The multi-objective approach to solve the (α, β)-k Feature Set Problem using Memetic Algorithms

by

Francia Jimenez

Thesis

submitted in partial fulfilment

of the requirements for the Degree of

Doctor of Philosophy in Computer Science

Supervisor: Prof. Pablo Moscato

Co-Supervisor: Prof. Regina Berretta

THE UNIVERSITY OF
NEWCASTLE
AUSTRALIA

School of Electrical Engineering

and Computing

March, 2019

Statement of Originality

I hereby certify that the work embodied in the thesis is my own work, conducted under normal supervision. The thesis contains no material which has been accepted, or is being examined, for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made. I give consent to the final version of my thesis being made available worldwide when deposited in the University's Digital Repository, subject to the provisions of the Copyright Act 1968 and any approved embargo.

Acknowledgment of Authorship

I hereby certify that the work embodied in this thesis contains published papers of which I am a joint author. I have included as part of the thesis a written declaration endorsed in writing by my supervisor, attesting to my contribution to the joint publications. By signing below I confirm that Francia Jimenez contributed as the first author of the publications entitled “A multi-objective approach for the (α, β)-k-feature set problem using memetic algorithms” and “Accelerating a multi-objective memetic algorithm for feature selection using hierarchical k-means indexes.”

Acknowledgments

Thanks for all the support, love and patience provided by my husband Claudio. Without you and your words of encouragement none of this would have been possible. You are really special to me, your kind words, and thoughts, complemented with your smile and hugs are the source of my daily energy. Thanks dear for those quick lunches and long conversations during the afternoons, those moments were the source of amazing ideas that helped me to make better my research. You really inspired me and strengthened my ideas during this Ph.D. I couldn't ask for a better partner, you are the best one in the world. Thanks honey for walking with me during this journey.

Many thanks to Regina Berretta and Pablo Moscato, my supervisors during the PhD. Their feedback and advice have significantly improved the quality of this work. Thank you for enriching this Ph.D. by sharing your knowledge, ideas, and life experiences with me.

To my mum Wilma, my dad Albeerto and my sister Belgica infinite thanks because this achievement is also part of you. Thank you for teaching me that in life everything can be achieved with perseverance. Thanks for your unconditional support, I will be there always for you. May love always be in our family, specially now that we have Carlos, Rafaela y Claudio.

I want to thank my friends, Aussies and Chileans, for always listening to the same and not getting bored with me. To those who helped me to dry more than one tear and those who helped me to reduce the stress during this journey. Let's take care of this special way of love that exists between us.

Finally, I would to thank the University of Newcastle for helping me to develop my career with a postgraduate study. Thanks to the support and daily work of the staff members of the university, this journey without them would be much more complicated. Thanks to the member of the research laboratory, who shared this 4 year journey with me. Ademir, Ahmed, Amer, Amir, Heloisa, Inna, Leila, Luke, Mohammad, Nader, Nasimul, Natalie, Nisha, and Shannon, all the best for you.

Francia Jimenez
The University of Newcastle
March 2019

Contents

List of Tables

List of Figures

List of Algorithms

Acronyms

α α. xiii, 2, 11–15, 17, 30, 135

α^* α-`constraint`. 29, 33–35, 38–40, 60, 64, 65, 73–75, 110, 131

β β. xiii, 2, 3, 11–17, 30

ClusterInit *Cluster Initialization heuristic*. 64, 66–69, 83

ClusterSH *Cluster Satisfaction heuristic*. 64–69, 71, 83

CoverSH *Cover Satisfaction heuristic*. 39–41, 44–46, 51, 60

GreedySH *Greedy Satisfaction heuristic*. 38, 41, 44–46, 51, 60, 66–69, 83, 89, 92

IndexInit *Index Initialization heuristic*. 73, 74, 76–78, 80–82, 89, 92

IndexSH(25%) *IndexSH(25%) heuristic*. 89, 92, 93

IndexSH(3) *IndexSH(3) heuristic*. 89, 92, 93

IndexSH *Index Satisfaction heuristic*. 74–77, 81, 93

RandomInit *Random Initialization procedure*. 34, 66–69, 71, 76–78, 80–83, 89, 92, 93

k-means *k-means* algorithm. 66, 67, 72, 73, 77

`ABkFS` (α, β)-k-Feature Set Problem. xii, xiii, 2–7, 11–17, 19, 27, 29–31, 35, 38, 42, 43, 60, 61, 63, 71, 74, 110, 133–136

`CABkFS` Coloured (α, β)-k Feature Set Problem. 16

`FSA` feature selection algorithms. 8, 33, 34

`FS` Feature Selection. xii, 2, 7, 9, 26, 32

`Fs` Feature selection. 2, 7

`MA` Memetic Algorithm. 31, 33

`MA` memetic algorithm. 4–6, 30–33, 37, 40, 44, 45, 60, 61, 66, 67, 69, 74, 75, 77, 89, 91, 133–136

`MO-ABK` Multi-objective (α, β)-k-Feature Set Problem. 17, 19, 27, 29, 30, 33, 135

`MOA` Multi-objective optimization algorithms. xii, 4

`MOA` multi-objective optimization algorithms. 5, 24, 26

`MOMA-ABK` Memetic Algorithm for Multi-objective (α, β)-k-Feature Set Problem. xiii, 5, 6, 30–42, 44, 46, 48, 50–52, 54, 56, 58, 60–80, 82, 84–92, 94, 96, 98, 100, 102, 104, 106, 108, 110, 112, 114, 116, 118, 120, 122, 124, 126, 128, 130, 132–135

`MOMA` multi-objective memetic algorithm. 31, 133

`MOPs` Multi-objective optimization problems. xii

`MOP` Multi-objective optimization problem. 20, 21, 23

`MOP` multi-objective optimization problem. xii, 4, 19–23, 28

`MO` Multi-Objective. 26

`MO` multi-objective optimization. 24, 27, 30

`MO` multi-objective. xiii, 4–6, 17, 19, 21, 23, 24, 26, 27, 30, 38, 43, 60, 71, 85, 87, 109, 110, 133–136

`SOO` single-objective optimization. 26

`SOP` single-objective optimization problem. 19, 21

`SO` Single-objective. 4, 24

`SO` single-objective. 4, 19, 21, 22, 26, 128, 134

`fs` feature selection. 2, 3, 5, 7, 8, 26, 32, 33, 52

`k-FS` k-Feature Set Problem. xii, 5, 7, 9–12, 17, 26, 33

GPF *Global Pareto front*. 88, 91, 101, 105, 110, 119, 123, 124, 128, 130

J48 C4.5-based decision trees. 91, 93–96, 98–104, 110–122, 128–130

KNN k-Nearest Neighbor. 33, 91, 93–104, 110–122, 128–131

MC *Minimum Coverage*. 88, 91, 93, 105, 110, 123

MF *Minimum Feature*. 88, 91, 97, 105, 110, 115, 123

NB Naive Bayes. 91, 93–104, 109–122, 124, 128–130

Satisfaction Heuristic satisfaction heuristic. 33, 35, 37, 38, 41, 44–46, 51, 61, 65, 66, 68, 69, 71, 74–76, 79, 81, 87, 89, 91–93, 131, 133–135

SVM Support Vector Machine. 91, 93–104, 109–122, 124, 128–131

Abstract

In many application areas, the decision-making process is enhanced by the information obtained from analyzing data. In fact, the process of improving digital products and services can be driven by insights from understanding complex relationships inside the data. Commonly to have a complete picture of the process, the data is obtained from multiple sources. Each source stores different type of data that it is essential for the specific data source. However, when we aggregate different sources, the new data can have some elements that can be considered as unreliable, irrelevant, or redundant for a specific problem. The previous challenge is known as Feature Selection (`FS`) and commonly presented during data integration. The k-Feature Set Problem (`k-FS`) is a problem in `FS`, that aims to find the minimum subset of features necessary to describe a dataset. Similarly, the (α, β)-k-Feature Set Problem (`ABkFS`) also aims to find the minimum subset of features, but in addition the subset of features needs to satisfy two conditions: α and β, where the α value is related with the differentiation power and the β value is related with the representation power of the subset of features. Commonly the `ABkFS` is used to reduce the number of features on datasets where the number of features is higher than the number of samples. This type of datasets can be found in bioinformatics where a few numbers of samples (e.g. corresponding to a set of biological samples obtained from individuals/patients) have their gene expression (features) measured in a quest to characterize a specific disease. In the literature, state-of-the-art feature selection techniques do not report good performance when analyzing this type of dataset because they use univariate tests which are commonly based on statistical measures across the samples. Currently, the `ABkFS` has been solved with exact models and also heuristics have been employed only based on single objective approach. However, there is a need to consider a multi-objective approach since the minimization of the number of features (usually required to achieve better generalization) "conspires" against the requirements of having a large value of α and β. This then constitutes a typical scenario in which the multi-objective approach is the most natural alternative.

Many engineering solutions are developed using optimization techniques where we formally define an optimization problem which is composed by an objective function (or metric of interest) that we will optimize (minimize or maximize). A more realistic strategy of modeling optimization problems is assessing many objectives simultaneously, formally known as Multi-objective optimization problems (`MOPs`), where the main goal is to optimize multiple and possibly conflicting objectives. The conflict between two objectives functions is when improving the value of one of them worsen the second one. A special type of algorithms has been developed to solve `MOP` which are known as Multi-objective optimization algorithms (`MOA`). As a result of this type of algorithms, we have a set of solutions that between them we can not establish which one is better, and the set represents the trade-off that exists between the objectives that are being optimized. In the literature, these multi-objective techniques are generating good results in a variety of complex problems. Commonly, multi-objective techniques are used to implement wrapper feature selection

approaches. Therefore, developing a multi-objective filter feature selection is a challenge and the exploration of this niche area with new optimization techniques is promising if we consider the benefits of multi-objective approaches.

In the first contribution of this dissertation, we design and implement an efficient Memetic Algorithm for Multi-objective (α, β)-k-Feature Set Problem (`MOMA-ABK`). The (α, β)-k-Feature Set Problem (`ABkFS`) aims to find a subset of features able to "cover" α times each pair of samples with different class values and each pair of samples with the same class value "covered" β times. We use a multi-objective optimization approach mainly because is unknown the relationship between α, β, and the number of features. Additionally, we improved the performance of our algorithm by including information during the optimization process. We considered information from the relationship between features by applying clustering techniques between the features and storing features efficiently on a search tree structure. We experiment with six real-life datasets and our results shown that the use of the search tree structure improves the performance of the algorithm.

Considering the challenging area of analyzing high-dimensional datasets, our second contribution is a novel multi-objective (`MO`) filter feature selection algorithm. We proposed a filter feature selection methodology based on the (α, β)-k-Feature Set Problem composed by four stages: preprocessing, `MOMA-ABK`, classification and postprocessing. In order to integrate several Pareto front into one set of representative solutions, we proposed and implemented three novel approaches. In addition, we studied the impact in the performance of the filter feature selection approach of the α value considered during the optimization process. Our experiments have shown that our approach has competitive performance in comparison with state-of-the-art algorithms.

Publications and Outcomes

The material presented in this thesis has been already published, or accepted for publication, in peer-reviewed journals and conferences. The list of publications is provided below.

Conference papers

Jiménez, Francia and Sanhueza, Claudio and Berretta, Regina and Moscato, Pablo A multi-objective approach for the (α, β)-k-feature set problem using memetic algorithms. *Proceedings of the Genetic and Evolutionary Computation Conference Companion 2017, ACM*

Jiménez, Francia and Sanhueza, Claudio and Berretta, Regina and Moscato, Pablo Accelerating a multi-objective memetic algorithm for feature selection using hierarchical k-means indexes. *Proceedings of the Genetic and Evolutionary Computation Conference Companion 2018, ACM*

Posters

Jiménez, Francia and Riveros, Carlos and Moscato, Pablo A multi-objective memetic algorithm for (α, β)-k-feature set problem. *Faculty of Engineering and Built Environment, University of Newcastle, 2015.*

Jiménez, Francia and Berretta, Regina and Moscato, Pablo A new filter feature selection algorithm for classification: An application of the (α, β)-k-feature set problem. *Faculty of Engineering and Built Environment, University of Newcastle, 2016.*

Jiménez, Francia and Sanhueza, Claudio and Berretta, Regina and Moscato, Pablo A multi-objective filter feature selection algorithm applied to high-dimensional microarray datasets. *Faculty of Engineering and Built Environment, University of Newcastle, 2017. Awarded the 2017 Research Poster Prize*

Other publications produced during the time of this dissertation

Sanhueza, Claudio and Jiménez, Francia and Berretta, Regina and Moscato, Pablo
mQAPViz: A divide-and-conquer multi-objective optimization algorithm to compute large data visualizations. *Congress on Evolutionary Computation (CEC), 2017 IEEE*

Sanhueza, Claudio and Jiménez, Francia and Berretta, Regina and Moscato, Pablo
PasMoQAP: a parallel asynchronous memetic algorithm for solving the Multi-Objective Quadratic Assignment Problem. *Proceedings of the Genetic and Evolutionary Computation Conference 2018, ACM*