A multi-objective approach for the (α, β)-k-Feature Set Problem using Memetic Algorithms

Francia Jiménez, Claudio Sanhueza, Regina Berretta, Pablo Moscato
School of Electrical Engineering and Computing, The University of Newcastle, NSW, Australia
{francia.jimenezfuentes,claudio.sanhuezalobos}@uon.edu.au,{regina.berretta,pablo.moscato}@newcastle.edu.au

ABSTRACT

Nowadays, the ceaseless data gathering in science and technology is bringing new challenges. Companies use the collected data to create new digital services and products. These services rely on innovations in data mining algorithms. The combinatorial search required to select a subset of features that best describe classes in a dataset is a challenging problem in this research field. The $(\alpha, \beta) - k-$ Feature Set Problem is a mathematical model proposed for addressing this task. In its most common optimization variant, the problem has always been to find the minimum number of features for given fixed values of α and β that satisfy the requirements of the model. However, the relation between the α and β parameters and the number of features is unknown. In the literature, multi-objective approaches have been used, with great success, to address problems that require optimizing several objectives simultaneously. In this study, we propose a novel multi-objective approach for solving the $(\alpha, \beta) - k-$ Feature Set Problem using memetic algorithms. We study and evaluate different local searches and initialization procedures using six well-known datasets. Our results show that the clustering-based local search heuristic has a positive impact on the quality of the solutions.

CCS CONCEPTS

•Applied computing → Multi-criterion optimization and decision -making;

KEYWORDS

(α, β)-k-Feature Set Problem, Multi-Objective Optimization, Memetic Algorithm.

ACM Reference format:
Francia Jiménez, Claudio Sanhueza, Regina Berretta, Pablo Moscato . 2017. A multi-objective approach for the (α, β)-k-Feature Set Problem using Memetic Algorithms. In *Proceedings of GECCO '17 Companion, Berlin, Germany, July 15-19, 2017,* 2 pages.
DOI: http://dx.doi.org/10.1145/3067695.3076106

1 INTRODUCTION

The (α, β)-k-Feature Set Problem is a generalization of the k- Feature Set Problem[4]. The problem aims to find a subset of features that can explain a *class* value, considering the support of samples with different class value (α) and the support of samples with same class value (β). Considering a dataset with m samples and n features, the formal definition of the (α, β)-k-Feature Set Problem is:

Instance:	A discrete valued $m \times n$ matrix $\mathcal{D}$, a discrete valued $m \times 1$ vector $\mathcal{T}$, and positive integers α, β and k.
Parameter:	k
Question:	$\exists\, S \subseteq [1, \ldots, n]$, $\lvert S\rvert \leq k$ such that $\forall\, i, j \in [1, \ldots, m]$ and $\circ$ if $t_i \neq t_j$, $\exists\, S' \subseteq S$ where $\lvert S'\rvert \geq \alpha$ and $\forall\, s \in S'\ d_{i,s} \neq d_{j,s}$ $\circ$ if $t_i = t_j$, $\exists\, S' \subseteq S$ where $\lvert S'\rvert \geq \beta$ and $\forall\, s \in S'\ d_{i,s} = d_{j,s}$?

The (α, β)-k-Feature Set Problem has been applied in Bioinformatics to find a subset of features (genes) over-expressed/under-expressed in a particular disease [1, 7]. The problem has been addressed as a single-objective optimization problem for which the objective is to minimize the number of features. This optimization problem has been solved using integer programming models [1, 7] and meta-heuristics [5]. However, we can define the problem as a multi-objective optimization problem, for which the number of features has a contradictory relation with the values of the parameters α and β. Since the mid-1980s, the research community has created a variety of techniques to deal with multi-objective optimization problems [3]. As a result, nowadays there is a set of techniques able to address multi-objective optimization problems without reducing their inherent complexity.

In this paper, we present a novel multi-objective approach which uses a memetic algorithm (MA) to optimize different objectives of the (α, β)-k-Feature Set Problem. We study the algorithm's performance of different local searches and initialization approaches. We evaluate our algorithms using six well-known datasets. Our experimental results show that our clustering-based approach is suitable to address the (α, β)-k-Feature Set Problem and brings a new multi-objective perspective to this domain.

2 MEMETIC ALGORITHM FOR MULTI-OBJECTIVE (α, β)-K- FEATURE SET PROBLEM

Our multi-objective optimization problem considers three objectives, where $S \subseteq \{1, ..., n\}$ and $\mathcal{X}$ is the set of all subsets of features:

$$\min_{S \in \mathcal{X}} \{k(S), -\beta(S), -C(S)\} \tag{1}$$

- $k(S)$ is the number of selected features. Formally, $k(S) = \sum_{j=1}^{n} f_j$ where $f_j = \begin{cases} 1, & \text{if } j \in S \\ 0, & \text{otherwise} \end{cases}$

GECCO '17 Companion, Berlin, Germany

DOI: http://dx.doi.org/10.1145/3067695.3076106

Table 1: Description of datasets and experimental results. For each dataset, we present the number of features and samples. As results, we present the mean of the normalized hypervolume and the execution time (secs.) accross 30 trials for each MA.

			Normalized Hypervolume				Execution Time in seconds			
Dataset	**Features**	**Samples**	**MA1**	**MA2**	**MA3**	**MA4**	**MA1**	**MA2**	**MA3**	**MA4**
Bruta	683	83	0.4551	**0.5451**	0.4087	0.4103	2305	2642	**1936**	2030
DownSyn	73	15	0.3313	0.4793	**0.5319**	0.4934	**57**	79	108	118
Parkinson	17099	105	0.3501	0.4347	0.3769	**0.5216**	182325	290295	**65282**	82332
PdParkinson	1674	25	0.4446	0.4612	0.4702	**0.4819**	2615	8882	**2206**	5006
Prostate	3556	171	0.3830	**0.5853**	0.4587	0.4804	13629	45482	**5117**	11204
Smoking	525	1219	0.4412	**0.5943**	0.4886	0.4731	**16680**	50515	42862	91548

- $\beta(S)$ is the minimum number of features with the same value in any pair of samples with the same class value. Formally, $\beta(S) = \min_{1<p,q<m} \sum_{j=1}^{n} b_{jpq}$,
  where $b_{jpq} = \begin{cases} 1, & \text{if } j \in S \text{ , } t_p = t_q \text{ and } d_{p,j} = d_{q,j} \\ 0, & \text{otherwise} \end{cases}$
- $C(S)$ is the *total* number of pair of samples connected with the selected features. Formally, $C(S) = \sum_{j=1}^{n} \sum_{p=1}^{m} \sum_{q=1}^{m} cov_{jpq}$,
  where $cov_{jpq} = \begin{cases} 1, & \text{if } j \in S \text{ , } t_p \neq t_q \text{ and } d_{p,j} \neq d_{q,j} \\ 1, & \text{if } j \in S \text{ , } t_p = t_q \text{ and } d_{p,j} = d_{q,j} \\ 0, & \text{otherwise} \end{cases}$

For our optimization problem, the value of α is fixed and computed as $\alpha^* = \alpha(\{1, 2, ..., n\})$. Formally, $\alpha(S) = \min_{1<p,q<m} \sum_{j=1}^{n} a_{jpq}$,
where $a_{jpq} = \begin{cases} 1, & \text{if } j \in S \text{ , } t_p \neq t_q \text{ and } d_{p,j} \neq d_{q,j} \\ 0, & \text{otherwise} \end{cases}$

We propose a novel memetic algorithm, where the individual is represented by a bit array of size n where each bit has the value of 1 if the feature is selected and 0 otherwise. To initialize the individuals, we use a random selection process. The recombination procedure is *Intersect*, which is the AND logical operator between two bit arrays. Our mutation strategy is Deterministic Bit Flip at u positions, *DetBitFlip(u)*, which performs a flip ($0 \rightarrow 1$; $1 \rightarrow 0$) over $u = 3$ features. Then, we perform a local search procedure over the individuals generated by the mutation procedure. Finally, we use *Elitism* to determine which individuals remain in the population and how to include the new individuals into the population for the next generation. Our stopping criteria is to run the algorithm for a maximum number of generations ($maxGen = 100$).

3 EXPERIMENTS AND RESULTS

To test our multi-objective memetic algorithm, we use two initialization procedures and two different local searches. The algorithms are: MA1: *RandomIni procedure* and *GreedyLS heuristic*; MA2: *RandomIni procedure* and *ClusterLS heuristic*; MA3: *ClusterIni procedure* and *GreedyLS heuristic*; MA4: *ClusterIni procedure* and *ClusterLS heuristic*. The procedures *ClusterIni procedure* and *ClusterLS heuristic* utilize clustering information to select the features. In our experiments, we use six life-science datasets used previously in [5]. In Table 1, we present the main characteristics of each dataset such as the total number of features and the number of samples. We use ***Hypervolume*** [2] to evaluate the performance of the multi-objective algorithm. A higher hypervolume value means a better performance of the algorithm.

In Table 1, we present the normalized hypervolume and the time in seconds in average for each dataset. We highlight the highest hypervolume and the lowest execution time per dataset. We apply the Wilcoxon signed-rank test [6] to evaluate if the differences are significant. The MA2 algorithm has the highest hypervolume, and it has a significant difference (p=0.031) with MA1, where the only difference between them is the local search procedure. The algorithm with the lowest average time is MA3, and it has a significant difference with all the others algorithms.

4 CONCLUSIONS AND FUTURE WORK

We proposed a novel multi-objective approach to tackling the (α, β)-k-Feature Set Problem using memetic algorithms.

As a result, we observe that our proposed heuristic for the local search, *ClusterLS heuristic*, has a positive impact on the performance when we use the *RandomIni procedure*. However, the use of *ClusterLS heuristic* increases the execution time.

Although our results show that our multi-objective approach is suitable to address the (α, β)-k-Feature Set Problem, more studies are needed to improve our approach. In particular, we want to process large-scale datasets which have thousand of samples and million of features.

REFERENCES

[1] Regina Berretta, Alexandre Mendes, and Pablo Moscato. 2007. Selection of discriminative genes in microarray experiments using mathematical programming. *Journal of Research and Practice in Information Technology* 39, 4 (2007), 287–299.

[2] Shi Cheng, Yuhui Shi, and Quande Qin. 2012. On the performance metrics of multiobjective optimization. In *Advances in Swarm Intelligence*. Springer, 504–512.

[3] Carlos A Coello and Nareli Cruz Cortés. 2005. Solving multiobjective optimization problems using an artificial immune system. *Genetic Programming and Evolvable Machines* 6, 2 (2005), 163–190.

[4] Carlos Cotta, Christian Sloper, and Pablo Moscato. 2004. Evolutionary Search of Thresholds for Robust Feature Set Selection: Application to the Analysis of Microarray Data. In *Applications of Evolutionary Computing*. Lecture Notes in Computer Science, Vol. 3005. Springer Berlin Heidelberg, 21–30.

[5] Mateus Rocha de Paula, Regina Berretta, and Pablo Moscato. 2015. A fast memetic approach for the (α,β)-k-feature set problem. *Journal of Heuristics* (2015), 1–22.

[6] Warren J Ewens and Gregory R Grant. 2006. *Statistical methods in bioinformatics: an introduction.* Springer Science & Business Media.

[7] Mou'ath Hourani, Alexandre Mendes, Regina Berretta, and Pablo Moscato. 2007. Genetic biomarkers for brain hemisphere differentiation in Parkinson's Disease. In *Computational Models For Life Sciences International Symposium on Computational Models of Life Sciences*, Vol. 952. AIP Publishing, 207–216.