-2312. Bridging Machine learning and Evolutionary Computation (BMLEC) Computational Collective Intelligence.

[3] GASPAR-CUNHA, A. et al. Many-objectives optimization: A machine learning approach for reducing the number of objectives. *Mathematical and Computational Applications*, v. 28, n. 1, 2023. ISSN 2297-8747.

[4] SANTOS, S. A. et al. *Recomendações técnicas para o planejamento da introdução de forrageiras exóticas de forma sustentável no Pantanal*. Corumbá, 2022. 26 p.